Automatic Adaptation of Sound Analysis and Synthesis

Marco Liuni

Co-directed thesis, to be defended on March the 9^{th} 2012, to obtain the grade of DOTTORE DI RICERCA IN MATEMATICA from the UNIVERSITÀ DEGLI STUDI DI FIRENZE - DIPARTIMENTO DI MATEMATICA "ULISSE DINI", and the grade of DOCTEUR DE L'UNIVERSITÉ PARIS VI - PIERRE ET MARIE CURIE (UPMC) with major in Signal Processing from the ECOLE DOCTORALE INFORMATIQUE, TÉLÉCOMMUNICATIONS ET ELECTRONIQUE (EDITE)

Reviewers	Peter Balazs	Director - ARI, Austrian Academy of Sciences
	Fulvio Gini	Professor - DII-EIT , University of Pisa
Examiners	Monika Dörfler	Researcher - NuHAG, University of Vienna
	Alvise Vidolin	Researcher - DEI, University of Padua
	Bruno Gas	Professor - UPMC
	Marco Romito	Professor - University of Pisa
	Axel Röbel	Researcher - IRCAM
	Xavier Rodet	Emeritus Researcher - IRCAM

This work has been supervised by Xavier Rodet, Marco Romito and Axel Röbel, and conducted at DIPARTIMENTO DI MATEMATICA U. DINI and INSTITUT DE RECHERCHE ET COORDINATION ACOUSTIQUE/MUSIQUE (IRCAM)

Ircam - CNRS-UMR9912-STMS
Sound Analysis/Synthesis Team
1, place Igor Stravinsky
75004 Paris, FRANCE

May 2, 2012

For any comment, suggestion, or correction please contact the author.
leehooni at gmail

To my parents and my grandparents, to my brother
To Elena, and to Nina

Contents

CHAPTER 1

Context, motivations and objectives of the work

Sound processing techniques are employed over a wide area of research and industrial applications: music first comes to mind, together with the community of composers, producers and multimedia artists as well as the professionals of entertainment; then we have speech, which is elaborated in many different ways in our everyday life. Smartphones, tablets and any other kind of mobile devices, as well as TVs and home theater set-ups, computers, digital equipment for music and film studios: all of them deal with sound in digital format and come with different and challenging needs, which rise many interesting research and technological issues. Several fields other than music or speech exploit sound analysis, transformation and synthesis: medical sciences, security instruments and communications, among others.

Traditional sound analysis methods, based on single sets of atomic functions like Gabor windows or wavelets, offer limited possibilities concerning the flexibility of their time-frequency precision. Moreover, fundamental analysis parameters have to be set a-priori, according to the signal characteristics and the quality of the representation required. Analyses with a non-optimal resolution lead to a blurring, or sometimes even a loss of information about the original signal, which affects every kind of later treatment.This problem concerns a large part of the technical applications dealing with signals: visual representation, feature extraction and processing among others; the community working on these issues is a very broad one, including telecommunications, sound and image processing as well as applied mathematics and physics. Our main interest is focused on sounds, and our questions principally rise from the musical and voice domains. The mainstream industrial fields more strictly related to this topic are signal transformation, music production, speech processing, source separation and music information retrieval, the latter covering a broad range of applications from classification, to identification, feature extraction and information handling about music: many of the algorithms applied within these processes rely on a given time-frequency representation of the signal, inheriting its qualities and drawbacks, and would therefore benefit from adapted analyses with optimized resolutions. This motivates the research for adaptive methods, conducted at present in both the signal and the applied mathematics communities: they lead to the possibility of analyses whose resolution locally changes according to the signal features.

This thesis starts from the main idea that algorithms based on adaptive representations will help to establish a generalization and simplification for the application of signal processing methods that today still require expert knowledge. An automatic

parameter selection would allow to achieve more robust methods with significantly less human effort. Our attention is focused in particular on advanced signal processing methods in applications designed for large communities: the need to provide manual low level configuration is indeed one of the main problems. The possibility to dispose of an automatic time frequency resolution drastically limits the parameters to set, without affecting, and even ameliorating, the analysis quality: the result is an improvement of the user experience with advanced signal processing techniques that require, at present, a high expertise.

The first and fundamental objective of our project (Chapter 2) is thus the formal definition of mathematical models whose interpretation leads to theoretical and algorithmic methods for adaptive analysis. *Gabor frames theory* constitutes a very natural mathematical context: one of its main subjects is the definition of redundant sets of atoms in Hilbert spaces, generally larger than orthonormal bases, together with the associated decomposition operators and their inverses. Actually, using that for sound processing requires the possibility of reconstructing a signal from its analysis coefficients: thus we need an efficient way to find an inverse of the adaptive decomposition operator, together with appropriate methods to manage adaptive analyses in order to preserve and improve the existing sound transformation techniques.

The second objective (Chapter 3) is to make this adaptation automatic; we aim to establish criteria to define the optimal local time-frequency resolution: we deduce such criteria from the optimization of given *sparsity measures*. We take into account both theoretical and application-oriented sparsity measures: entropies and other quantities borrowed from information theory and probability belong to the first class. When dealing with concrete sounds, information measures may not always be well-suited, since some of their characteristics do not find a direct interpretation in the signal domain. Thus, it is often useful to give application-driven definitions of sparsity, depending on the particular features that the system should privilege.

This first chapter deals with the scientific and historical motivations of the work, while Chapters 4 and 5 present the algorithms that we have realized, together with a description of their properties, applications and results.

1.1. List of publications issued from this work

[**Liuni et al., 2011c**] M. Liuni, A. Röbel, M. Romito, X. Rodet, “Rényi information measures for spectral change detection,” in Proc. of ICASSP11, Prague, Czech Republic, May 22-27, 2011
[**Liuni et al., 2011a**] M. Liuni, P. Balazs, A. Röbel, “Sound analysis and synthesis adaptive in time and two frequency bands,” in Proc. of DAFx11, Paris, France, September 19-23, 2011
[**Liuni et al., 2010**] M. Liuni, A. Röbel, M. Romito, X. Rodet, “A reduced multiple Gabor frame for local time adaptation of the spectrogram,” in Proc. of DAFx10, Graz, Austria, September 6-10, 2010, pp. 338 – 343
[**Liuni et al., 2011b**] M. Liuni, A. Röbel, M. Romito, X. Rodet, "An entropy based

method for local time-adaptation of the spectrogram. In S. Ystad, M. Aramaki, R. Kronland-Martinet and K. Jensen, editors, "Exploring Music Contents", volume 6684 of Lecture Notes in Computer Science, pages 60–75. Springer Berlin / Heidelberg.

1.2. Sounds and Music

In this section, a few historical guidelines are given: to make a computer deal with sounds has been a challenge almost as soon as the first computers have been conceived; a complete review of the pioneers' different approaches is out of the scope of this work (see [**Chadabe, 1997**] for a survey, including details about composers and scientists mentioned in this section), but it is worth to mention some researchers and composers who have delineated the fields of Sound Processing and Computer Music where this thesis is inscribed. In particular, we focus on the origins of the application of Fourier theory to the representation, transformation and synthesis of sounds through computer programs: stationary sinusoids can be considered as elementary stimuli, whose superpositions and modulations originate sounds of higher complexity; from an engineering point of view, sinusoids could be easily generated, either in the first electroacoustic studios by means of analogical oscillator, or by the first computers with look up tables of a few points. Moreover, western music theory and instruments are grounded on harmonic principles, whose interpretation can easily be formulated in terms of superposed sinusoids. For all of these reasons, since the origins of electroacoustic laboratories, Fourier-based representations have been adopted and intensively experimented for sound analysis and synthesis.

The problem of conceiving a sound representation by means of a computer has first been handled by scientists who were musicians, too: not all of them were composers, but their work laid the foundations for a considerable part of the next generations composers. Max Mathews and John Pierce worked at one of the first sound-generating computer program, in the sixtieth of the past century: they both were scientists of the Bell Telephone Laboratories (also known as Bell Labs), in Murray Hill New Jersey, and originated the *Music-N* series of programs, whose principles are still adopted in many real-time music softwares. Their work inspired several other people: Jean Claude Risset, a composer and physician graduated at the École Normale Supérieure in Paris, came to the Bell Labs for the first time in 1964. He was interested in the timbre of musical instruments, and his first work at the Bel Labs was on the synthesis of trumpet sounds: this is a hard task, disposing uniquely of an additive synthesis with a small number of voices, that is, sinusoids to add together. His strategy consisted of three main steps: the analysis of spectra from different trumpet samples, the experimental deduction of a small number of relevant components, then the synthesis of a mixture of sinusoids whose parameters were set according to the analytical results.

Risset's method is an example of constructive approach to sound synthesis: a complex spectrum is obtained as a sum of several basic components. John Chowning, a composer graduated at Stanford University, read about Mathews' research in the same period, and started working with his programs too: his *FM Synthesis* technique (Frequency Modulation), whose patent was issued in 1975 and licensed to Yamaha in

1977, came from his research about the generation of fast vibratos, and is an example of alternative approach to sound generation, still in a constructive sense: complex spectra are generated with a given model based on sinusoids and depending on a few parameters, whose values determine a predictable structure of the synthesized sound spectrum. In 1975, Chowning formed the CCRMA (Center for Computer Research in Music and Acoustics) in Stanford, together with other researchers: James A. Moorer, who was co-director and co-founder, developed advanced DSP techniques for analysis and synthesis of musical sounds. A section of his PhD thesis ([**Moorer, 1975**]), presented in 1975, is dedicated to the *Heterodyne filter*, which is a first example of improvement of the Fourier classic transform towards a time-frequency varying representation: the input is a tone whose fundamental frequency is known, the output is a series of sinusoids with amplitude and phase varying over time, which are harmonics of the fundamental frequency. In a following paper ([**Moorer, 1976**]), published in 1976, he introduces a class of synthesis techniques based on discrete summations of time-varying sinusoids, whose capabilities and control are similar to those of Chowning's frequency modulation technique, with the advantage that the signal can be exactly limited to a specified number of partials.
He also worked at improvements of the *Phase Vocoder* technique ([**Moorer, 1978**]): originally introduced in 1966 ([**Flanagan, 1966**]) by Flanagan, working at the Bell Labs, this technique is based on the STFT (Short Time Fourier Transform, see Section 1.5); it began to be widely exploited when the dedicated algorithms were made computationally fast enough (see [**Cooley and Tukey, 1965**] for the Fast Fourier Transform original algorithm, and [**Portnoff, 1976**] for an implementation of STFT taking advantage of the FFT). The input of the STFT is a generic sound, the output is a set of coefficients which allow a perfect reconstruction of the original sound in terms of atomic signals, which are weighted modulated sinusoids. The advantage, here, is that no knowledge about the fundamental frequency is needed, thus making the method well suited for a broad range of sounds. On the other hand, the representation is no longer related to sinusoids, so that they have to be deduced from the coefficients by means of *sinusoidal modeling* techniques (see Section 1.3). Despite of its drawbacks, that we detail in the words of a composer later in this section, a broad range of current sound processing techniques are based on phase vocoder and its improvements (see [**Griffin and Lim, 1984**, **Laroche and Dolson, 1999**] and the related bibliographies).

When his work about the phase vocoder was published, Moorer was working at IRCAM (Institut de Recherche et Coordination Acoustique/Musique) in Paris, France. The creation of the institute started in 1970, by Pierre Boulez, who received the invitation of the French president Georges Pompidou; there was a continuity with the activities and researches going on at Bell Labs and CCRMA, and Boulez entrusted several people from these laboratories in charge of direction: Risset was the head of the computer department, while Mathews was appointed scientific advisor in 1974. From 1977 to 1979, Moorer had the role of scientific advisor and researcher. The official opening was held in 1977, and by 1978 the IRCAM three levels underground building included laboratories dedicated to the research activity, as well as recording studios, an anechoic chamber and a concert hall with advanced possibilities for the design of

the internal acoustic. The research about spectral processing techniques based on the phase vocoder has been, and still is, a main topic for the computer department and the Analysis/Synthesis Team: the latter has been created when the computer department was divided in specialized branches. This research has lead to the *SuperVP* library [1], an extended phase vocoder which is used by a large number of composers and integrated in *AudioSculpt* [2], a software for viewing, analysis and processing of sounds.

Time-frequency analysis is a natural context for the modeling of time-evolving spectra, thus in particular for sounds and music. Since their introduction, STFT-based models for sound analysis and processing have been applied in a wide range of musical communities, determining new paradigms of thinking sounds. An example is given by the musical current of *Spectralism*, where time-frequency representations of sounds become organizing principles for the structure, the formal articulation, and the materials of a piece of music. Originated in early seventies by the works of Gérard Grisey and Tristan Murail, spectralism constituted an attitude rather than a school: from the analysis point of view, the main interests were a quantitative description of sound spectra and a rigorous characterization of timbre; from the compositional one, the attitude consists in the inference of rules and relations from the spectral analysis to the orchestration and sound manipulation level.

Since the origins of the movement, many works of spectral music have been realized at IRCAM, and have introduced compositional techniques shared by composers not directly involved in the group: in Jonathan Harvey's *Mortuos Plango Vivos Voco* (1980, for computer elaborated concrete sounds, see [**Harvey, 1981**, **Clarke, 2006**]), the sounds taken as analytical references are samples of the tenor bell of Winchester Cathedral, and the voice of the composer's child Dominic, who were a chorister in the same cathedral. Assisted in the technical realization by Stanley Haynes and Xavier Rodet, Harvey uses sound and voice synthesis together with manipulations of the original samples, controlled and articulated by criteria deduced from the spectral analysis: pitches, as well as the global structure and other parameters, are deduced from the analysis of a half-second fragment of the bell sample, right after the toll.

The peculiarity, in spectral music, is to establish relations between a sound analysis and a musical score: an approach which moves from sound to formal choices. But on a different level, the interest of time-frequency analysis and processing is to establish relations between sounds themselves: a sound representation makes it easier to define and work with classes of timbre, and to visualize sound components. Once defined a target sound or effect to realize, this knowledge is a useful tool for the orientation among the large range of processing and re-synthesis methods available. Several composers have developed a deep musical experience of these techniques; among them, the approach of Marco Stroppa gives a special outlook on the musical potential offered by a complete framework for sound analysis, manipulation and re-synthesis:

[1] see http://anasynth.ircam.fr/home/english/software/supervp

[2] see http://anasynth.ircam.fr/home/english/software/audiosculpt

I use SuperVP as an instrument, like I use musical instruments; I know which are the notes that can be played by a flute, as well as I know which sounds and transformations I can get with SuperVP: if I have a piano playing many superposed fast notes in the low register, I know that I will have a bad analysis with any fixed resolution, and therefore bad transformations.[3]

In his work *Zwielicht* (1994-99, for double bass, two percussions, electronics and 13-D sound projection), the whole electronic material is obtained with AudioSculpt, elaborating uncommon instrumental sounds, that can rarely be heard in concerts because they are too soft or too unpredictable to be reproduced: they appear when the edge of a crotale is gently scrubbed with a knitting needle, or when a double bass phrase is played using an extremely light, fast bow. The role of the elaboration is inspired by alchemy, the science of transforming matter: here, *matter* is any sounds coming from the various instruments, and the transformation is taking place in the electronic part and in the sound projection.

1.3. Sound signals

Complex-valued functions of a real variable having finite energy form a Hilbert space; this space is typically adopted as a model for physical quantities which change in time, as sounds, which are called *waveforms* or *signals*. Thus, a signal is denoted as $f(t) \in L^2(\mathbb{R})$, a potentially complex function of time with finite energy. This space is formally obtained by choosing $p = 2$ in the following definition.

Definition 1.3.1. The space of complex-valued functions of a real variable having finite norm $\|\cdot\|_p$ is indicated as $L^p(\mathbb{R})$,

$$L^p(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{C} \; s.t. \; \|f\|_p = \left(\int_{\mathbb{R}} |f(t)|^p \mathrm{d}t \right)^{\frac{1}{p}} < \infty \right\} . \tag{1.3.1}$$

In analogy with other wave fields, the energy in a unit time for a sound wave is the signal squared $|f(t)|^2$. Since power is the amount of work per unit time, it can be called *instantaneous power* or *energy density*. So the energy in the time interval Δt is given by $|f(t)|^2 \Delta t$, and the total energy E is $\int_{\mathbb{R}} |f(t)|^2 \mathrm{d}t = \|f\|_2^2$. Similarly, we can compute means with respect to time, defining the measure $\mathrm{d}T = |f(t)|^2 \mathrm{d}t$; therefore, the concepts of mean time $\langle t \rangle$ and standard deviation σ_t apply, the first in relation with the time of highest energy density, the second with the time spread of the energy around its mean time,

$$\langle t \rangle = \int_{\mathbb{R}} t \, \mathrm{d}T \; , \; \sigma_t^2 = \int_{\mathbb{R}} (t - \langle t \rangle)^2 \, \mathrm{d}T = \langle t^2 \rangle - \langle t \rangle^2 \; .$$

A frequency description of the signal leads to a deeper knowledge of its structure according to our perception. Through the Fourier transform and expansion, a signal is decomposed in terms of sinusoids of different frequencies. The Fourier transform or *spectrum* of a function $f(t)$ in $L^1(\mathbb{R})$ at the frequency ω is defined as follows,

$$\mathscr{F}(f)(\omega) = \widehat{f}(\omega) = \int_{\mathbb{R}} f(t) \, \mathrm{e}^{-2\pi \mathrm{i} \omega t} \, \mathrm{d}t \tag{1.3.2}$$

[3] The quoted sentences are taken from a private interview with Marco Stroppa (author's translation).

while the Fourier expansion of $f(t)$ is given by

$$f(t) = \int_{\mathbb{R}} \widehat{f}(\omega)\, \mathrm{e}^{2\pi \mathrm{i}\omega t}\, \mathrm{d}\omega \,. \tag{1.3.3}$$

This definition can be extended to signals as $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ is densely embedded in $L^2(\mathbb{R})$. Both the signal and the spectrum can be represented in the complex form,

$$f(t) = a(t)\, \mathrm{e}^{\mathrm{i}\phi(t)} \,,\ \widehat{f}(\omega) = b(\omega)\, \mathrm{e}^{\mathrm{i}\psi(\omega)} \,, \tag{1.3.4}$$

where $a(t)$ and $b(\omega)$ are the amplitude and spectral amplitude of the signal, while $\phi(t)$ and $\psi(\omega)$ are its phase and spectral phase. In analogy with the time domain we consider $|\widehat{f}(\omega)|^2$ as the *spectral energy density* per unit, and $|\widehat{f}(\omega)|^2 \Delta\omega$ is the spectral energy in the frequency interval $\Delta\omega$. From the Plancherel identity (see [Gröchenig, 2001b], Theorem 1.1.2), we have that the integral of the spectral energy density over all frequencies gives the total energy of the signal,

$$E = \int_{\mathbb{R}} |\widehat{f}(\omega)|^2 \mathrm{d}\omega = \int_{\mathbb{R}} |f(t)|^2 \mathrm{d}t \,; \tag{1.3.5}$$

finally, by writing $\mathrm{d}\Omega = |\widehat{f}(\omega)|^2 \mathrm{d}\omega$, the mean frequency $\langle\omega\rangle$ and standard deviation σ_ω^2 of the spectral density can be defined as well,

$$\langle\omega\rangle = \int_{\mathbb{R}} \omega \,\mathrm{d}\Omega \,,\ \sigma_\omega^2 = \int_{\mathbb{R}} (\omega - \langle\omega\rangle)^2 \,\mathrm{d}\Omega \,.$$

The simplest time-varying signal is the sinusoid, characterized by a constant amplitude a and a constant frequency ω,

$$f(t) = a \sin \omega t \,, \tag{1.3.6}$$

where the amplitude is the modulus of the minima and maxima of the oscillations, while the frequency is the number of oscillations per unit time. This representation can be extended to a larger class of signals, the functions whose time-varying amplitude and frequency is expressed in the following form,

$$f(t) = a(t) \cos \phi(t) \,, \tag{1.3.7}$$

where the instantaneous frequency is given by the first derivative of the phase function $\phi(t)$. As seen in Section 1.2, one of the first general model to be adopted is based on a representation of sounds as a sum of slowly time-varying functions of type (1.3.7): the decomposition of audio spectra in sinusoids is used to improve the results of signal manipulation algorithms. This model is not meaningful when sounds present sharp onsets, as well as significative inharmonic or noise components, which cannot be efficiently represented in such a form. An improved model in this sense is given by the *deterministic plus stochastic decomposition* introduced in [Serra and Smith, 1990, McAulay and Quatieri, 1986]: the signal is represented as a sum of time-varying sinusoids plus a noise component,

$$f(t) = \sum_{p=1}^{P} a_p(t) \cos\left(\phi_p(t)\right) + e(t) \,, \tag{1.3.8}$$

where $a_p(t)$ and $\phi_p(t)$ are the instantaneous amplitude and phase of the p-th sinusoid, while $e(t)$ is the noise component at time t. This representation can be deduced by the STFT of the signal f, as well as from other representations. Even if this model has been largely integrated with further strategies to deal with nonstationary components (see [**Roads et al., 1997**, **Röbel, 2003**] and the bibliography in the latter), it still constitutes a reference for a wide range of efficient high-quality sound processing and parameter estimation techniques. In this work, we aim to define representations which are well-suited for sinusoidal models, thus providing optimal readability and separation of their fundamental elements: sinusoidal components, transient attacks, noise.

1.4. Time-frequency representations and energy densities

Time-frequency representations (see [**Cohen, 1995**, **Cohen, 1989**, **Mallat, 1999**] for the theory and the motivations beyond this approach), briefly indicated as *TFR*, are employed for several different signals: sound, light, image, video, every kind of phenomenon which is interpretable as a function with finite energy on a real or complex space. The starting point of this prolific field is the work of the french mathematician and physicist Jean Baptiste Fourier, together with the improvements of computer science techniques for the fast application of models and tools stemming from his results. The first goal of a signal representation is to increase its readability: the spectrum of a sound is a fundamental characterization of its features in the frequency domain, but it is not enough to have a local complete information; if we consider a signal and its Fourier transform separately, we cannot observe the evolution of its spectral content over the time. With TFRs, a further characterization is provided, increasing the dimension of the representation domain: for a mono-dimensional signal, a TFR is a two-dimensional space which jointly describes its time and frequency content. In this section we investigate the relations between the physical and the probabilistic concepts of density; in the approach we adopt, this motivates the idea of the *spectrogram* (introduced in Section 1.5) as a time-frequency density and the use of certain mathematical tools to analyze its features.

The spectrogram is an example of a TFR defined from a decomposition of the signal within a set of elementary atoms. This strategy is largely employed in signal processing, because the information is distributed among different basic functions, which are easier to deal with: depending on the application, we can select only a few of them carrying the most information we are interested in (data compression), or define a transformation of the atoms which determines a transformation of the original signal. Therefore, the interest of the atoms resides in their capability to separate and make intelligible basic properties and components of the signal: onsets, noise, sinusoids or resonant structures for sounds; colors or edges for images.

1.4.1. Densities and distributions. To define a probability density we consider the real numbers $\mathbb{R}$ with the usual Borel σ-algebra and Lebesgue one-dimensional measure; a *probability density* on $\mathbb{R}$ is a positive continuous function f such that $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$. Given a probability density, a *probability law* based on such density is defined as a positive function μ which associates to every Borel set A the number

$$\mu(A) = \int_A f(x)\mathrm{d}x \ .$$

The *cumulative distribution* F of μ is defined from $\mathbb{R}$ in $[0,1]$, as follows

$$F(x) = \mu(]-\infty, x]) \ ;$$

it verifies the following identity,

$$\lim_{h\to 0^+} \frac{F(x+h)-F(x)}{h} = \lim_{h\to 0^+} \frac{\mu((x,x+h])}{h} \ ,$$

that defines the value of a probability density in a point as the derivative of its cumulative distribution. A first notational problem occurs: in signal processing applications the cumulative distribution is not often employed; on the other hand, the term *distribution* is always used as a synonym of density. In this work, the term *distribution* indicate the density, otherwise it will be specified. A second remark concerns the distinction between the density function and the probability law that it defines; as our interest is mainly focused on densities, many quantities and functions defined on the probability law will be straightly related to the density itself.
Defining a probability density on $\mathbb{R}^2$, and generally speaking in $\mathbb{R}^d$, requires nothing more of what has just been observed, considering the multi-dimensional Borel σ-algebra and Lebesgue measure. So let now f be a probability density in $\mathbb{R}^2$; thanks to the Fubini theorem the following identities hold,

$$f_1(x) = \int_{\mathbb{R}} f(x,y)\mathrm{d}y \ , \ f_2(y) = \int_{\mathbb{R}} f(x,y)\mathrm{d}x \ , \tag{1.4.1}$$

and the functions f_1, f_2 are one-dimensional densities, called *marginal densities* or *marginals*, while f is their *joint density*. If we consider now a TFR as a probability density on $\mathbb{R}^2$, and interpret the two dimensions as time and frequency, we would like to use the properties (1.4.1) to deduce the instantaneous energy and spectral density per unit: in the following, we will give examples of TFR for which properties (1.4.1) hold, and others, like the spectrogram (see Section 1.5), for which they do not hold.

1.4.2. Overview of some time-frequency transforms and distributions. After the introduction of several different TFRs with specific features, a first general approach is established by Cohen (see [**Cohen, 1989**, **Cohen, 1995**]): given a signal $f(t)$, the *Cohen's class* is composed by time-frequency representations C_f such that,

$$C_f(t,\omega) = \iint_{\mathbb{R}^2} A_f(\theta,\tau)\Phi(\theta,\tau)\,\mathrm{e}^{-2\pi\mathrm{i}(\theta t+\tau\omega)}\,\mathrm{d}\theta\mathrm{d}\tau \ ; \tag{1.4.2}$$

here, A_f is the *ambiguity function* of f, defined as follows,

$$A_f(\theta,\tau) = \int_{\mathbb{R}} f(t+\frac{\tau}{2}) f^*(t-\frac{\tau}{2})\, \mathrm{e}^{2\pi \mathrm{i}\theta t}\, \mathrm{d}t\ , \tag{1.4.3}$$

while Φ and the product $A\Phi$ are called the *kernel* and the *characteristic function* of the distribution, respectively. The bilinear distributions of Cohen's class can be seen as the 2-D Fourier transform of a weighted version of A_f, where the weighting function is Φ; in the same way, the characteristic function is obtained as the inverse 2-D Fourier transform of the distribution.

There exist TFRs (see [**Cohen, 1995**]) $C_f(t,\omega)$ whose marginal properties parallel those of probability densities (1.4.1),

$$\int_{\mathbb{R}} C_f(t,\omega)\mathrm{d}\omega = |f(t)|^2\ ,\ \int_{\mathbb{R}} C_f(t,\omega)\mathrm{d}t = |\widehat{f}(\omega)|^2\ , \tag{1.4.4}$$

$$\iint_{\mathbb{R}^2} C_f(t,\omega)\mathrm{d}t\mathrm{d}\omega = \int_{\mathbb{R}} |f(t)|^2 \mathrm{d}t =: \|f\|_2^2\ . \tag{1.4.5}$$

Consider, for instance, the *Wigner distribution*, defined as

$$W_f(t,\omega) = \int_{\mathbb{R}} f(t+\frac{\tau}{2}) f^*(t-\frac{\tau}{2})\, \mathrm{e}^{-2\pi \mathrm{i}\tau\omega}\, \mathrm{d}\tau\ . \tag{1.4.6}$$

The kernel of the Wigner distribution is the constant one function, and its characteristic function is A_f. This distribution verifies equations (1.4.4) and (1.4.5), making it possible to deduce exact time or frequency information about the signal from the joint representation. The disadvantage of the Wigner distribution is that the time-frequency distribution may reveal components that do not correspond to the analyzed signal, the so-called *cross components* (see Figure 1.1 for an example).

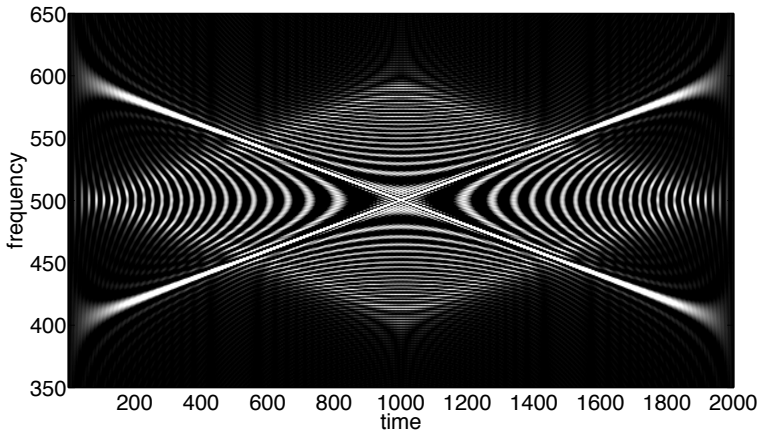

FIGURE 1.1. *Wigner distribution of a sum of complex sinusoids with linear frequency modulation (linear chirps): we would expect the representation to show only the two diagonals, the other white zones are interference components.*

The role of the kernel, in general distributions in the Cohen's class, is to smooth these interference components, giving a coherent representation of the signal. Considering a non constant kernel in the Wigner distribution, the so-called *Pseudo Wigner* distribution is obtained (see [**Martin and Flandrin, 1985**]), with better robustness to cross components, but which in general does not verify equations (1.4.4). Several different approaches have been considered for the design of appropriate kernels, leading to the conclusion that fixed kernels are in general well suited for limited class of signals; in Section 1.6 we mention some adaptive strategies in this sense.

According to the marginal analogy, as well as others, between certain TFRs and probability densities, it makes sense to investigate properties of TFRs by means of probabilistic tools: we are interested in particular to entropy measures (see [**Baraniuk et al., 2001**]). The Shannon entropy of a TFR C_f is given by the following integral, considering a unit-energy signal f,

$$H(C_f) := -\iint_{\mathbb{R}^2} C_f(t,\omega)\log_2 C_f(t,\omega)\mathrm{d}t\mathrm{d}\omega\ . \tag{1.4.7}$$

Rényi entropies (see [**Rényi, 1961**, **Beck and Schlögl, 1993**, **Zibulski and Zeevi, 1997**] for its definition and general properties, and Section 3.2), which are an extension of the Shannon one, can be calculated as well,

$$H_\alpha(C_f) := \frac{1}{1-\alpha}\log_2 \iint_{\mathbb{R}^2} (C_f(t,\omega))^\alpha \mathrm{d}t\mathrm{d}\omega\ , \tag{1.4.8}$$

given $\alpha > 0\ ,\ \alpha \neq 1$. These quantities are not defined for every TFR, in particular for some of those which are not positive. In Chapter 3, we detail how entropies can be used to measure the concentration of a spectrogram (see Section 1.5), considered as a TFR: this information can then be used to set an automatic adaptive framework for the analysis of a signal.

Like the spectrogram, several representations originated by time-frequency transforms belong to the Cohen's class; a typical way to define a linear time-frequency transform is to set a *dictionary* of functions $\{\phi_\gamma\}_{\gamma\in\Gamma}$ with a localized support, called *atoms*; then, for every function f the corresponding time-frequency transform T is defined as

$$Tf(\gamma) = \int_{\mathbb{R}^d} f(t)\overline{\phi_\gamma(t)}\mathrm{d}t. \tag{1.4.9}$$

As we are working with functions in the Hilbert space $L^2(\mathbb{R})$, we look to the case $d = 1$, and the integral in the right side of equation (1.4.9) can be written as $\langle f, \phi_\gamma\rangle$. From the Parseval's formula (see [**Gröchenig, 2001b**], Theorem 1.1.2) we know that

$$Tf(\gamma) = \int_{\mathbb{R}} f(t)\overline{\phi_\gamma(t)}\mathrm{d}t = \int_{\mathbb{R}} \widehat{f}(\omega)\overline{\widehat{\phi}_\gamma(\omega)}\mathrm{d}\omega\ , \tag{1.4.10}$$

so we have that if $\phi_\gamma(t)$ is null outside a time interval, then $\langle f, \phi_\gamma\rangle$ depends only on the values taken by f in that interval. Similarly, if $\widehat{\phi}_\gamma(\omega)$ is null outside a frequency interval, from the right side of (1.4.10) we have that $\langle f, \phi_\gamma\rangle$ depends only on the values taken by $\widehat{f}$ in that interval. As the target is to obtain strongly localized informations on f and $\widehat{f}$

simultaneously, we would like to narrow both the intervals at once. In what follows, we show the existing limit to the tightness we can get, that is to the information we can deduce about f, using the coefficients of a time-frequency transform.

As seen, the information earned on f by the product $\langle f, \phi_\gamma \rangle$ is related to a time-frequency region whose dimensions depend on $\phi_\gamma(t)$ and $\widehat{\phi_\gamma}(\omega)$. If we suppose $\int_{\mathbb{R}} |\phi_\gamma(t)|^2 \mathrm{d}t = 1 = \|\phi_\gamma\|_2$, then $|\phi_\gamma(t)|^2$ can be seen as an energy distribution on $\mathbb{R}$ whose *central time* τ_γ and spread around τ_γ are given, respectively, by the following average and variance:

$$\tau_\gamma = \int_{\mathbb{R}} t|\phi_\gamma(t)|^2 \mathrm{d}t \ , \ \sigma_t^2(\gamma) = \int_{\mathbb{R}} (t-\tau_\gamma)^2 |\phi_\gamma(t)|^2 \mathrm{d}t \ . \tag{1.4.11}$$

Similarly, the *central frequency* ω_γ and spread around ω_γ of $|\widehat{\phi_\gamma}|^2$ are given by

$$\omega_\gamma = \int_{\mathbb{R}} \omega|\widehat{\phi_\gamma}|^2 \mathrm{d}\omega \ , \ \sigma_\omega^2(\gamma) = \int_{\mathbb{R}} (\omega-\omega_\gamma)^2 |\widehat{\phi_\gamma}(\omega)|^2 \mathrm{d}\omega \ . \tag{1.4.12}$$

Thus we have that the information we can get on f through $\phi_\gamma(t)$ is concentrated in the so-called *Heisenberg box* associated to the atom (see Figure 1.2), that is a rectangle in the time-frequency plane, centered in $(\tau_\gamma, \omega_\gamma)$, whose time and frequency sides are $\sigma_t(\gamma)$ and $\sigma_\omega(\gamma)$, respectively.

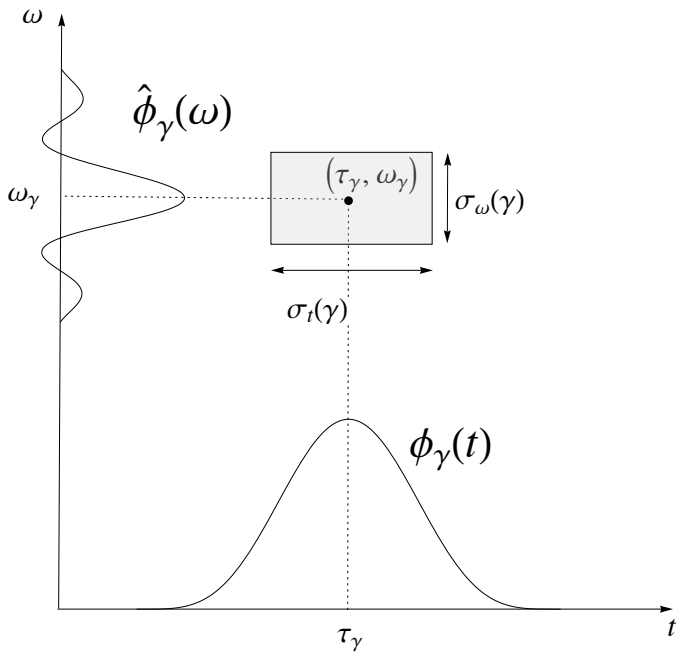

FIGURE 1.2. *Heisenberg box associated to the atom ϕ_γ centered in $(\tau_\gamma, \omega_\gamma)$, whose sides are given by the time and frequency spreads $\sigma_t(\gamma)$ and $\sigma_\omega(\gamma)$, respectively.*

Theorem 1.4.1 (Heisenberg principle). *Let f be a function in $L^2(\mathbb{R})$ and $\widehat{f}$ its Fourier transform. The temporal variance σ_t^2 and the frequency variance σ_ω^2 of f satisfy the inequality*

$$\sigma_t^2 \sigma_\omega^2 \geq \frac{1}{4}\,, \tag{1.4.13}$$

with equality if and only if there are $(u, \xi, a, b) \in \mathbb{R}^2 \times \mathbb{C}^2$, with $\Re b > 0$, such that

$$f(t) = a e^{i\xi t - b(t-u)^2}.$$

Theorem 1.4.1 is a statement of the *Heisenberg principle* (see [**Flandrin, 1999**] for other different ones); from the point of view of a signal analysis, it says that given an atom $\phi_\gamma(t)$, the quantities $\sigma_t^2(\gamma)$ and $\sigma_\omega^2(\gamma)$ limit the time and frequency precisions of the information we can earn about the signal from the product $\langle f, \phi_\gamma \rangle$. When the goal is to increase the analysis precision, Theorem 1.4.1 is often cited as the main tie: the more details we want to see in time, the coarser our resolution is bound to become in frequency. The fact, that we can not simultaneously achieve arbitrarily high time- and frequency-resolution of a particular signal component, is less troublesome, if we are able to vary the time-frequency resolution over the time-frequency plane. Some steps towards this idea have been made, as explained in Section 1.6. However, many problems, both theoretical and practical, remain open, and constitute one of the main interest of this work.

The *Short Time Fourier Transform (STFT)* is probably the most used dictionary-based time-frequency transform in Computer Music and Sound Processing: we introduce in more details the STFT and the related time-frequency distribution in Section 1.5, as it is the tool we focus on, throughout this work. Another example of dictionary-based transform is given by the *Wavelet Transform* (see [**Daubechies, 1992**, **Daubechies, 1990**, **Coifman et al., 1992**]), largely exploited in Image and Video Processing, and in Sound Processing, too. A *wavelet* is a function $\psi \in L^2(\mathbb{R})$ such that $\int_{\mathbb{R}} \psi(t) \mathrm{d}t = 0$, and we can always assume $\|\psi\|_2 = 1$. The dictionary $\{\psi_{a,b}\}_{(a,b)\in\mathbb{R}^+\times\mathbb{R}}$ is obtained here with a scaling and a time translation of factors a and b, respectively, given by

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right), \tag{1.4.14}$$

where ψ is called *mother wavelet*. The related time-frequency transform is thus defined as follows,

$$W f(a,b) = \langle f, \psi_{a,b} \rangle = \int_{\mathbb{R}} f(t) \frac{1}{\sqrt{a}} \overline{\psi}\left(\frac{t-b}{a}\right) \mathrm{d}t\,. \tag{1.4.15}$$

If the *admissibility condition* (see [**Mallat, 1999**]) is fulfilled, that is if

$$C_\psi = \int_{\mathbb{R}^+} \frac{|\widehat{\psi}(\omega)|^2}{\omega} \mathrm{d}\omega < +\infty\,, \tag{1.4.16}$$

then the Wavelet transform admits an inversion formula which allows to reconstruct f from the coefficients $W f(a,b)$.

Varying a and b the Heisenberg boxes related to the wavelets have different sides; if ψ and $\widehat{\psi}$ are centered in 0 and ξ, respectively, with time spread σ_t and frequency spread σ_ω, then the Heisenberg box associated to $\psi_{a,b}(t)$ is centered in $(b, \xi/a)$, with sides $a\sigma_t$ and σ_ω/a. Thus we see that different scalings of a wavelet determine different boxes, with the same surface: that is, the global resolution is not increasing, but the time or frequency resolutions can individually increase, to the detriment of the other. In particular, when a is large, the box is located in a low frequency range, and provides a lower frequency spread σ_ω/a than the mother wavelet, that is a higher frequency resolution; on the other hand, a small scaling factor a places the box in a high frequency range, with a better time resolution (see Figure 1.3).

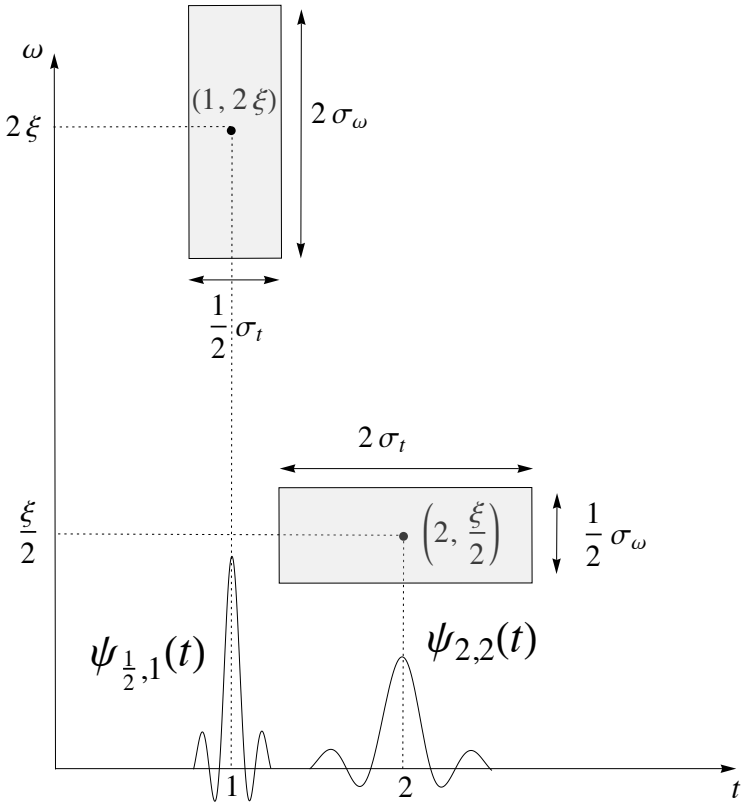

FIGURE 1.3. *Heisenberg boxes associated to two scaled wavelets, whose mother wavelet is centered in* $(0, \xi)$ *and has time and frequency spreads* σ_t *and* σ_ω*, respectively.*

Our auditory system perceives sounds in a similar way, with a lower capability to distinguish close tones when their frequencies grow. This characteristic is exploited, for example, in some audio coding techniques (see [**Painter and Spanias, 2000**] for a survey), which aim to reduce the digital size of an audio file for the purpose of efficient transmission or storage: the MP3 (MPEG-1 Audio Layer III) audio codec is one of the most popular. In most of these algorithms, a time-frequency transform of the signal

is calculated, and only a few of its coefficients are selected to be stored, according to specific criteria of perceptual relevance. The loss of information is thus controlled in order to obtain a reconstructed signal perceptually close to the original one. The varying-resolution property has shown to be well-suited for image processing, too: here, colors play the role of frequency, and typical problems are related to contour detection, preservation or reconstruction.

In this work, the purpose is not to reduce the size of the audio file through an appropriate representation: we aim, instead, to define representations which increase the readability of a sound, separating its elementary components, and making them easier to manipulate. For this reason, we look for a representation whose coefficients have a direct interpretation in terms of time-varying frequency spectrum, and we rather adopt the STFT transform (Section 1.5) as a starting point.

1.5. The spectrogram of a sound

The time-frequency transform we focus on is an extension of the classical Fourier transform, realized by first multiplying the signal by another function and then taking the Fourier transform.

Definition 1.5.1. Given a function $g \neq 0$, the *Short Time Fourier Transform (STFT)* of a function f is defined as

$$\mathcal{V}_g f(t,\omega) = \int_{\mathbb{R}^d} f(x)\overline{g(t-x)}\,\mathrm{e}^{-2\pi \mathrm{i}\omega x}\,\mathrm{d}x\ .$$

The function g is called *window function* and is used to localize the spectral information given by the transform. When dealing with signals, we consider $d = 1,\ f, g \in L^2(\mathbb{R})$ with $\|g\|_2 = 1$ and $g(t) = g(-t)$.
We introduce the time and frequency shifts operators,

$$T_x f(t) = f(t-x),\ M_\omega f(t) = \mathrm{e}^{2\pi \mathrm{i}\omega t}\, f(t)\ , \tag{1.5.1}$$

and consider time-frequency shifts of g as follows

$$M_\xi T_x g = \mathrm{e}^{2\pi \mathrm{i}\xi t}\, g(t-x)\ . \tag{1.5.2}$$

For every $(x,\xi) \in \mathbb{R}^2$ we have $\|M_\xi T_x g\| = 1$, and it is easy to see that the Heisenberg box associated to the atom $M_\xi T_x g$ is centered in (x,ξ) itself, with sides which do not depend on x and ξ, as

$$\begin{aligned} \sigma_t^2(x,\xi) &= \int_{\mathbb{R}} (t-x)^2 |g(t-x)|^2 \mathrm{d}t \\ &= \int_{\mathbb{R}} t^2 |g(t)|^2 \mathrm{d}t = \sigma_t^2(0,0) \end{aligned} \tag{1.5.3}$$

and

$$\begin{aligned} \sigma_\omega^2(x,\xi) &= \int_{\mathbb{R}} (\omega-\xi)^2 |\widehat{g}(\omega-\xi)\,\mathrm{e}^{-2\pi \mathrm{i}x(\omega-\xi)}\,|^2 \mathrm{d}\omega \\ &= \int_{\mathbb{R}} \omega^2 |\widehat{g}(\omega)|^2 \mathrm{d}t = \sigma_\omega^2(0,0)\ . \end{aligned} \tag{1.5.4}$$

STFT is the time frequency transform corresponding to the dictionary $\{M_\xi T_x g\}_{(x,\xi)\in\mathbb{R}^2}$: as we have seen, the Heisenberg box of an atom tells us the precision, in time and frequency, of the information we get about the signal, when taking its scalar product with that atom. We thus deduce that, unlike wavelets, in the STFT the precision of the original window and those of its shifted and modulated versions are the same.

The following result ([**Gröchenig, 2001b**], Corollary 3.2.3) provides an inversion formula for the STFT.

Theorem 1.5.2. *Given $g, h \in L^2(\mathbb{R}^d)$ such that $\langle g, h\rangle \neq 0$, for every $f \in L^2(\mathbb{R}^d)$ we have*

$$f(t) = \frac{1}{\langle g, h\rangle} \iint_{\mathbb{R}^{2d}} \mathcal{V}_g f(x,\xi) M_\xi T_x h \, \mathrm{d}\xi \mathrm{d}x \, . \tag{1.5.5}$$

When working with signals, it states that every signal $f \in L^2(\mathbb{R})$ can be reconstructed from its STFT; moreover, it shows that the window used for the reconstruction may be different from the one used for the analysis, and nearly every synthesis window is possible to still get perfect reconstruction.

As usual with time-frequency transforms, we can deduce a TFR taking the squared modulus of the STFT; we obtain this way the *spectrogram* of the signal, that we indicate as

$$\mathrm{PS}_g f(t,\omega) = |\mathcal{V}_g f(t,\omega)|^2 \, , \tag{1.5.6}$$

omitting the indication of the window function g if no ambiguity occurs. The spectrogram can be seen as a surface with finite energy, as $\mathrm{PS}f \in L^2(\mathbb{R}^2)$; the following proposition (see [**Mallat, 1999**]) shows that, on the other hand, not every $\Phi \in L^2(\mathbb{R}^2)$ is the spectrogram of a signal $f \in L^2(\mathbb{R})$.

Proposition 1.5.3. *Let $\Phi \in L^2(\mathbb{R}^2)$. There exists $f \in L^2(\mathbb{R})$ such that $\Phi(t,\omega) = \mathrm{PS}f(t,\omega)$ if and only if*

$$\Phi(t,\omega) = \iint_{\mathbb{R}^2} \Phi(x,\xi) K(t,x,\omega,\xi) \mathrm{d}x \mathrm{d}\xi \, , \tag{1.5.7}$$

where

$$K(t,x,\omega,\xi) = \langle M_\omega T_t g, M_\xi T_x g\rangle \, . \tag{1.5.8}$$

The function K is called *reproducing kernel* and it measures the time-frequency overlap of the two atoms $M_\omega T_t g$ and $M_\xi T_x g$. Its amplitude decays with $t-x$ and $\omega-\xi$ at a rate that depends on the energy concentration of g and $\widehat{g}$. The characteristic function $M_{\mathrm{PS}f}$ of a spectrogram $\mathrm{PS}f$ is given by

$$\begin{aligned} M_{\mathrm{PS}f}(t,\omega) &= \iint_{\mathbb{R}^2} \mathrm{PS}f(\tau,\theta) \, \mathrm{e}^{2\pi\mathrm{i}(\tau t+\theta\omega)} \, \mathrm{d}\tau \mathrm{d}\theta \\ &= A_f(t,\omega) A_g(-t,\omega) \, , \end{aligned} \tag{1.5.9}$$

where A_f and A_g are the ambiguity functions of the signal and the window function, respectively. Therefore, we deduce that the kernel of the spectrogram is the ambiguity function of the window function; we can also deduce, by evaluating $M_{\mathrm{PS}f}$ in $(t,0)$ and

$(0, \omega)$, that if we consider the spectrogram as a joint distribution, the marginals are in general different from $|f(t)|^2$ and $|\widehat{f}(\omega)|^2$, and equations (1.4.4) do not hold; this is due to the contribution of the window spectrum to the spectrum of the windowed signal; and it shows that the information we can get from the spectrogram, about the energy and the spectral energy of f, is necessarily altered by the window function.

1.5.1. The role of the window function. As we have seen, a window function g identifies a corresponding rectangular area in the time-frequency plane, its Heisenberg box; the dimensions of the box are determined by the time and frequency variance of the window, and the minimum area is fixed by the Theorem 1.4.1, when g is a Gaussian. Depending on the desired resolution of the analysis, the ratio between the box sides can be changed. Let g be a function whose energy and spectral energy have variance σ_t and σ_ω, respectively. If we consider a scaling of g with a factor $s \in \mathbb{R}^+$,

$$g_s(t) = \frac{1}{\sqrt{s}} g\left(\frac{t}{s}\right) , \tag{1.5.10}$$

we obtain an atom whose time and frequency spread are $s\sigma_t$ and $\frac{\sigma_\omega}{s}$, respectively. As seen for wavelets, the area of the Heisenberg box associated to the scaled window g_s is the same as the one corresponding to g, but the sides have changed: the amount of information we get from an STFT taken with g_s or g is the same, but the time and frequency precisions, considered separately, are different.

We now describe some features of the Fourier transform $\widehat{g}$ of a window function, whose proofs are also given for completeness. The following classical theorem states that they cannot both have a compact support.

Theorem 1.5.4. *If $g \in L^1(\mathbb{R})$ is a not identically null function, then g and $\widehat{g}$ cannot both have compact support.*

PROOF. We prove the statement by contradiction. Suppose that $\widehat{g}$ has a compact support $[-\omega_0, \omega_0]$ and that g is null over the whole interval $[t_1, t_2]$, which is a consequence of assuming that g has compact support. Then,

$$g(t) = \int_{-\omega_0}^{\omega_0} \widehat{g}(\omega)\, \mathrm{e}^{2\pi \mathrm{i}\omega t}\, \mathrm{d}\omega \; ; \tag{1.5.11}$$

consider $t_0 \in (t_1, t_2)$; by differentiating n times under the integral, we have

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} g(t_0) = \int_{-\omega_0}^{\omega_0} \widehat{g}(\omega)(2\pi \mathrm{i}\omega)^n\, \mathrm{e}^{2\pi \mathrm{i}\omega t_0}\, \mathrm{d}\omega = 0 \, . \tag{1.5.12}$$

Since

$$g(t) = \int_{-\omega_0}^{\omega_0} \widehat{g}(\omega)\, \mathrm{e}^{2\pi \mathrm{i}\omega(t-t_0)}\, \mathrm{e}^{2\pi \mathrm{i}\omega t_0}\, \mathrm{d}\omega \, , \tag{1.5.13}$$

developing $\mathrm{e}^{2\pi \mathrm{i}\omega(t-t_0)}$ in t_0 we have, for every $t \in \mathbb{R}$,

$$g(t) = \sum_{n=0}^{\infty} \frac{[2\pi \mathrm{i}(t-t_0)]^n}{n!} \cdot \int_{-\omega_0}^{\omega_0} \widehat{g}(\omega)\omega^n\, \mathrm{e}^{2\pi \mathrm{i}\omega t_0}\, \mathrm{d}\omega = 0 \tag{1.5.14}$$

contradicting the assumption that g is not identically null.

□

In digital signal processing applications, the window function has a compact support by necessity: in this case, $\widehat{g}$ has an unlimited support; by the definition of Fourier transform, considering the absolute value, we see that if g is positive, then $|\widehat{g}(\omega)|$ has an absolute maximum in $\omega = 0$; moreover, as g is real, $\widehat{g}$ is symmetric. It is an oscillating function (its first derivative has an infinite number of zeros), which decays to zero as $|\omega|$ goes towards infinity, at a rate which is studied in the next proposition.

Proposition 1.5.5. *Consider $g \in L^1(\mathbb{R})$; then g is p times continuously differentiable, $g \in \mathscr{C}^p(\mathbb{R})$, $p \geq 1$ if the following inequality holds,*

$$\int_{\mathbb{R}} |\widehat{g}(\omega)|(1 + |\omega|^p)\mathrm{d}\omega < +\infty \tag{1.5.15}$$

PROOF. By the assumption on g we have that $\widehat{g}$ is defined, and

$$\begin{aligned} |g(t)| &= \left| \int_{\mathbb{R}} \widehat{g}(\omega)\, \mathrm{e}^{2\pi \mathrm{i}\omega t}\, \mathrm{d}\omega \right| \\ &\leq \int_{\mathbb{R}} |\widehat{g}(\omega)| \mathrm{d}\omega < \infty \, ; \end{aligned} \tag{1.5.16}$$

by the expression of the Fourier transform of the n-th derivative of a function,

$$\widehat{g^{(n)}}(\omega) = (2\pi)^n \omega^n \widehat{g}(\omega) \, , \tag{1.5.17}$$

we obtain

$$|g^{(n)}(t)| \leq (2\pi)^n \cdot \int_{\mathbb{R}} |\widehat{g}(\omega)||\omega|^n \mathrm{d}\omega < \infty \, , \tag{1.5.18}$$

and the last inequality holds for every $n \leq p$ as a consequence of the hypothesis (1.5.15).

□

Remark 1.5.6. The result implies that a sufficient condition for g to be in $\mathscr{C}^p(\mathbb{R})$ is that there exist constants K and $\epsilon > 0$ such that

$$|\widehat{g}(\omega)| \leq \frac{K}{1 + |\omega|^{p+1+\epsilon}} \, . \tag{1.5.19}$$

So the regularity of g depends on the decay of $\widehat{g}(\omega)$ at infinity. As an example, if $r(t) = \mathbf{1}_{[-T,T]}$ the rectangular window, then $r(t)$ is discontinuous in T and $-T$, and $|\widehat{r}(\omega)|$ decays like $|\omega|^{-1}$ when ω tends to infinity. The proposition still holds taking $\widehat{g}$ and the inverse Fourier transform, so the regularity of $\widehat{g}$ depends on the decay of g at infinity. ■

We assume now to work with windows $g \in L^2(\mathbb{R})$ with compact support; the characteristics of $\widehat{g}$ tell us which is the biasing of the signal spectral information introduced by the window. As seen, $\omega = 0$ is the value where $|\widehat{g}(\omega)|$ has its absolute maximum; the value $\widehat{g}(0)$ identifies the so called *main lobe* of the window, and a main peak in the spectrogram. Beside the absolute maximum, $|\widehat{g}(\omega)|$ has an infinite number of local maxima, which determine further peaks in the spectrogram on both sides of the central

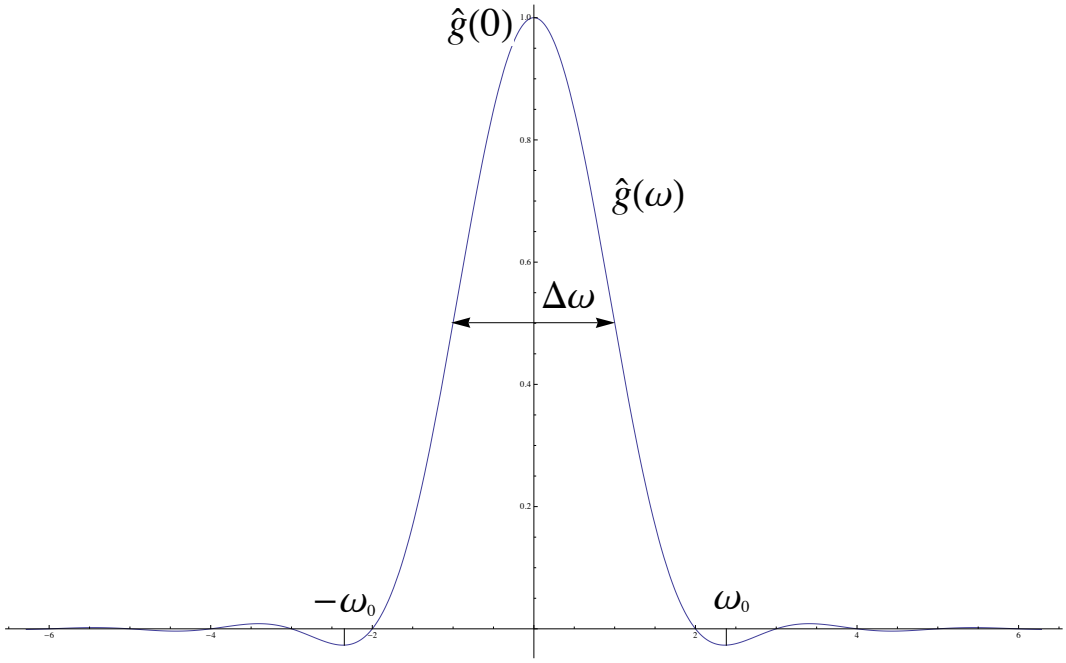

FIGURE 1.4. *Fourier transform $\widehat{g}(\omega)$ of a Hanning window with compact support, $g(t) = 2 \cdot \cos^2(\pi t)\chi_{[-\frac{1}{2},\frac{1}{2}]}$: the amplitude of the main lobe is $\widehat{g}(0)$, that of the side lobes is $A = |\widehat{g}(\omega_0)| = |\widehat{g}(-\omega_0)|$, the root mean-square bandwidth is $\Delta\omega$.*

frequency of the window (see Figure 1.4): they can be seen as a delocalization of the spectral information, and if their amplitude is considerable they become hard to distinguish from main peaks. It is thus desirable that the ratio between the amplitudes of the absolute maximum and those of local maxima is small. To measure this quantity, as $|\widehat{g}(\omega)|$ decays with oscillations, we can consider the two frequencies $\pm\omega_0$ where the first local maxima are reached: they are called *side lobes*, and the ratio of their amplitude with the one at the main lobe is measured in decibels,

$$A = 10 \log_{10} \frac{|\widehat{g}(\omega_0)|^2}{|\widehat{g}(0)|^2} \,. \tag{1.5.20}$$

Together with a small A, a measure of the spectral localization provided by the window is given by the amount of energy concentrated within the main lobe. The *root mean-square bandwidth* $\Delta\omega$ is defined by

$$\frac{|\widehat{g}(\Delta\omega/2)|^2}{|\widehat{g}(0)|^2} = \frac{1}{2} \,, \tag{1.5.21}$$

and measure the width of the mainlobe (see Figure 1.4). A detailed comparison of these features for different windows is out of the scope of this work (see [**Mallat, 1999**] Section 4.2.2 for a comparison of the quantities considered here, and [**Harris, 1978**] for

a complete survey in the discrete case): for the tests in Chapter 4, we choose a fixed window function often adopted in the applications, the *Hanning window* with compact support $[-\frac{1}{2}, \frac{1}{2}]$,

$$g(t) = \cos^2(\pi t)\chi_{[-\frac{1}{2},\frac{1}{2}]} ;$$

it provides a good localization of the spectral energy, as $A = -32$dB and $\Delta(\omega) = 1.44$, considering that $\widehat{g}(0) = \frac{1}{2}$ (see Figure 1.4, where $2\widehat{g}(\omega)$ is plotted, to enhance the lobes).

1.6. Adaptive time-frequency representations

STFT is often adopted in the sound processing domain as it closely reflects the concept of time-varying spectrum: its coefficients give the local amplitudes and phases of sinusoids, with a direct interpretation in terms of the sound components they refer to. But as we have seen in Section 1.5, it is a transform with constant resolution over the whole time-frequency plane. This is a limit, as the precision needed to separate the information coming from different components of complex sounds may vary significantly.
As a basic example, we can consider a percussion sample with fast sequences of transients, that we may want to fit for a different tempo than the original one; if the support of the analysis window used is too large, it is possible that a given time-shift includes several transients. From the analysis point of view, these components are indivisible, which means that every treatment concerning their analysis frame applies to them all: in the case of a time-stretch of the original sample, this is particularly inappropriate, because it makes impossible to situate different transients independently.
A symmetric basic example is the case where a small analysis window is used with instruments having close partials: as the frequency resolution of a function with short compact support is low, the value of a frequency bin in the analysis may be influenced by different partials, thus degrading the accuracy of spectral processing techniques like a pitch-shift.

The concept of adaptivity is related to the the intrinsic flexibility of the model, allowing to conceive set of atoms and operators which fulfill certain desired characteristics. In particular, we look for methods that provide a flexible choice of the local time-frequency resolution. As we have seen in Subsection 1.4.2, the classical wavelet transform cannot be considered adaptive in the sense just mentioned, because the resolution varies according to a fixed rule.

The limits about the fixed resolution of standard analysis methods have been overcome following different approaches. We consider in particular the ones related to Gabor Frame theory (Chapter 2), as this is the context where this work is included; from the point of view of adaptive kernel design (see Subection 1.4.2), we refer to [**Jones and Baraniuk, 1994**, **Jones and Baraniuk, 1995**] and the related bibliographies.

There are three main aspects we consider: first, the adaptivity as the possibility to deal with different resolutions locally within a sound; then, a criterium to choose the best local resolution which provides the adapted representation; and finally, the

possibility to define a reconstruction method from the adapted analysis. The idea of gathering a sparsity measure from information measures, and Rényi entropies in particular, is detailed in [Baraniuk et al., 2001]. In [Jaillet, 2005, Jaillet and Torrésani, 2007] a local time-frequency adaptive framework is presented exploiting this concept: automatic local adaptation and reconstruction are both developed, the latter being realized through a recursive algorithm whose general convergence is not investigated.

The definition of *multiple* Gabor frames, which is comprehensively treated in [Dörfler, 2002], provides Gabor frames with analysis techniques with multiple resolutions; an approach where sparse analyses are obtained through a regression model is introduced in [Wolfe et al., 2001]. The *nonstationary* Gabor frames (see [Jaillet et al., 2009, Balazs et al., 2011, Søndergaard et al.,] for their definition and implementation) are a further development in this sense; they fully exploit theoretical properties of the analysis and synthesis operator, and extend the *painless case* introduced in [Daubechies et al., 1986]: if the analysis respect certain conditions, they provide a class of FFT-based algorithms for analysis adaptation, in the time or frequency dimension separately, together with perfect reconstruction formulas.
The technique developed in [Rudoy et al., 2010] belongs to this same class but presents several novelties in the construction of the Gabor multi-frame, and in the method for automatic local time-adaptation. In [Lukin and Todd, 2006] a time-frequency adaptive spectrogram is defined considering a sparsity measure called *energy smearing*, without taking into account the re-synthesis task.
The concept of *quilted frame*, recently introduced in [Dörfler, 2011], is a promising effort to establish a unified mathematical model for all the various frameworks cited above.

1.7. Contributions of this work to the state of the art

We detail here the main contributions of this work, concerning the three aspects of adaptation, automatic choice of the best resolution, and reconstruction from adapted analyses. For the first two points, the strategy we adopt is the same as the one in [Jaillet, 2005, Jaillet and Torrésani, 2007], giving new results on two main subjects:

- in Section 3.3 we give new results on the existence of Rényi entropy measures of spectrograms in the continuous case, thus extending the results of [Baraniuk et al., 2001]; in the same section, we give new results about the convergence of discrete versions of these measures to their continuous one, when the sampling grid becomes infinitely dense;
- in Sections 3.4 and 3.6, we deduce some properties about the Rényi entropies and the parameter they depend on, which are useful for the interpretation of this parameter in applicative contexts; we define a novel method for spectral change detection based on these properties, as well as a particular normalization of the Rényi entropy detailed in Section 4.1, which is appropriate for the comparison of the entropy of discrete finite TFRs with different dimensions.

Concerning the reconstruction from adapted analyses, in Chapter 2 we focus on methods allowing for FFT-based implementations, dividing the cases where they can give perfect reconstruction or not:

- if the resolution of the adapted analyses changes as a function of time only, nonstationary Gabor frames are used, guaranteeing perfect reconstruction;
- if the resolution of the adapted analyses changes depending on time **and** frequency, we define in Sections 2.5 and 2.6 two new reconstruction methods giving an approximation of the original signal: we analyze the reconstruction error they give, by means of some tests, and provide a theoretical bound of the error for the second one, the *filter bank* method.

We have implemented new Matlab code for the whole framework of analysis, automatic adaptation and reconstruction; the different FFT-based reconstruction functions, which vary depending on the time or time-frequency adaptation, are new extensions of the existing ones (see [**Balazs et al., 2011**, **Søndergaard et al.,**]).

CHAPTER 2

Frame theory in sound analysis and synthesis

In Mathematics, Time-frequency Analysis is a branch of Harmonic Analysis that characterizes functions and operators considering the structure of their translations and modulations, that is time-frequency shifts. In the first decades of the last century, it has originally been formulated in the field of quantum mechanics, while the work of Dennis Gabor established its theoretical foundations in information theory and signal analysis, some years later (see [**Gabor, 1946**] for the original formulation by Gabor, and [**Gröchenig, 2001b**] for a survey about the origins of time-frequency analysis).

Typical problems of time-frequency signal processing, and in particular sound processing and computer music, can be modeled in a formal mathematical framework. Given a set of atomic functions in a Hilbert space, the related decomposition operator is called *Analysis* operator, while an expansion one is the *Synthesis* operator. They are the basic tools for a complete scheme for the analysis, transformation, and re-synthesis of a sound, which can be sketched as follows:

(1) a representation is obtained decomposing the sound by means of a given set of atoms, the result being a set of *analysis coefficients*;
(2) the analysis coefficients are interpreted to deduce information about the original sound;
(3) the analysis coefficients are modified to transform specific features of the representation;
(4) a new sound is constructed as an expansion of the modified coefficients within a certain set of atoms, not necessarily the same used for the analysis.

The four points of the scheme concern several different applications: sound visualization processes deal just with the first one, while feature extraction techniques exploit the first two; more complicated processes, such as source separation or vocal transformation, have to handle them all. One of the principal focus of our research is making the scheme adaptive: the analysis and synthesis operators have to change according to the characteristics of the signal.

In Chapter 1, we introduced several time-frequency representations with their characteristics for sound analysis and reconstruction. We did it for the continuous version of these TFRs, while real-world applications have to deal with discrete finite TFRs. Frame theory (see [**Gröchenig, 2001b**, **Christensen, 2003**, **Casazza, 1999**] for the general theory) is a general theoretical approach to the discretization of TFRs, including both Wavelet and Short Time Fourier Transforms (see [**Mallat, 1999**] for a comprehensive survey of theory and applications): it investigates, in particular,

the conditions for the sampled analysis and synthesis operators to preserve perfect reconstruction of the original signal.

In this chapter, we focus on Gabor frames; in the Gabor transform (see [**Gabor, 1946**] for the original article), the analysis atoms are obtained with time-frequency shifts of a Gaussian function, which provides for an optimal time-frequency localization (see Theorem 1.4.1); along with this approach, the STFT (Section 1.5) is the continuos version of the Gabor transform, with a generic symmetric window function. In the nineties of the last century, a different approach has led to the Wavelet Transform (Subsection 1.4.2), where a new paradigm is introduced; symmetric windows are replaced by wavelets, and the related analyses are not expressed within the time-frequency plane: the related two-dimensional space is indicated as *time-scale* (see [**Flandrin, 1999**] and [**Mallat, 1999**] for a comprehensive review of the two models). As detailed in Subsection 1.4.2, the main difference lays in the different resolution offered by the two bases: while the elements of a stationary Gabor frame have all the same time-frequency concentration, the different atoms in a wavelet set vary their concentration depending on their position in the time-scale space.

The first and fundamental objective of this thesis is the formal definition of mathematical models whose interpretation leads to theoretical and algorithmic methods for adaptive analysis. Such models have to take into account the necessity of reconstructing the original signal from the analysis coefficients, thus the problem of re-synthesis. We deal with two principal cases, both dealing with compactly supported analysis atoms: when the atoms change depending on their time location, with nonstationary Gabor frames it is possible to define efficient reconstruction methods giving a perfect reconstruction of the original signal (see Subsection 2.2.1); when the atoms change depending on both their time and frequency locations, the reconstruction cannot in general be made with efficient procedures, as it requires the inversion of an operator which may not have any regular structure. For this case, we define in Sections 2.4, 2.5 and 2.6 two new approximation methods extending the approach adopted for time adaptation. In Section 4.4 we measure the reconstruction error by means of several applications, while in Section 2.6 we give theoretical bounds for the error performed by the method that we indicate as *filter bank approach*.

2.1. Frame theory: basic definitions and results

The Fourier representation of a signal is based on sinusoids: these functions have an unbounded time support, which is not well-suited when we are interested in the local behavior of the signal. Moving from this drawback of the classical Fourier transform, frame theory enlarges the possible choices of bases and decomposing systems in a Hilbert space. Here we summarize the basic definitions and theorems, which are useful to the introduction of Gabor frames with multiple resolutions (see [**Dörfler, 2002**] Chapter 3 for a comprehensive survey).
Given a separable Hilbert space H, with its structure of vector space on $\mathbb{C}$ and its own scalar product $\langle \cdot, \cdot \rangle$, we consider a set of vectors $\{\phi_\gamma\}_{\gamma \in \Gamma}$ in H, where the index set Γ is countable and may be infinite, and γ can also be a multi-index.

Definition 2.1.1. The sequence $\{\phi_\gamma\}_{\gamma\in\Gamma}$ is a *frame* for H if there exist two positive non zero constants $A > 0$ and $B < \infty$, called *frame bounds*, such that for all $f \in H$,

$$A\|f\|^2 \leq \sum_{\gamma\in\Gamma} |\langle f, \phi_\gamma\rangle|^2 \leq B\|f\|^2 \,. \tag{2.1.1}$$

If $A = B$ the frame is *tight*; moreover, this definition includes orthonormal basis, as with $A = B = 1$ equation (2.1.1) is the Plancherel identity (see [Gröchenig, 2001b], Theorem 1.1.2). The frame bounds A and B are the infimum and supremum, respectively, of the eigenvalues of the *frame operator* S, defined as follows.

Definition 2.1.2. Given a set $\{\phi_\gamma\}_{\gamma\in\Gamma}$, the *analysis operator* C_ϕ is given by

$$\mathrm{C}_\phi f = \{\langle f, \phi_\gamma\rangle, \gamma \in \Gamma\} \,, \tag{2.1.2}$$

while for every sequence $c = (c_k)_{k\in\Gamma}$, the *synthesis operator* D_ϕ is defined as

$$\mathrm{D}_\phi c = \sum_{\gamma\in\Gamma} c_\gamma \phi_\gamma \,, \tag{2.1.3}$$

and for every $f \in H$ the *frame operator* S is given by

$$\mathrm{S}f = \sum_{\gamma\in\Gamma} \langle f, \phi_\gamma\rangle \phi_\gamma \,. \tag{2.1.4}$$

The synthesis operator is the adjoint of the analysis one, $\mathrm{D}_\phi = \mathrm{C}_\phi{}^*$, and the frame operator $\mathrm{S} = \mathrm{C}_\phi{}^*\mathrm{C}_\phi = \mathrm{D}_\phi\mathrm{D}_\phi{}^*$ is a positive invertible operator. For any frame $\{\phi_\gamma\}_{\gamma\in\Gamma}$ there exist dual frames $\{\widetilde{\phi}_\gamma\}_{\gamma\in\Gamma}$, such that for all $f \in H$ we have

$$f = \mathrm{D}_{\widetilde{\phi}}(\mathrm{C}_\phi f), \, = \mathrm{D}_\phi(\mathrm{C}_{\widetilde{\phi}} f) \,, \tag{2.1.5}$$

so that given a frame it is always possible to perfectly reconstruct a signal f using the coefficients of its decomposition through the frame. The inverse of the frame operator allows the calculation of the *canonical* dual frame, given by

$$\widetilde{\phi}_\gamma = \mathrm{S}^{-1}\phi_\gamma \tag{2.1.6}$$

which provides the minimal-norm analysis coefficients, in the ℓ^2-sense. The frame operator for the frame $\{\widetilde{\phi}_\gamma\}_{\gamma\in\Gamma}$ is S^{-1}.

We are interested in the case $H = L^2(\mathbb{R})$, as it represents the standard situation where a signal f is decomposed through a countable dictionary of atomic functions $\{\phi_k\}_{k\in\mathbb{Z}}$. In particular, a *Gabor system* is obtained by time-shifting and frequency-transposing a real window function g, such that $g(t) = g(-t)$ and $\|g\|_2 = 1$, according to a regular lattice $\Lambda = a\mathbb{Z} \times b\mathbb{Z}$. We say that the Gabor system $\mathbf{G}(g, a, b)$ is a *Gabor frame* if it satisfies the frame condition (2.1.1). We will also indicate such a frame as *stationary*, since the window used for the time-frequency shifts does not change. For a Gabor frame, the reconstruction formula (2.1.5) takes the following form,

$$f = \sum_{(l,k)\in\mathbb{Z}^2} \langle f, M_{bk}T_{al}g\rangle M_{bk}T_{al}\widetilde{g} = \sum_{(l,k)\in\mathbb{Z}^2} \mathcal{V}_g f(al, bk)\; M_{bk}T_{al}\widetilde{g} \,, \tag{2.1.7}$$

thus we see that using a Gabor frame $\mathbf{G}(g, a, b)$ we are able to perfectly reconstruct the signal f from a discrete sampling of its STFT with window g, according to the

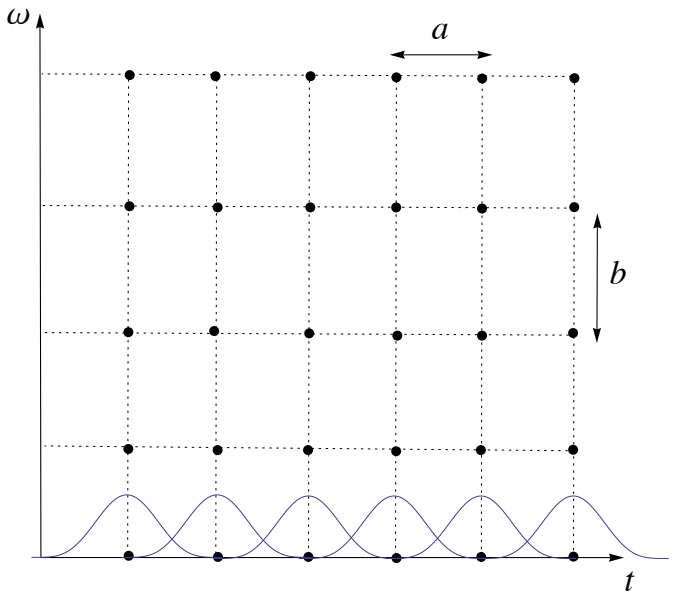

FIGURE 2.1. *Time-frequency centers for a stationary Gabor frame* $\mathbf{G}(g, a, b)$ *with time and frequency steps* a *and* b*, respectively.*

nodes of the lattice Λ: indeed, given a time step a and a frequency step b, the sequence (al, bk) with $(l, k) \in \mathbb{Z}^2$ generates the nodes of the time-frequency lattice Λ for the frame $\{g_{k,l}\}_{(k,l)\in\mathbb{Z}^2}$ defined as

$$g_{k,l}(t) = M_{bk}T_{al}g = g(t - al)\,\mathrm{e}^{2\pi ibkt}\,, \tag{2.1.8}$$

where the nodes are the centers of the Heisenberg boxes associated to the windows in the frame (Figure 2.1). We use in what follows a compact form of the reconstruction formula (2.1.7), to focus on the analysis and synthesis operator,

$$f = \mathrm{D}_{\tilde{\mathrm{g}}}(\mathrm{C}_{\mathrm{g}}f)\,, \tag{2.1.9}$$

specifying the lattice when ambiguity occurs.

Given a window g, the lattice has to satisfy certain conditions for $\{g_{k,l}\}$ to be a frame; the basic principle, that will be further analyzed in Section 3.3, is that if g is sufficiently regular, then $\mathbf{G}(g, a, b)$ is a Gabor frame as long as a and b are small enough: that is, if the sampling is sufficiently dense. Theorem 2.1.3 and 2.1.5 (see [**Daubechies, 1992**, **Daubechies, 1990**]), which follow, give necessary and sufficient conditions on a and b for $\mathbf{G}(g, a, b)$ to be a frame for $L^2(\mathbb{R})$.

Theorem 2.1.3. *The Gabor system* $\mathbf{G}(g, a, b)$ *is a frame for* $L^2(\mathbb{R})$ *only if* $ab \leq 1$*. The frame bounds* A *and* B *necessarily verify the following inequalities*

$$A \leq \frac{1}{ab} \leq B, \tag{2.1.10}$$

$$\forall t \in \mathbb{R},\ A \leq \frac{1}{b}\sum_{l\in\mathbb{Z}} |g(t-al)|^2 \leq B, \tag{2.1.11}$$

$$\forall \omega \in \mathbb{R},\ A \leq \frac{1}{a}\sum_{k\in\mathbb{Z}} |\widehat{g}(\omega-bk)|^2 \leq B, \tag{2.1.12}$$

thus we have that a Gabor frame $\mathbf{G}(g,a,b)$ is an orthonormal system for $L^2(\mathbb{R})$ only if $ab = 1$. The following theorem (see [**Daubechies, 1990**]) further characterize this particular case, showing that, in these hypotheses, g cannot provide a good time-frequency localization of the associated transform.

Theorem 2.1.4 (Balian - Low - Coifman - Semmes)**.** *If $\mathbf{G}(g,a,\frac{1}{a})$ is a Gabor frame, then either $tg \notin L^2(\mathbb{R})$, or $\omega\widehat{g} \notin L^2(\mathbb{R})$.*

We show now sufficient conditions for $\mathbf{G}(g,a,b)$ to be a frame for $L^2(\mathbb{R})$.

Theorem 2.1.5. *Define*

$$\beta(u) = \sup_{0\leq t\leq a}\sum_{l\in\mathbb{Z}} |g(t-al)||g(t-al+u)|\,,$$

$$\Delta(b) = \sum_{\substack{k\,\in\,\mathbb{Z}\\ k\,\neq\,0}} \left[\beta\left(\frac{k}{b}\right)\beta\left(-\frac{k}{b}\right)\right]^{1/2}.$$

If a and b are such that

$$A_0 = \frac{1}{b}\left(\inf_{0\leq t\leq a}\sum_{l\in\mathbb{Z}} |g(t-al)|^2 - \Delta(b)\right) > 0$$

and

$$B_0 = \frac{1}{b}\left(\sup_{0\leq t\leq a}\sum_{l\in\mathbb{Z}} |g(t-al)|^2 + \Delta(b)\right) < +\infty\,,$$

then $\mathbf{G}(g,a,b)$ is a frame for $L^2(\mathbb{R})$; A_0 is the upper bound for the lower frame bound A, and B_0 is the lower bound for the upper frame bound B.

Remark 2.1.6. Theorems 2.1.3 and 2.1.5 are classical density results proved by Ingrid Daubechies; several other conditions have been given later, for a Gabor system to be a frame (see [**Gröchenig, 2001b**]); in particular, if φ is a Gaussian window, a stronger result exists (see Theorem 7.5.3 in the previous reference) proving that a necessary and sufficient condition for $\mathbf{G}(\varphi,a,b)$ to be a frame is that $ab < 1$. ■

In some particular cases, which are often adopted in standard applications, the frame operator takes the form of a multiplication, as stated in the following theorem (see [**Daubechies et al., 1986**] for the original formulation).

Theorem 2.1.7. *Consider $g \in L^\infty(\mathbb{R})$ with $\mathrm{supp}(g) \subset [0, L]$; if $a \leq L,\ b \leq \frac{1}{L}$, then $\mathbf{G}(g, a, b)$ is a Gabor frame, and the frame operator S is the following multiplication operator,*

$$\mathrm{S}f(t) = \left(b^{-1} \sum_{l \in \mathbb{Z}} |g(t - al)|^2 \right) f(t) \ . \tag{2.1.13}$$

The hypotheses of Theorem 2.1.7 define the *painless case*, where the dual window $\widetilde{g}$ is easy to calculate by means of a multiplication of the original one,

$$\widetilde{g}(t) = \mathrm{S}^{-1} g(t) = \frac{g(t)}{b^{-1} \sum_{l \in \mathbb{Z}} |g(t - al)|^2} \ . \tag{2.1.14}$$

Remark 2.1.8. As we see from formula (2.1.7), the atoms needed for the reconstruction of f are the time-frequency shifts of $\widetilde{g}$ according to the lattice Λ. From the identity (2.1.14) which expresses $\widetilde{g}$ in the painless case, we have that in these conditions the whole analysis-reconstruction scheme can be implemented with fast FFT-based methods: the input for transform to take is a short one, as both the analysis and reconstruction steps are limited to the short-length support of the window g. Throughout the work, we will indicate as *fast* those algorithms whose computational order is due to the FFT of short-length signals. ■

2.2. Extensions of stationary Gabor frames

The limit of stationary Gabor frames is that the decomposing atoms are defined from the same original function, thus constraining the type of information we can deduce from the analysis coefficients; if we were able to consider frames where several families of atoms coexist, than we would have an analysis with variable information, at the price of a higher redundancy. In our adaptive framework, we look for a method to achieve analyses with multiple resolutions, combining the information coming from the decompositions of a signal in several frames with different window functions. *Multi-window* Gabor frames have been introduced in [**Zibulski and Zeevi, 1997**] to provide the original Gabor analysis with more flexible multi-resolution techniques: given a finite set of index S and different Gabor frames $\mathbf{G}(g_s, a_s, b_s)$ with $s \in S$, a multi-window Gabor frame is obtained as the union of the single given frames. Similarly, the analysis operator C is given by the union of the analysis coefficients obtained with the individual frames $\mathbf{G}(g_s, a_s, b_s)$.

The different g_s do not necessarily share the same type or shape: a typical strategy inspired by the wavelet approach is to scale an original window with a finite number of scaling,

$$g_s(t) = \frac{1}{\sqrt{s}}\, g\left(\frac{t}{s}\right) ; \tag{2.2.1}$$

therefore, in such a multi-window frame the signal is analyzed with several different tradeoff between time and frequency resolution. The disadvantage is that a significant redundancy is introduced, which lowers the readability of the analysis. Moreover, each individual frame give coefficients over the whole time-frequency space, while we

would like to locally select coefficients from a unique analysis. This approach has been introduced with *quilted frames* (see [Dörfler, 2011]): in these systems, the choice of the analysis window in the STFT depends on both the time and the frequency location of the considered coefficient. If this strategy is plain from the analysis point of view, for a quilted frame there is not a general painless case for the inversion of the frame operator: this implies that finding the dual frame for the reconstruction can be difficult, and dedicated algorithms are in general computationally expensive.

In this work, we focus on analysis methods providing an analytic fast computation of a dual frame, in the sense of Remark 2.1.8: when such methods are not theoretically achievable, we consider strategies for a good approximation of the dual frame, still using fast algorithms. The rest of this section and the next ones describe the details of our approach.

2.2.1. Nonstationary Gabor frames. A strategy to get an adaptive framework preserving a fast reconstruction method, in the sense of Remark 2.1.8, is given by Nonstationary Gabor frames (NGF, see [Jaillet et al., 2009, Balazs et al., 2011, Søndergaard et al.,]): we first consider the so-called *time case*, where the starting point is a set of different window functions. A unique analysis window is chosen depending on the time location of the coefficient, originating a globally irregular lattice Λ (see Figure 2.2): for each time index l, a window g_l is chosen among the different set considered, which is centered at time a_l; then g_l is modulated according to a frequency step, indicated with b_l as it depends on the time index l, too; therefore, Λ is irregular over time, with regular frequency sampling at each time position.
We have a similar configuration for NGF in the frequency case, where the analysis window is chosen depending on the frequency location of the coefficient, thus originating a lattice Λ which is irregular over time, with regular time sampling at each frequency point (see Figure 2.3).

Referring to the time case, a nonstationary Gabor frame is thus given by the atoms

$$g_{k,l}(t) = M_{b_l k} g_l(t) = g_l(t)\, \mathrm{e}^{2\pi i k b_l t}\ , (l,k) \in \mathbb{Z}^2\ , \tag{2.2.2}$$

where b_l is the frequency step associated to the window g_l. For NGFs there exist a painless case for the calculation of the dual, whose conditions are detailed in the following theorem ([Balazs et al., 2011], Theorem 1).

Theorem 2.2.1. *Suppose that the windows $g_l \in L^2(\mathbb{R})$ have compact support, $\mathrm{supp}(g_l) \subseteq [c_l, d_l]$, and that the frequency steps b_l are chosen such that $d_l - c_l \leq \frac{1}{b_l}$; then the frame operator S is the following multiplication operator,*

$$\mathrm{S}f(t) = \left(\sum_{l \in \mathbb{Z}} \frac{1}{b_l} |g_l(t)|^2 \right) f(t)\ . \tag{2.2.3}$$

As a consequence, if there exist two constants C, D such that

$$0 < C < \sum_{l \in \mathbb{Z}} \frac{1}{b_l} |g_l(t)|^2 < D < \infty\ , \tag{2.2.4}$$

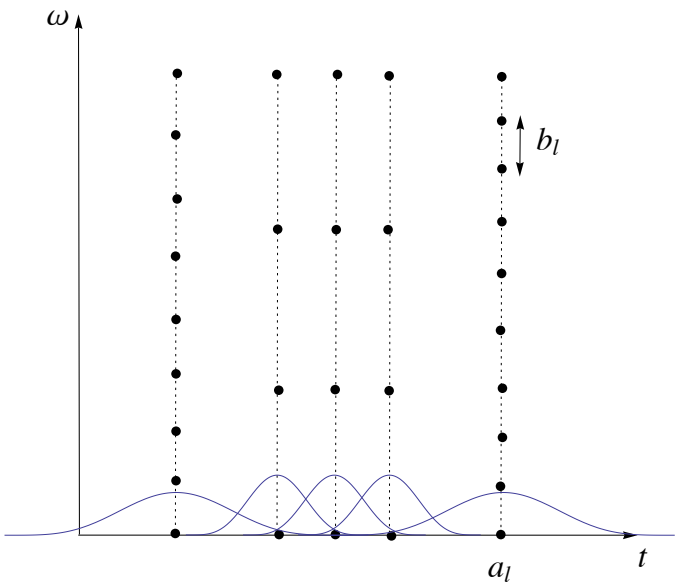

FIGURE 2.2. *Time-frequency centers for a Nonstationary Gabor frame in the time case, with variable time locations a_l and frequency steps b_l, depending on the time index l.*

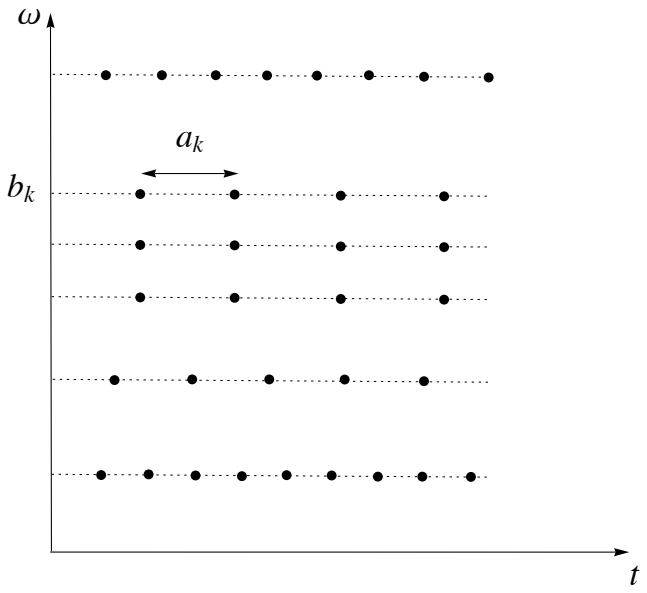

FIGURE 2.3. *Time-frequency centers for a Nonstationary Gabor frame in the frequency case, with variable frequency locations b_k and time steps a_k, depending on the frequency index k.*

then the set (2.2.2) is a frame whose dual frame is given by

$$\widetilde{g}_{k,l}(t) = \frac{g_l(t)\,\mathrm{e}^{2\pi i k b_l t}}{\sum_{l\in\mathbb{Z}} \frac{1}{b_l}|g_l(t)|^2}\,. \tag{2.2.5}$$

Having an expression of the dual frame, it is now possible to define a reconstruction formula; we can still make use of the compact form,

$$f = \mathrm{D}_{\widetilde{\mathrm{g}}\mathrm{l}}(\mathrm{C}_{\mathrm{g}\mathrm{l}} f)\,, \tag{2.2.6}$$

appropriately considering the window and lattice variations at each time location. As we see from the expression of the dual frame, this formula can be implemented with a fast FFT-based algorithm.

2.3. Gabor Multipliers

In the Gabor framework described till now, the analysis and synthesis operator are the only elements considered; that is, the analysis coefficients are not modified before the reconstruction. Spectral processing techniques are based on analysis manipulations, which determine the desired effect in the re-synthesized signal. Gabor multipliers (see [**Feichtinger and Nowak, 2002**] for a complete survey) provide a mathematical model to manipulate the analysis coefficients by means of multiplications, and to define operators in the signal domain from a modeling in the analysis domain. We consider the definition of Gabor multiplier in $L^2(\mathbb{R})$, which can be generalized to the $L^2(\mathbb{R}^d)$ general case.

Definition 2.3.1. Let g^1, g^2 be two functions in $L^2(\mathbb{R})$, Λ a time-frequency lattice and $\mathbf{m} = (m_\lambda)_{\lambda\in\Lambda}$ a complex-valued sequence; the *Gabor multiplier* $\mathbf{G}_{m,\Lambda}^{g^1,g^2}$, with *upper symbol* $\mathbf{m}$, is given by

$$\mathbf{G}_{\mathbf{m},\Lambda}^{g^1,g^2}(f) = \mathrm{D}_{\mathrm{g}^2}(\mathbf{m}\cdot \mathrm{C}_{\mathrm{g}^1} f)\,, \tag{2.3.1}$$

where $\mathbf{m}\cdot \mathrm{C}_{\mathrm{g}^1} f$ is the pointwise multiplication of $\mathbf{m}$ and $\mathrm{C}_{\mathrm{g}^1} f$.

In particular, if $\mathbf{G}(g,a,b)$ is a Gabor frame with $\Lambda = a\mathbb{Z}\times b\mathbb{Z}$, and $\mathbf{m}\in\ell^\infty(\Lambda)$, then the frame condition implies that $\mathbf{G}_{\mathbf{m},\Lambda}^{g,\widetilde{g}}$ is a bounded operator.

2.3.1. Weighted Frames.

The definition of spectral manipulations can be also approached from the point of view of the decomposing atoms; in [**Balazs et al., 2010**], the concept of weighted frame is introduced.

Definition 2.3.2. Consider a separable Hilbert space H and a set of atoms $\{\phi_\gamma\}_{\gamma\in\Gamma}$ in H, and a sequence $(w_\gamma)_{\gamma\in\Gamma}$ of complex numbers. The set $\{w_\gamma\phi_\gamma\}_{\gamma\in\Gamma}$ is a *weighted frame* for H if there exist two positive non zero constants A and B such that for all $f\in H$,

$$A\|f\|^2 \le \sum_{\gamma\in\Gamma} |\langle f, w_\gamma\phi_\gamma\rangle|^2 \le B\|f\|^2\,. \tag{2.3.2}$$

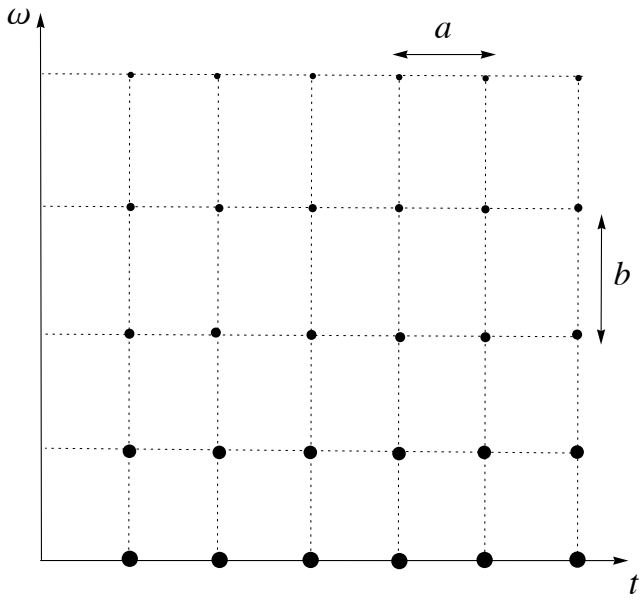

FIGURE 2.4. *Visual representation of time-frequency centers for a weighted stationary Gabor frame* $\mathbf{G}^w(g,a,b)$*: here, the weighting sequence* $(w_\lambda)_{\lambda\in\Lambda}$ *is a function of frequency, and is sketched through the points size.*

If a Gabor frame $\mathbf{G}(g,a,b)$ forms a weighted frame with a sequence $(w_\lambda)_{\lambda\in\Lambda}$, we indicate with $\mathbf{G}^w(g,a,b)$ the corresponding weighted Gabor frame (see Figure 2.4 for a graphical representation). If we indicate with C_g^w the analysis operator associated to $\mathbf{G}^w(g,a,b)$ and consider $\mathbf{m}=(w_\lambda)_{\lambda\in\Lambda}$, then we can write

$$\mathbf{G}_{\mathbf{m},\Lambda}^{g,\tilde{g}} f = \mathrm{D}_{\tilde{\mathrm{g}}}(\mathrm{C}_\mathrm{g}^w f)\ , \tag{2.3.3}$$

showing the relation between a Gabor multiplier and a weighted Gabor frame. In the following Lemma (Lemma 4.3 in [**Balazs et al., 2010**]), the structure of the dual of a weighted frame is considered, when there exist constants $0<E\leq F$ such that the sequence $(w_\lambda)_{\lambda\in\Lambda}$ verifies $E\leq|w_\lambda|\leq F$ for every λ (such sequences are called *semi-normalized*).

Lemma 2.3.3. *Let* $(w_\gamma)_{\gamma\in\Gamma}$ *be a semi-normalized sequence with bounds* E,F*. If* $\{\phi_\gamma\}_{\gamma\in\Gamma}$ *is a frame with bounds* A *and* B*, then* $\{w_\gamma\phi_\gamma\}_{\gamma\in\Gamma}$ *is also a frame with bounds* E^2A *and* F^2B*. The sequence* $\{w_\gamma^{-1}\widetilde{\phi}_\gamma\}_{\gamma\in\Gamma}$ *is a dual frame of* $\{w_\gamma\phi_\gamma\}_{\gamma\in\Gamma}$*.*

The reconstruction formula for a weighted frame is therefore the standard one given in (2.1.5), with non weighted atoms. In Section 2.5, we define an approximation method allowing the weighting sequences to be non semi-normalized, and to have zero values in particular: this reflects the standard technique of suppressing spectral component by setting to zero the corresponding analysis coefficients before the re-synthesis.

2.4. Sound transformation and re-synthesis by means of adaptive representations

Having defined adaptive analyses, there are two major problems to solve: the interpretation for the individual coefficients, and the definition of a reconstruction method. For the former, we choose to develop our framework in the Gabor analysis context to take advantage of the STFT interpretation of the coefficients, as motivated in Chapter 1; but still, having analyses with varying resolution require changes of the standard spectral processing techniques: if the lattice is irregular along frequency, for instance, phase relations between different bins have to be interpreted, locally, considering their variable spacing. In this work, spectral processing techniques are not extensively treated, but the algorithms we develop are designed to allow reimplementations of existing methods, such the ones available in the phase vocoder approach.

For the reconstruction task, we have to distinguish two cases: if the analysis window varies depending on time or frequency individually, or if it depends on them both. For the first case, nonstationary Gabor frames provide fast reconstruction algorithms within the painless conditions (see Subsection 2.2.1). In particular, we use windows with compact support, thus forming NGF in the time case, and we provide the adaptive framework with the automatic decision routine detailed in Chapter 4.

In several situations, the optimal resolution to separate independent sound components varies locally, both depending on time and frequency (this case has been detailed in [**Dörfler, 2011**] among others): for instance, if a bass and a drum are playing together, we wish to use a better frequency resolution at low frequencies, where most of the bass partials lay; on the other hand, in spectral regions where bass harmonics are negligible, we would like to privilege time resolution for a more precise identification of drum hits; but if we have zones where one of the two instruments plays alone, then time or frequency resolution should be privileged over the whole spectrum. In such cases, fast methods guaranteeing a perfect recover of the original signal are in general not possible. The difficulties arise when we want to define a fast FFT-based reconstruction formula merging the different analysis coefficients: we would like to separate and use them depending on their band of pertinence. On the contrary, reconstructing a limited time-frequency portion of the signal, we are not allowed to neglect the contribution coming from *far* coefficients: analysis windows with compact time support cannot have a compactly supported Fourier transform; from the analysis point of view, this means that a spectrogram coefficient affects the signal reconstruction across the whole frequency dimension. If the coefficients outside a certain band are neglected, the reconstruction error comes mainly from the fact that we are setting to zero the contribution of atoms whose Fourier transforms spread into the band of interest. We can limit such an influence with a choice of well-localized time-frequency atoms: even if their frequency support is not compact, they have a fast decay outside a certain compact region; therefore, if the atoms are well-localized, only a few of the *far* coefficients actually are involved. Thus, fast methods provide a reconstruction error which should be kept small in order to preserve a perceptual perfect reconstruction: that is, the re-synthesis does not exactly recover the original sound, but the error

remains perceptually negligible.

Even in cases where the resolution changes both depending on time and frequency, if no information is lost, frame theory provides synthesis methods with perfect reconstruction; however, this is a typical case where the calculation of the dual frame for the signal reconstruction cannot, in general, be achieved with a fast algorithm: thus a choice must be done between a slow analysis/re-synthesis method guaranteeing perfect reconstruction and a fast one giving an approximation with a certain error. We consider two different approaches to obtain fast algorithms, which are sketched in the following subsections.

2.4.1. Filter bank. The signal is first filtered with an invertible bank of P pass band filters, to obtain P different band limited signals; for each of these bands a different nonstationary Gabor frame $\{g^p_{k,l}\}$ of windows with compact time support is used: the index k may thus indicate different time centers depending on the individual NGF, and the time-dependent window functions $g^p_k,\ p = 1, ..., P$, may also be different . The other members of the frame are time-frequency shifts of g^p_k,

$$g^p_{k,l} = g^p_k(t)\, \mathrm{e}^{2\pi i b^p_k l t}\ , \tag{2.4.1}$$

where $k,\ l \in \mathbb{Z}$ and b^p_k is the frequency step associated to the p-th NGF at the time index k. We always assume to be in the time painless case, so each band-limited signal is perfectly reconstructed with an FFT-based expansion of the analysis coefficients in the dual frame $\{\widetilde{g_{k,l}}^p\}$. Note that by this notation we denote the dual frame for a fixed p.

By summing the reconstructed bands, we obtain a perfect reconstruction of the original signal. An important remark is that the reconstruction of the individual filtered signals is perfect as long as all the frequency coefficients within all the P analyses are used. On the other hand, for every analysis we are interested in considering only the coefficients within the corresponding frequency band, thus introducing a reconstruction error (see Figure 2.5). The results detailed in [**Matusiak and Eldar, 2010**] provide a useful tool: they give an exact upper bound of the reconstruction error when reconstructing a compactly supported and essentially band-limited signal from a certain subset of its analysis coefficients within a Gabor frame (see Section 2.6).

2.4.2. Analysis–weighting. The signal is first analyzed with P different NGFs $\{g^p_{k,l}\}$ of windows with compact time support . Each analysis is associated to a certain frequency band, and its coefficients are weighted to match this association. We look for a reconstruction formula to minimize the reconstruction error when expanding the weighted coefficients within the union of the P individual dual frames $\cup_{p=1}^{P}\{\widetilde{g_{k,l}}^p\}$ (see Example 2.4.1).

Example 2.4.1. To give a visual interpretation of the analysis-weighting approach, we consider the basic example where $P = 2$ and the frames are stationary. Thus, we can think of associating a certain frame $\mathbf{G}(g^1, a_1, b_1)$ to the lower frequency band, and another frame $\mathbf{G}(g^2, a_2, b_2)$ to the upper one. The association is realized by means of

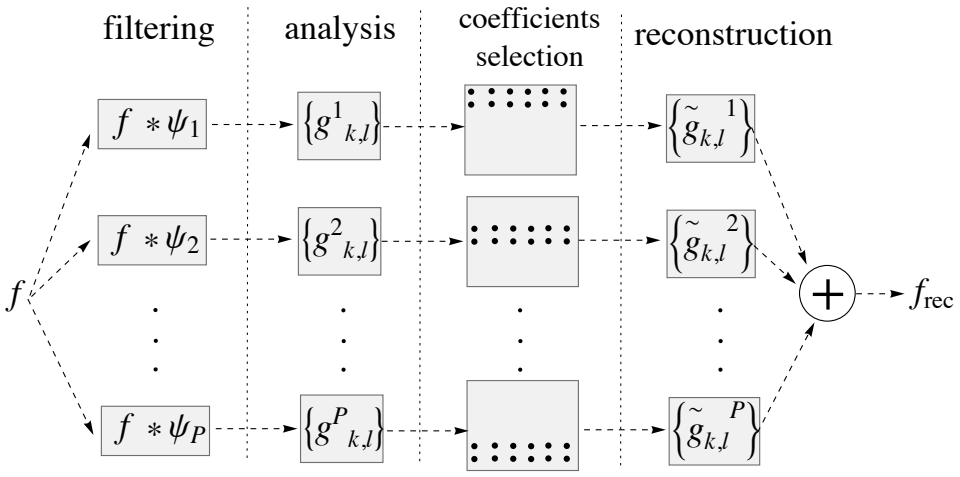

FIGURE 2.5. *Block diagram detailing the steps of the filter bank approach (see Subsection 2.4.1): the signal is first filtered with a bank of P band-pass filters; the filtered signals are analyzed with P NGFs; the coefficients in the analyses are selected depending on the corresponding frequency band; the filtered signal are approximated with expansions of the selected coefficients with the corresponding dual frames; the reconstructed signals are summed to give an approximation of the original signal.*

weighting sequences depending on the frequency location of a coefficient: on top of Figure 2.6, two complementary binary masks are used, setting to zero the coefficients which do not belong to the appropriate frequency band; at the bottom, the frequency supports of the weighting sequences have a certain overlap. ■

2.5. Extended weighted frames approach

In [**Liuni et al., 2011a**], we propose a first intuitive solution for the reconstruction task outlined in Section 2.4, adapting the analysis window in time *and* frequency. We focus here on the analysis-weighting approach, in the basic case of two bands; so we split the frequency dimension into high and low frequencies, with $P = 2$. The two corresponding NGFs are given by the automatic adaptation routine described in Chapter 4. The analysis-weighting method is treated with an extension of the weighted Gabor frames approach, which will give us a closed reconstruction formula.

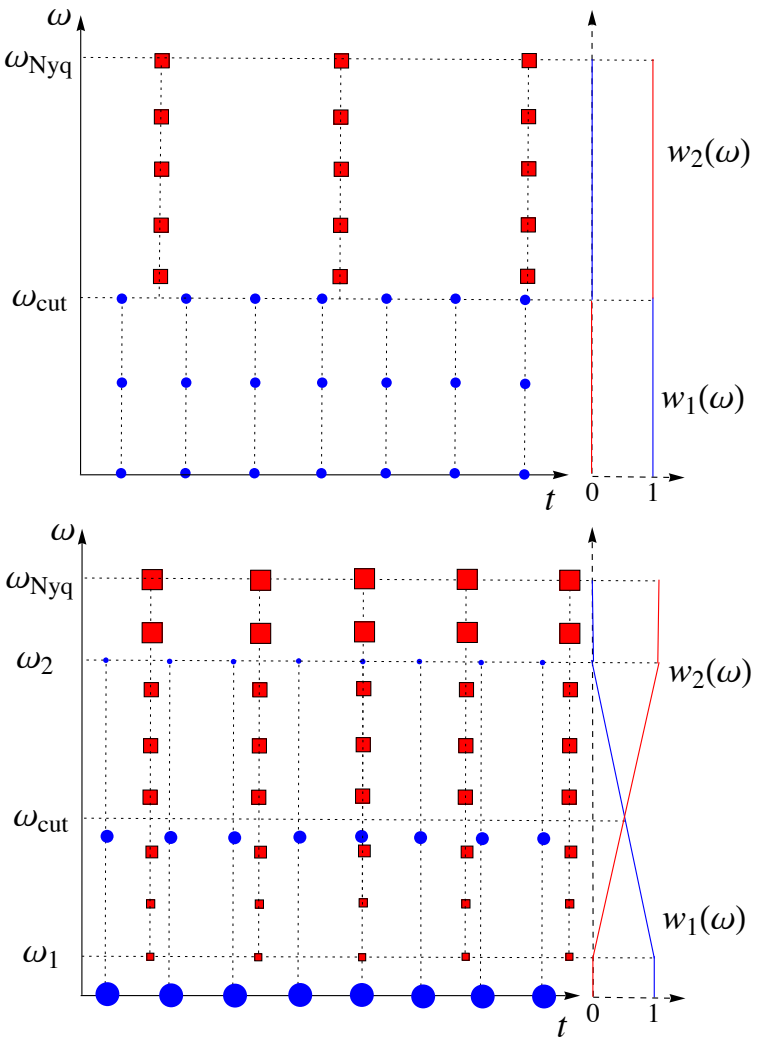

FIGURE 2.6. *Visual representation of time-frequency centers for the analysis-weight approach (see Example 2.4.1): the coefficients of the two Gabor frames are represented with different colors and shapes, their size is related to the value of the corresponding weight functions $w_1(\omega)$ and $w_2(\omega)$; ω_{Nyq} is the Nyquist frequency, ω_{cut} is the cut frequency for the binary masks (on top), ω_1 and ω_2 are the bound of the frequency overlap between the two weights (at the bottom).*

2.5.1. Reconstruction from Weighted Frames. Let $P \in \mathbb{N}$ and $\{g^p_{k,l}\}$ be different NGFs, $p = 1, \dots, P$, where k and l are the time and frequency index, respectively. We consider weight functions $0 \leq w^p(\nu) \leq \infty$: for every p, they only depend on the frequency location ν. The idea is to set to zero the coefficients not belonging to the frequency portion which the p-th analysis has been assigned to; in this way, every analysis just contributes to the reconstruction of the signal portion of its pertinence: when $P = 2$, we divide the plane into two portions, high and low frequencies. For each NGF

$\{g^p_{k,l}\}$, we write $c^p_{k,l} = w^p(b^p_k l)\langle f, g^p_{k,l}\rangle$ to indicate the weighted analysis coefficients, and we consider the following reconstruction formula:

$$\widetilde{f} = \mathscr{F}^{-1}\left(\frac{1}{\mathrm{p}(\nu)}\mathscr{F}\left(\sum_{p=1}^{P}\sum_{k,l}\mathrm{r}(p,k,l)\right)\right) , \tag{2.5.1}$$

where $\mathrm{p}(\nu) = \sharp\{p : w^p(\nu) \geq \epsilon\}$ and for every $\epsilon > 0$, $\mathrm{r}(p,k,l)$ is 0 if $w^p(b^p_k l) < \epsilon$, else

$$\mathrm{r}(p,k,l) = \left(w^p(b^p_k l)\langle f, g^p_{k,l}\rangle\right)\frac{1}{w^p(b^p_k l)}\widetilde{g_{k,l}}^p . \tag{2.5.2}$$

We see that non-zero weights cancel each other: this reconstruction formula still makes sense, as the goal is exactly to find a reconstruction as an expansion of the $c^p_{k,l}$. We give now an interpretation of the introduced formula. If w^p is a semi-normalized sequence for each p, that is there exist constants m_p, n_p and $\epsilon > 0$ such that

$$\epsilon < m_p \leq w^p(b^p_k l) \leq n_p < \infty \tag{2.5.3}$$

for all p, then $\mathrm{p}(\nu) = p$ and the equation (2.5.1) becomes

$$\widetilde{f} = \frac{1}{P}\sum_{p=1}^{P}\sum_{k,l}\left(w^p(b^p_k l)\langle f, g^p_{k,l}\rangle\right)\frac{1}{w^p(b^p_k l)}\widetilde{g_{k,l}}^p = f . \tag{2.5.4}$$

Here, the perfect reconstruction is guaranteed, as detailed in Subsection 2.3.1 about weighted frames: indeed, in the hypothesis of semi-normalization the sequence $w^p(b^p_k l)g^p_{k,l}$ is a frame, with $\frac{1}{w^p(b^p_k l)}\widetilde{g_{k,l}}^p$ as one of its dual.
For weights which are not bounded from below, but still non-zero, the reconstruction still works for $\epsilon = 0$: the sequences $w^p(b^p_k l)\cdot g^p_{k,l}$ are not frames anymore (for each p), but complete Bessel sequences (also known as upper semi-frames [**Antoine and Balazs, 2011**]). This reconstruction can be unstable, though.

In our case, these hypotheses are not verified, as we need to set to zero a certain subset of the coefficients within both of the analyses; thus the equation (2.5.1) will in general give an approximation of f. In Section 4.4 we give several examples of reconstruction following this approach, evaluating the reconstruction error depending on:

- the signal spectral features at frequencies ν where $\mathrm{p}(\nu) > 1$;
- the features of the w^p sequences and the $\mathrm{p}(\nu)$ function.

A first natural choice for the weights w^p is a binary mask, as the reconstruction formula takes the very simple form detailed in equation (2.5.6), allowing a direct interpretation of the contributions coming from the two bands; moreover, a fast implementation can be deduced from the general full band algorithm. So we consider $P = 2$ and ω_c a certain cut value, then

$$w^1(\nu) = \begin{cases} 1 & \text{if } \nu \leq \omega_c \\ 0 & \text{if } \nu > \omega_c \end{cases} \tag{2.5.5}$$

and $w^2(\nu) = 1 - w^1(\nu)$. In this case $\mathrm{p}(\nu) = 1$ for every frequency ν and the equation (2.5.1) becomes

$$\widetilde{f} = \sum_{b_k^p l \leq \omega_c} \langle f, g_{k,l}^1 \rangle \widetilde{g_{k,l}}^1 + \sum_{b_k^p l > \omega_c} \langle f, g_{k,l}^2 \rangle \widetilde{g_{k,l}}^2 \,. \tag{2.5.6}$$

The reconstruction error in this case will in general be large at frequencies corresponding to coefficients close to the cut value ω_c; depending on the sound spectral features, a way to reduce this error is to allow the w^p weights to have an overlap (see Section 4.4); this implies that more coefficients from different analyses contribute to the reconstruction of a same portion of signal, thus weakening their interpretation. There are still several open problems in this sense, a basic one being how to display a representation of the signal such the one described; there are, at least, two possibilities involving weighted means of the coefficients at a certain time-frequency location:

- $d_{k,l} = \frac{1}{\sum_p w^p} \cdot \sum_p c_{k,l}^p$, displaying $|d_{k,l}|$, or
- $d_{k,l}^{(A)} = \frac{1}{\sum_p w^p} \cdot \sqrt{\sum_p \left| c_{k,l}^p \right|^2}$.

The second one is the one we adopt (see Section 4.4); as our algorithm keeps the original coefficients in memory, we can still use the reconstruction scheme (2.5.1). If the original coefficients were not available, a further question would be how to reconstruct the signal from an expansion of the $d_{k,l}$ or $d_{k,l}^{(A)}$ coefficients. If $d_{k,l}^{(A)}$ is used, we also have to address the problem of the phase. This approach is useful when dealing with representations where the phase information is lost, as with reassigned spectrogram or spectral cepstrum. A solution could be to use an iterative approach, like the one described in [**Griffin and Lim, 1984**] adapted to frame theory, or use a system with a high redundancy (see [**Balan et al., 2006**]).

From a computational point of view, we are interested in limiting the size of the signal for the direct and inverse Fourier transforms in (2.5.1), as this would largely improve the efficiency of the algorithm. An equivalent form of the formula (2.5.1) in this sense is the following,

$$\widetilde{f} = \sum_{p,k,l} c_{k,l}^p \mathscr{F}^{-1} \left(\frac{1}{\mathrm{p}(\nu)} \mathscr{F} \left(\frac{\widetilde{g_{k,l}}^p}{w^p(b_k^p l)} \right) \right) \,. \tag{2.5.7}$$

As the weights depend only on frequency, so does the normalizing function $\mathrm{p}(\nu)$: the approximate dual frames for the expansion in (2.5.7) are thus calculated only once, with Fourier direct and inverse transforms limited to the short-length normalization and window functions: therefore, the computational cost due to the approximation of the dual frame is small, and the reconstruction formula (2.5.7) has the same complexity order as the standard fast inversion of NGFs.

2.6. Filter bank approach

In this section, we define a second new approximation method based on analyses with resolution changing in time and frequency, together with theoretical bounds

for the reconstruction error. In Subsections 2.6.1 and 2.6.3, we extend the results in [Matusiak and Eldar, 2010] to the case of filtered signals, obtaining the approach we indicate as *filter bank* in both the cases of stationary and nonstationary Gabor frames. This part of the work is the result of a collaboration with Ewa Matusiak and Monika Dörfler.

In Subsection 2.6.2, we extend the estimates obtained in the stationary case, defining a different version of the introduced approximation method: Gabor multipliers are used instead of filters, leading to the analysis-weighting approach implemented in our adaptive framework (See Section 4.4).

We consider here a finite duration signal f supported on the interval $[-\beta/2, \beta/2]$, and ϵ_Ω-bandlimited to the interval $[-\Omega/2, \Omega/2]$, $\beta, \Omega \in \mathbb{R}^+$, which is the case we are interested in when working with music signals; this implies that $|\widehat{f}(\omega)| < \epsilon_\Omega$ for every $\omega \notin [-\Omega/2, \Omega/2]$. We first want to reconstruct f using different STFTs of a certain number of its filtered versions; in particular, we use different window functions for each different version, and compute the reconstruction error based on the estimates in [Matusiak and Eldar, 2010].

2.6.1. Filter bank approach with stationary Gabor frames. Given $P \in \mathbb{N}$, consider the functions ψ_p, $p = 1, ..., P$, which are the impulse responses of P filters with finite time supports $[-T_p/2, T_p/2]$, whose essential frequency supports are $[\Omega_p^1, \Omega_p^2]$ and cover the essential bandwidth of f. We assume also that at most two essential frequency supports of $\widehat{\psi_p}$ overlap at the same time, and that they satisfy $\widehat{\psi}_1(\omega) + ... + \widehat{\psi}_P(\omega) = 1$ on $[-\Omega/2, \Omega/2]$. We consider P windows g^p compactly supported on $[-W_p/2, W_p/2]$ such that $\|g^p\|_2 = 1$ and the STFT of the signal f ,

$$\mathcal{V}_p f(t, \omega) = \int_{\mathbb{R}} f(\tau) \overline{g^p(\tau - t)}\, \mathrm{e}^{-2\pi i \omega \tau}\, \mathrm{d}\tau \,, \tag{2.6.1}$$

using the compact form $\mathcal{V}_p f(t, \omega) = \langle f, M_\omega T_t g^p \rangle$. We denote by f_p a filtered version of f, $f_p = f * \psi_p$ and $\widehat{f} = \sum_p \widehat{f_p}$ on $[-\Omega/2, \Omega/2]$. Each one of the f_p filtered versions is a finite duration signal, supported on the interval $[-\beta/2 - T_p/2, \beta/2 + T_p/2]$, and ϵ_p-bandlimited to the interval $[\Omega_p^1, \Omega_p^2]$.
Now, if we consider P stationary Gabor frames $\mathbf{G}(g^p, a_p, b_p)$, we obtain a sampling of $\mathcal{V}_p f_p$ composed by the values

$$c_{k,l}^p = \langle f_p, M_{b_p k} T_{a_p l} g^p \rangle \,, \ (k, l) \in \mathbb{Z}^2 \,; \tag{2.6.2}$$

here, the time step a_p and the frequency step b_p depend on the window function, and are chosen in order for the sampled analysis to be more redundant than the critical case, $a_p b_p < 1$: the goal is to have a stable frame with well concentrated windows, hence overcompleteness is necessary. In these hypotheses, the estimates in [Matusiak and Eldar, 2010] allow to approximate f_p with a finite expansion involving the sampled analysis coefficients and the dual window. In particular, if we indicate

with $\widetilde{g^p}$ the dual of g^p, then for every $\epsilon > 0$ there exist two finite sets $K^p, L^p \subset \mathbb{Z}$ such that the truncated expansion $\overline{f_p}$, given by

$$\overline{f_p} = \sum_{k \in K^p} \sum_{l \in L^p} c^p_{k,l} M_{b_p k} T_{a_p l} \widetilde{g^p} \,, \tag{2.6.3}$$

verifies the following inequality,

$$\left\| f_p - \overline{f_p} \right\|_2 \leq C_p(\epsilon_p + \epsilon) \|f_p\|_2 \,, \tag{2.6.4}$$

where $C_p = (1 + 1/a_p)(1 + 1/b_p)\|\widetilde{g^p}\|_{S_0} \|g^p\|_{S_0}$ and $\|g\|_{S_0} = \|\mathcal{V}_{g_0} g\|_1$ with g_0 Gaussian. The set L^p contains the time positions la_p for which support of f_p overlaps with support of g^p shifted by la_p; the set K^p contains the frequency positions kb_p for which essential support of $\widehat{f_p}$ overlaps with essential support of $\widehat{g^p}$ shifted by kb_p . Then the cardinality of L_p equals

$$|L^p| = 2 \left\lceil \frac{\beta + T_p + W_p}{2a_p} \right\rceil - 1 \,; \tag{2.6.5}$$

if g^p_c is a $[-\alpha_p/2, \alpha_p/2]$ −bandlimited approximation of g^p in S_0, meaning $\|g^p - g^p_c\|_{S_0} \leq \epsilon \|g\|_{S_0}$, then the cardinality of K_p equals

$$|K^p| = \left\lceil \frac{\Omega^2_p - \Omega^1_p + \alpha_p}{b_p} \right\rceil \,. \tag{2.6.6}$$

Given these estimates, we want to approximate the original signal summing the truncated expansions; therefore, the reconstruction error we obtain is bounded by the sum of the error bounds for the filtered components. We indicate with C_P and ϵ_P the maxima over all C_p and ϵ_p, respectively. We can thus determine an upper bound directly from equation (2.6.4): for every $\epsilon > 0$, with the appropriate sets and constants we have

$$\begin{aligned} \left\| f - \sum_p \overline{f_p} \right\|_2 &\leq \left\| f - \sum_p f_p \right\|_2 + \left\| \sum_p f - \sum_p \overline{f_p} \right\|_2 \\ &\leq \epsilon_\Omega \|f\|_2 + C_P(\epsilon_P + \epsilon) \sum_p \|f_p\|_2 \,. \end{aligned} \tag{2.6.7}$$

We want to express the error as a function of $\|f\|_2$: by applying triangle inequality, we have that $\sum_p \|f_p\|_2 \leq \|f\|_2 \cdot P \max_p \|\widehat{\psi_p}\|_\infty$; so, writing $C_\psi = P \cdot \max_p \|\widehat{\psi_p}\|_\infty$, we have

$$\left\| f - \sum_p \overline{f_p} \right\|_2 \leq (\epsilon_\Omega + C_\psi C_P(\epsilon_P + \epsilon)) \|f\|_2 \,. \tag{2.6.8}$$

Remark 2.6.1. The choice of the ψ_p functions has an influence on the error we obtain: assuming to work with S_0 windows (introduced in Subsection 2.6.3), that have "nice" time-frequency properties guaranteed, the ϵ_p constant, which concerns the essential frequency support of $\widehat{f_p}$, depends on the regularity of ψ_p : the smoother it is, the faster $\widehat{f_p}$ decays of out of its essential support, and then the smaller ϵ_p.

On the other hand, ϵ depends on the number of coefficients used in the expansion (2.6.3); obviously, considering more frequency coefficients we obtain a better approximation, reducing ϵ. In this sense, an interesting perspective is to implement an automatic method to determine the number of coefficients needed to achieve a desired precision, given the analysis parameters. ■

2.6.2. Filter bank approach and Gabor multipliers. Spectral processing techniques often avoid manipulations in the signal domain, privileging modifications of the analysis coefficients, followed by the re-synthesis. Therefore, we look for an estimate like the one in equation (2.6.7) when working with Gabor multipliers instead of filters. In particular, we want to replace each filter ψ_p with a Gabor multiplier $\mathbf{G}^{g^p,\widetilde{g^p}}_{\mathbf{m}_p,\Lambda_p}$, whose symbol $\mathbf{m}_p$ does not depend on time, and matches the frequency response $\widehat{\psi}_p$ of the filter, $\mathbf{m}_p(t,\omega) = \widehat{\psi}_p(\omega)$. We thus obtain weighted versions of the STFTs of the signal f,

$$W_p f(t,\omega) = \mathcal{V}_p f(t,\omega)\mathbf{m}_p(t,\omega). \tag{2.6.9}$$

Our aim is to replace $\mathcal{V}_p f_p(t,\omega)$ by the weighted analyses $W_p f(t,\omega)$, and we write their sampling according to the lattices Λ_p as follows,

$$d^p_{k,l} = W_p(a_p l, b_p k)\ ; \tag{2.6.10}$$

indeed, if we write $g(\tau - t) = g_t(\tau)$, we see that

$$\mathcal{V}_p f_p(t,\omega)(t,\omega) = ((\widehat{f}\cdot\widehat{\psi}_p) * \widehat{g^p_t})(\omega)\ ,\ W_p f(t,\omega) = (\widehat{f} * \widehat{g^p_t})(\omega)\cdot\widehat{\psi}_p(\omega)\ ; \tag{2.6.11}$$

therefore, the difference depends on how similar multiplication and convolution with the atoms are, if their roles are switched. To quantify this difference, we need to clarify the relation between a time invariant filter and a Gabor multiplier. Hilbert-Schmidt operators, as well as a larger class of operators called *underspread*, can be well approximated by means of Gabor multipliers (see [**Dörfler and Torrésani, 2007**, **Dörfler and Torresani, 2010**, **Matz and Hlawatsch, 1998**]): given an underspread operator H, its best approximation by a Gabor multiplier $\mathbf{G}^{g_1,g_2}_{\mathbf{m},\Lambda}$ can be calculated, with an error depending on the *spreading function* η_H of H and $\mathcal{V}_{g_1}g_2$. Time invariant convolution operators, such as filters, are not underspread; but still, we envisage that it is possible to estimate the error when approximating a convolution operator A with a Gabor multiplier G of the type we are considering: this result is the object of an ongoing collaborative work (see [**Engelputzeder, 2011**, **Balazs et al., 2012**]). Knowing that the Hilbert-Schmidt norm of the difference $\|A - G\|_{HS}$ is conveniently small, the aim is to deduce a pointwise inequality for the sampled analyses we work with, that is for each $(k,l) \in \mathbb{Z}^2$, the following inequality must hold for a small ϵ^*_p,

$$|c^p_{k,l} - d^p_{k,l}| \leq \frac{\epsilon^*_p}{PKL}\ , \tag{2.6.12}$$

where KL is the number of coefficients in the expansion (2.6.3); here, we assume this inequality to hold. Using the coefficients $d^p_{k,l}$ in the same expansion, we obtain

$$\overline{f^*_p} = \sum_{k\in K^p}\sum_{l\in L^p} d^p_{k,l} M_{b_p k} T_{a_p l} \widetilde{g^p} \,, \tag{2.6.13}$$

and

$$\|\overline{f_p} - \overline{f^*_p}\|_2 \leq \frac{\epsilon^*_p}{P} \cdot \|\widetilde{g^p}\|_2 \,. \tag{2.6.14}$$

We can thus estimate the further approximation error introduced by considering the Gabor multiplier $\mathbf{G}^{g^p,\widetilde{g^p}}_{m_p,\Lambda_p}$ instead of the filter ψ_p,

$$\left\|f_p - \overline{f^*_p}\right\|_2 \leq C^p(\epsilon_p + \epsilon)\|f_p\|_2 + \frac{\epsilon^*_p}{P}\|\widetilde{g^p}\|_2 \,. \tag{2.6.15}$$

Writing $\epsilon^*_P = \max_p \epsilon^*_p$ and $\|\widetilde{g^P}\|_2 = \max_p \|\widetilde{g^p}\|_2$, we can rewrite the estimate (2.6.8) as follows,

$$\left\|f - \sum_p \overline{f^*_p}\right\|_2 \leq C_\psi C_P(\epsilon_P + \epsilon)\|f\|_2 + \epsilon^*_P\|\widetilde{g^P}\|_2 \,. \tag{2.6.16}$$

As we are working with Gabor frames in the painless case, we can further specify the estimation without need to calculate the dual, as we know that $\|\widetilde{g}\|_2 \leq \frac{\|g\|_2}{A_p}$, for each p, where A_p is the lower frame bound. In Section 4.4, we provide examples of the reconstruction error obtained for given choice of the above functions.

2.6.3. Filter bank approach with nonstationary Gabor frames. We want now to extend the inequality (2.6.7) to the case of nonstationary Gabor frames. We consider a signal f^c bandlimited to the interval $[-\Omega/2, \Omega/2]$, such that

$$\|f - f^c\|_2 \leq \epsilon_\Omega \|f\|_2. \tag{2.6.17}$$

For each filtered version f^c_p we consider now P different NGTs $\{g^p_{k,l}\}$ with windows g^p_k compactly supported in time (painless case). Each one of the filtered versions can be written as follows,

$$f^c_p = \sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} z^p_{k,l}\widetilde{g^p_{k,l}} \,, \tag{2.6.18}$$

where $z^p_{k,l} = \langle f^c_p, M_{b_p k} T_{a_p l} g^p\rangle$ are the coefficients of the p-th analysis. As done before, we would like to approximate this function only with the relevant coefficients, those that correspond to the frequency support of f^c_p, meaning

$$f^c_p \approx \overline{f^c_p} = \sum_{k\in\mathbb{Z}}\sum_{l\in I^p_k} z^p_{k,l}\widetilde{g^p_{k,l}} \,, \tag{2.6.19}$$

where the sets I^p_k are finite and depend on the time index k. In the case of stationary Gabor frames, the sets I^p_k are the same for all $k \in \mathbb{Z}$. These sets store the ℓ indexes of

Gabor coefficients that correspond to the relevant frequency bands. Before specifying I_k^p we need to introduce a norm on the family of Gabor atoms. Define

$$S_0 = \{f \in L^2(\mathbb{R}^d) : \mathcal{V}_{g_0} f \in L^1(\mathbb{R}^{2d})\}; \tag{2.6.20}$$

here, $d = 1$, and $g_0 = \mathrm{e}^{-\pi\|t\|^2}$ is the Gaussian function; we observe that the S_0 norm $\|\cdot\|_{S_0}$ is the L^1 norm of $\mathcal{V}_\phi(\cdot)$ in $\mathbb{R}^2$, hence usual compactly-supported smooth functions belong to S_0 (a complete characterization of this space is given in [**Feichtinger and Strohmer, 1998**], Chapter 3). For a fixed family of functions $\{g_{k,l}^p\}_{k,l\in\mathbb{Z}}$ we define a norm $\|\cdot\|_V$ as

$$\|g^p\|_V = \max\left\{ \sup_{k\in\mathbb{Z}} \|g_k^p\|_{S_0} \ , \ \Big\| \sum_{k,l\in\mathbb{Z}} |\mathcal{V}_\phi g_{k,l}^p| \Big\|_\infty \right\} \ , \tag{2.6.21}$$

where $\phi \in S_0$; different ϕ give rise to equivalent norms. Let $\|g\|_V$ denote the maximum over all $\|g^p\|_V$, and similarly, $\|\widetilde{g}\|_V$ to be the maximum over all $\|\widetilde{g^p}\|_V$. To achieve a good reconstruction we assume that for every $g_k^p \in S_0$ there exists a band-limited approximation $h_k^p \in S_0$ such that

$$\|g^p - h^p\|_V \le \epsilon_p \|g^p\|_V. \tag{2.6.22}$$

Let ϵ_P be the maximum over all ϵ_p . Let $[\theta_{p,k}^1, \theta_{p,k}^2]$ denote the bandwidths of h_k^p. Define

$$I_k^p = \{l \in \mathbb{Z} \ : \ [\Omega_p^1, \Omega_p^2] \cap [\theta_{p,k}^1 + b_k l, \theta_{p,k}^2 + b_k l] \neq \emptyset\} \ , \tag{2.6.23}$$

that is the set of those ℓ for which the shifts of the essential bandwidth of g_k^p overlap the bandwidth of f_p^c. Then,

$$\|f_p^c - \overline{f_p^c}\|_2 \ \le \ \|g^p - h^p\|_V \|\widetilde{g^p}\|_V \|f_p^c\|_2 \ \le \ \epsilon_p \|g^p\|_V \|\widetilde{g^p}\|_V \|f_p^c\|_2 \ . \tag{2.6.24}$$

Therefore,

$$\begin{aligned} \|f^c - \sum_p \overline{f_p^c}\|_2 \ &\le \ \sum_p \|f_p^c - \overline{f_p^c}\|_2 \ \le \ \sum_p \epsilon_p \|g^p\|_V \|\widetilde{g^p}\|_V \|f_p^c\|_2 \\ &\le \ \epsilon_P \|g\|_V \|\widetilde{g}\|_V \sum_p \|f_p^c\|_2 \ \le \ 2\epsilon_P \|g\|_V \|\widetilde{g}\|_V \|f^c\|_2 \ , \end{aligned}$$

and

$$\|f - \sum_p \overline{f_p^c}\|_2 \le (2\epsilon_P \|g\|_V \|\widetilde{g}\|_V + \epsilon_\Omega) \|f\|_2. \tag{2.6.25}$$

CHAPTER 3

Entropy and sparsity measures

Chapter 2 has detailed the concept of adaptivity in sound analysis and synthesis, from a frame theory point of view. This one concerns the way that the adaptation is performed in our framework, the challenge being to realize an *automatic* process: our research is focused on models and tools for the local automatic adaptation of the atoms used in the decomposition of the signal. By defining appropriate measures to evaluate the local concentration of a given time-frequency representation, the sparsest analysis can be automatically achieved with less parameters to be specified; and most of all, without any a-priori knowledge of the signal properties.

The main point about adaptation is to understand what we are adapting to, and why: the concept of optimal resolution is highly signal- and application-dependent. This is, actually, the aspect that requires the highest care and expertise coming from the sound processing and musical worlds: we mainly need to define adaptation criteria matching the envisaged application, and to give formal definitions of the optimal time-frequency resolution we are interested in. We deduce such criteria from the optimization of specific *sparsity measures*. We take into account both theoretical and application-oriented sparsity measures: entropies and other quantities borrowed from information theory and probability belong to the first class; they provide the adaptive framework with a decisional structure whose mathematical properties are defined regardless of the specific application: hence we look for sufficiently flexible tools, whose parameters may be set in order for the measure to match the required concept of sparsity.

When dealing with real-world sounds, the characteristics of information measures may not find a direct interpretation in the signal domain. Human voice, for instance, has a periodic nature given by sequences of glottal pulses: depending on the application, we may be interested in privileging partials or pulses, thus considering as best representation the one where the desired component is more readable. These two cases determine different concepts of time-frequency concentration, which is not straightforward to express in terms of entropy-based sparsity measures. As an alternative to information measures, we give an application-driven definition of sparsity, depending on the particular features that the system should privilege: this measure is based on the classification of the sound components into sinusoids and noise, which is deduced from a time-frequency representation of the sound; disposing of several analyses, and given the ratio between the energies of sinusoids and noise according to this classification, we define the best analysis resolution to be the one maximizing this ratio. This choice

determines an adaptive model privileging analyses with a stronger sinusoidal content, designed to be optimal in most of the applications dealing with sinusoidal models.

3.1. Sparse problems and algorithms

Since the main goal of a representation is to increase the readability of the observed phenomenon, the concept of *sparsity* (see [**Gribonval and Nielsen, 2007**] and the related bibliography for a comprehensive survey) plays an important role: it concerns the efficiency of a given representation, when certain features have to be optimized. The definition of a sparsity measure depends on the problem: for sparse approximation, the optimal analysis is the one with the minimum number of coefficients still allowing an approximation of the original signal within a certain tolerated error. In other kind of *inverse problems* (see later in this section), instead of taking the minimal ℓ^0-norm of the analysis coefficients, sparsity is defined from ℓ^p-norms, as well as different types of measures. The main open problems are of three different orders: the choice of the analysis atoms, the quantity to be optimized among the possible representations, the efficiency of the algorithm to obtain the sparsest solution (see [**Tošić and and Frossard, 2011**] for a complete survey, and [**Tropp, 2004**] for several algorithms exploiting different approaches).

Let f be a signal in an Hilbert space $\mathcal{H}$, and Γ an index set; a dictionary $\mathcal{D}$ for $\mathcal{H}$ is a collection of vectors $\{\phi_\gamma\}_{\gamma\in\Gamma}$, called atoms, which spans the whole space, that is, every signal can be represented as a linear combination of atoms in the dictionary; $\mathcal{D}$ is *overcomplete* when its atoms are linearly dependent. The problem of representing f in a dictionary $\mathcal{D}$ of atoms may be approached decomposing f, through the usual scalar product $\langle f, \phi_\lambda\rangle$ in $\mathcal{H}$, with all of the atoms in the dictionary. A dual approach is to look for a vector of coefficients c such that f can be written as a linear combination of the atoms weighted by the coefficients,

$$f = \sum_{\gamma\in\Gamma} c_\gamma \phi_\gamma \ . \tag{3.1.1}$$

In this formulation, the signal f has to be perfectly reconstructed by means of a linear combination of the atoms: the representation is sparse if the number of non-zero coefficients $\|c\|_0$ in the expansion is much smaller than the dimension of f in $\mathcal{H}$. The general *sparse representation* problem is defined as follows,

$$\min_c \|c\|_0 \ s.t. \ \left\| f - \sum_{\gamma\in\Gamma} c_\gamma \phi_\gamma \right\|_2 \leq \epsilon \ , \tag{3.1.2}$$

where the target is to minimize the number of atoms keeping the reconstruction error ϵ small. This problem is known to be NP-hard, but there exist several polynomial time algorithms which give suboptimal solutions: a first possible relaxation consists in the iterative selection of appropriate atoms from the dictionary, until the desired precision is reached; the Matching Pursuit, as well as its orthogonal variation, and the Basis Pursuit belong to this class of algorithms, which are known as *greedy* (see

[Davis et al., 1997, Chen, 1998]).

A different strategy is the convex relaxation of the ℓ^0 minimization, adopted by algorithms like LASSO (see [Tibshirani, 1994]), solving the problem

$$\min_c \left(\left\| f - \sum_{\gamma \in \Gamma} c_\gamma \phi_\gamma \right\|_2 + \lambda \|c\|_1 \right) , \tag{3.1.3}$$

where λ is a regularization parameter: its value determines if the optimal solution privileges the sparsity or the quality of the reconstruction. If we a replace the ℓ^1 norm with the ℓ^p one, for $p > 0$, we obtain a general form for the *sparse approximation* problem: when $0 < p < 1$ the norm is not convex, therefore local minima are searched; when $p > 1$, the problem admits a global optimum. There exist a further form of the minimization (3.1.3), used to solve inverse problems: here, the original signal f is unknown, and the target is to obtain its sparse approximation disposing of the output f^* of a known operator A, with $\mathrm{A}f = f^*$. The problem takes the following form,

$$\min_h \left(\left\| \mathrm{A}h - f^* \right\|_2 + \lambda \|h\|_p \right) . \tag{3.1.4}$$

The main assumption, here, is that sparse solutions guarantee the unicity of the representation, so that $(\mathrm{A}h_1 = \mathrm{A}h_2) \Rightarrow (h_1 = h_2)$; in this sense, if f admits a sparse representation within the dictionary, then the solution h is a sparse approximation of f.

There is a large range of applications taking advantage of sparse representation, including relevant up-to-date topics in music industry or telecommunications (see [Plumbley et al., 2010]). We formulate our automatic selection of the best local resolution in terms of sparsity: the dictionary we use is composed of time-frequency shifts of a finite number of window functions; these windows are obtained as scalings of a same window. The solutions we look for are highly structured, as we adopt STFT-based reconstruction formulas: this implies that the choice of a certain atom, at a given time-frequency location, influences the effect of a large number of close atoms. For this reason, even if we define a sparsity criterium, our problem would not benefit of the sparse algorithms machinery to find optima: a direct comparison of the sparsity of all the finite possible solutions remains efficient.

Instead of minimizing the ℓ^p norm of the coefficients vector, we use entropy-based and application-oriented measures (see Section 3.2 and 3.7), whose properties are well-suited to the kind of sparsity we look for: in our framework, a representation is sparse if the *elementary* components of the analyzed signal are resolved within the representation. The concept of elementary signal is mathematically ill-posed, as it includes classes whose characteristics may be completely different: sinusoids as well as instantaneous events, for instance. On the other hand, most of the sound processing techniques we deal with have a precise operational definition of elementary signal: the main interest is thus to define a sparsity measure which reflects the needs of sound spectral processing, using a flexible mathematical framework.

3.2. Rényi entropies as sparsity measures

We first restate some definitions given in Chapter 1 and 2, to introduce a more convenient notation for the proofs in the following. Given a window function $g \neq 0$, the STFT of a function f with respect to g is defined as follows,

$$\mathcal{V}_g f(t,\omega) = \langle f, g_{t,\omega} \rangle = \int_{\mathbb{R}^d} f(s)\overline{g(s-t)}\, \mathrm{e}^{-2\pi \mathrm{i}\omega s}\, ds \,,$$

while the spectrogram of f with window g is the squared modulus of the STFT,

$$\mathrm{PS}_g f(t,\omega) = |\mathcal{V}_g f(t,\omega)|^2 = |\langle f, g_{t,\omega}\rangle|^2 \,. \tag{3.2.1}$$

Given a set $\Gamma \subset \mathbb{Z}^{2d}$, a sequence $\{g_\gamma\}_{\gamma\in\Gamma} \subset L^2(\mathbb{R}^d)$ is a frame for $L^2(\mathbb{R}^d)$ if there are two constants $0 < A_\Gamma \leq B_\Gamma$, called *frame bounds*, such that for every $f \in L^2(\mathbb{R}^d)$,

$$A_\Gamma \|f\|_{L^2}^2 \leq \sum_{\gamma\in\Gamma} |\langle f, g_\gamma\rangle|^2 \leq B_\Gamma \|f\|_{L^2}^2 \tag{3.2.2}$$

In the following we shall consider only sets Γ which are lattices, and in particular the lattices $\Lambda_{a,b} = a\mathbb{Z}^d \times b\mathbb{Z}^d$, for $a, b > 0$; these systems are called Gabor systems, and we shall indicate them with $\mathbf{G}(g, a, b)$. A Gabor system defines a discrete version of the STFT and the spectrogram of a signal, as sampling of their continuous versions; so a discrete spectrogram is given by

$$\mathrm{PS}_g f[n,k] = |\mathcal{V}_g f(an, bk)|^2 \,, \tag{3.2.3}$$

and we omit to indicate the window g if ambiguity does not occur.

With an appropriate normalization both the continuous and discrete spectrogram can be interpreted as probability densities. Thanks to this interpretation, some techniques belonging to the domains of probability and information theory can be applied to our problem: in particular, the concept of entropy can be extended to give a sparsity measure of a time-frequency density. The approach we adopt (see [**Baraniuk et al., 2001**] for the original formulation) takes into account Rényi entropies, a generalization of the Shannon entropy: the application to our problem is related to the concept that minimizing the complexity, or information, of a set of time-frequency representations of a same signal, is equivalent to maximizing the concentration, peakiness, and therefore the sparsity of the analysis. Thus we consider as *best* analysis the sparsest one, according to the minimal entropy evaluation.

Definition 3.2.1. Given a finite discrete probability density $P = (P_1, ..., P_N)$ and a real number $\alpha \geq 0,\ \alpha \neq 0$, the *Rényi entropy* of P is defined as follows,

$$\mathrm{H}_\alpha[P] = \frac{1}{1-\alpha} \log_2 \sum_{k=1}^{N} P_k^\alpha \,, \tag{3.2.4}$$

where P is in square brackets to indicate that discrete densities are considered.

Among the general properties of Rényi entropies (see [Rényi, 1961], [Beck and Schlögl, 1993] and [Zyczkowski, 2003]) we recall in particular those directly related with our problem. It is easy to show that for every finite discrete probability density P the entropy $\mathrm{H}_\alpha[P]$ tends to the Shannon entropy of P as the order α tends to one. Moreover, $\mathrm{H}_\alpha[P]$ is a non increasing function of α, so

$$\alpha_1 < \alpha_2 \Rightarrow \mathrm{H}_{\alpha_1}[P] \geq \mathrm{H}_{\alpha_2}[P] \,. \tag{3.2.5}$$

When working with finite discrete densities, the case $\alpha = 0$ can also be considered, which simply gives the logarithm of the number of elements in P; as a consequence $\mathrm{H}_0[P] \geq \mathrm{H}_\alpha[P]$ for every admissible order α.
A third basic fact is that for every order α the Rényi entropy H_α is maximum when P is uniformly distributed, while it is minimum and equal to zero when P has a single non-zero value.

All of these results give useful information on the values of different measures on a single density P, while the relations between the entropies of two different densities P and Q are in general hard to determine analytically; in our problem, P and Q are two spectrograms of a same signal, based on two window functions with different scaling as in equation (2.2.1). Therefore, we first need to extend the entropy definition to continuous densities, and in particular to the spectrogram.

Definition 3.2.2. Given a signal f and its spectrogram $\mathrm{PS}f$ as in equation (3.2.1), the *Rényi entropy* of $\mathrm{PS}f$ is defined for an *order* $\alpha > 0,\ \alpha \neq 1$ as follows,

$$\mathrm{H}^R_\alpha(\mathrm{PS}f) = \frac{1}{1-\alpha}\ \log_2 \iint_R \left(\frac{\mathrm{PS}f(t,\omega)}{\iint_R \mathrm{PS}f(t',\omega')\mathrm{d}t'\mathrm{d}\omega'} \right)^\alpha \mathrm{d}t\mathrm{d}\omega \,, \tag{3.2.6}$$

where $R \subseteq \mathbb{R}^{2d}$ and we omit its indication if equality holds.

In general terms, our problem can be written as follows,

$$\min_{s \in S} \mathrm{H}_\alpha(\mathrm{PS}_s f) \tag{3.2.7}$$

where S is a certain set of indexes for the window functions g_s, and $\mathrm{PS}_s f$ is the spectrogram of f with window g_s. The optimal choice of g depends on the signal f, and the search for analytical solutions would imply limitations on the signal domain. We give in the following subsection a simple example where this is achievable, but in general we are not interested in the analytical solutions of the problem: we rather focus our investigation on the solutions provided by the algorithm we have developed, to verify that the optimal choice determined by the measure gives the desirable resolution in terms of sound processing. This is one of the reason why the Rényi entropies are considered: by the dependence on the order α, they constitute a class of different sparsity measures, determining a particular concept of sparsity for each value of α. We thus can refine the choice of the best solution depending on the specific application requirements, keeping the framework unaltered. The relation between the solutions to problem (3.2.7) and the entropy order α is detailed in Section 3.4.

3.2.1. Best window for stationary sinusoids. For some basic stationary signals, it is possible to find the solution to the problem (3.2.7) analytically: for example, let f be a complex stationary sinusoid $f(t) = \mathrm{e}^{2\pi \mathrm{i} \omega_0 t}$ and g a window function of compact support; then

$$\mathcal{V}_g f(t,\omega) = \mathrm{e}^{-2\pi \mathrm{i}(\omega-\omega_0)t} \cdot \widehat{g}(\omega-\omega_0) \ , \ \mathrm{PS}f(t,\omega) = |\widehat{g}(\omega-\omega_0)|^2 \ ,$$

and $\mathrm{PS}f$ is therefore time-independent. We choose now a bounded set S of positive scaling factors; the spectrogram $\mathrm{PS}_s f$ taken with a scaled window g_s (see (2.2.1)) is thus given by

$$\mathrm{PS}_s f(t,\omega) = s \cdot |\widehat{g}(s \cdot (\omega-\omega_0))|^2 \ ,$$

therefore the following relation holds for every $s \in S$,

$$\mathrm{H}_\alpha(\mathrm{PS}_s f) = \mathrm{H}_\alpha(\mathrm{PS}f) - \log_2 s \ . \tag{3.2.8}$$

The solution to the problem (3.2.7) is given by the window minimizing the entropy measure: we deduce from equation (3.2.8) that it is the one obtained with the largest scaling factor available, therefore with the largest time-support. This is coherent with our expectation: the information of a stationary signal, such as a sinusoid, is completely determined by its frequency spectrum, which is time-independent, and is thus best represented with the highest possible frequency resolution. Moreover, this is true for any order α used for the entropy calculus. Symmetric considerations apply whenever the spectrogram of a signal does not depend on frequency, as for impulses.

3.3. Rényi entropy measures of a spectrogram

We now look closer to the problem of the existence of the measure defined in (3.2.6) with regard to the signal f, the window g and the order α. Our results about the STFT and the spectrogram complete the ones presented in [**Baraniuk et al., 2001**], where only integer values of α are considered. For the class of signals and windows typically considered in real-world applications, by the regularity of the STFT operator, which is investigated in Subsection 3.3.1, we see that the Rényi entropy of a spectrogram is well-defined for every $\alpha \geq \frac{1}{2}$. As we have to deal with discrete spectrograms, we also have to define a discrete version of the measure in (3.2.6), and find the dependance of the discretized measure on the sampling procedure: when comparing the entropies of discrete spectrograms with different sampling lattices, this aspect is fundamental to understand if the comparison makes sense. Therefore, we prove as well some results about the convergence of the discrete Rényi entropies of a spectrogram: we show in Subsections 3.3.3 and 3.3.4 that as the sampling grid increases its density, the discrete entropy converges to its continuos version (3.2.6); we obtain similar results for the Shannon entropy, which is the limit case when α tends to 1.

The proofs of these results are based on frame theory, and we first introduce an important class of window functions: they are useful to verify the existence of Gabor frames and to investigate the properties of discretized STFTs with varying lattices. A

function $g \in L^\infty(\mathbb{R}^d)$ belongs to the *Wiener amalgam space* $W = W(\mathbb{R}^d)$ if

$$\|g\|_W = \sum_{k\in\mathbb{Z}^d} \operatorname{ess\,sup}_{t\in Q} |g(t+k)| < \infty \,, \tag{3.3.1}$$

where $Q = [0,1]^d$ is the unit cube. Bounded functions with compact support belong to W, which is therefore a dense subspace of $L^p(\mathbb{R}^d), 1 \le p < \infty$; in particular, S_0 (introduced in Subsection 2.6.3) is a proper subspace of W.

The basic idea is, if the chosen window g is sufficiently regular, then $\mathbf{G}(g,a,b)$ is a Gabor frame whenever $ab > 0$ is small enough. To formally state this general concept we consider the following theorem (by Walnut, [**Walnut, 1992**]), whose proof is based on the properties of the correlation function, defined as

$$G_n^{(a,b)}(t) = \sum_{k\in\mathbb{Z}^d} \overline{g}(t - \frac{n}{b} - ak) g(t-ak) \,. \tag{3.3.2}$$

Theorem 3.3.1. *Suppose that $g \in W$ and that $a > 0$ is chosen such that for constants $C, D > 0$*

$$C \le \sum_{k\in\mathbb{Z}^d} |g(t-ak)|^2 \le D < \infty \ a.e. \tag{3.3.3}$$

Then there exist a value $b_0 = b_0(a) > 0$ such that $\mathbf{G}(g,a,b)$ is a Gabor frame for all $b \le b_0$. Moreover, b_0 can be chosen such that $\mathbf{G}(g,a,b)$ is a frame for all $b \le b_0$ with frame bounds

$$A = b^{-d}\left(C - \sum_{n\neq 0} \|G_n^{(a,b)}\|_\infty\right) \tag{3.3.4}$$

and

$$B = b^{-d} \sum_{n\in\mathbb{Z}^d} \|G_n^{(a,b)}\|_\infty \tag{3.3.5}$$

where $G_n^{(a,b)}$ is the correlation function defined in (3.3.2)

The hypotheses of Theorem 3.3.1 are satisfied by the windows used in most part of the applications: for example, if $|g(t)| \ge c > 0$ on a cube $t_0 + Q_{a_0} = t_0 + [0,a_0]^d$ for some $t_0 \in \mathbb{R}^d$, then condition (3.3.3) is verified for every a with $0 < a < a_0$. In what follows, we need the following corollary of this theorem.

Corollary 3.3.2. *Suppose that $g \in W$ satisfies the hypotheses of Theorem 3.3.1 for every $a < a'$, $a' \in (0,1]$; then there exists a positive constant c such that for every $0 < a \le a'$ the system $\mathbf{G}(g,a,b)$ is a Gabor frame for $0 < b \le b_0(a)$, whose upper frame bound B_{ab} verifies*

$$(ab)^d B_{ab} \le c \,. \tag{3.3.6}$$

PROOF. By equation (3.3.5) we have that $B_{ab} = b^{-d}\sum_{n\in\mathbb{Z}^d} \|G_n^{(a,b)}\|_\infty$, and by the properties of the correlation function $G_n^{(a,b)}$ ([**Gröchenig, 2001a**, Lemma 6.3.1]) we have

$$\sum_{n\in\mathbb{Z}^d} \|G_n^{(a,b)}\|_\infty \le \left(\frac{1}{a}+1\right)^d (2b+2)^d \|g\|_W^2 \,, \tag{3.3.7}$$

so that

$$(ab)^d B_{ab} \le (1+a)^d (2b+2)^d \|g\|_W^2 \le c\,, \tag{3.3.8}$$

with $c = 2^{3d}\|g\|_W^2$. □

Given a window g which satisfies the hypotheses of Corollary 3.3.2, we denote as Ψ_g the surface of $\mathbb{R}^2$ obtained as follows,

$$\Psi_g = \prod_{0<a\le a'} (0, b_0(a)]\,, \tag{3.3.9}$$

and we see that $(0,0)$ belongs to the closure $\overline{\Psi}_g$ of this set.

3.3.1. Regularity of $\mathcal{V}$. We denote by $UC_b(\mathbb{R}^d)$ the space of bounded uniformly continuous functions defined on $\mathbb{R}^d$.

Lemma 3.3.3. *Let $1 \le p \le \infty$ and consider $f \in L^p(\mathbb{R}^d)$, $g \in L^q(\mathbb{R}^d)$, where q is the Hölder conjugate exponent of p. Then $\mathcal{V}_g f \in UC_b(\mathbb{R}^{2d})$ and*

$$\|\mathcal{V}_g f\|_\infty \le \|f\|_{L^p}\|g\|_{L^q}. \tag{3.3.10}$$

PROOF. The inequality follows easily from the Hölder inequality ([Brezis, 1983, Theorem IV.6]). We prove that $\mathcal{V}_g f$ is uniformly continuous. Fix (t_1,ω_1) and (t_2,ω_2) and set $\tau = t_2 - t_1$ and $\theta = \omega_2 - \omega_1$; let $1 < p < \infty$, then using again the Hölder inequality,

$$\begin{aligned}
|\mathcal{V}_g f(t_2,\omega_2) - \mathcal{V}_g f(t_1,\omega_1)| &\le \Big|\int_{\mathbb{R}^d} f(s)\overline{g(s-t_2)}\big(\mathrm{e}^{-2\pi \mathrm{i}\omega_2\cdot s} - \mathrm{e}^{-2\pi\mathrm{i}\omega_1\cdot s}\big)\,ds\Big| + \\
&\quad + \Big|\int_{\mathbb{R}^d} f(s)\big(\overline{g(s-t_2)} - \overline{g(s-t_1)}\big)\,\mathrm{e}^{-2\pi\mathrm{i}\omega_1\cdot s}\,ds\Big| \\
&\le \|g\|_{L^q}\Big(\int_{\mathbb{R}^d} |f(s)|^p\,|\,\mathrm{e}^{2\pi\mathrm{i}\theta\cdot s} - 1|^p\,ds\Big)^{\frac{1}{p}} + \\
&\quad + \|f\|_{L^p}\Big(\int_{\mathbb{R}^d} |g(s-\tau) - g(s)|^q\,ds\Big)^{\frac{1}{q}}.
\end{aligned} \tag{3.3.11}$$

The right–hand side of the inequality above depends only on τ, θ and not explicitly on (t_1,ω_1) and (t_2,ω_2), hence it is sufficient to show that it converges to 0 as $\tau \to 0$ and $\theta \to 0$. For the first term of the right–hand side this follows from Lebesgue's theorem, for the second it follows from Beppo Levi's theorem [Brezis, 1983, Theorem IV.2 and IV.1, respectively].
For $p = 1$ and $p = \infty$, we develop the right–hand side of (3.3.11) similarly, and the conclusion follows, respectively, by the fact the $\operatorname{ess\,sup}_{s\in\mathbb{R}^d}(g(s-\tau) - g(s))$ tends to 0 as τ tends to 0, and $\operatorname{ess\,sup}_{s\in\mathbb{R}^d}(f(s)(\mathrm{e}^{2\pi\mathrm{i}\theta\cdot s} - 1))$ tends to 0 as θ tends to 0. □

Lemma 3.3.4. *Let $1 \le p \le \infty$ and consider $f \in L^p(\mathbb{R}^d)$, $g \in L^q(\mathbb{R}^d)$, where q is the Hölder conjugate exponent of p. Then $\mathcal{V}_g f \in L^{\max(p,q)}(\mathbb{R}^{2d})$ and there exists a constant $c_p > 0$ such that*

$$\|\mathcal{V}_g f\|_{L^{\max(p,q)}} \le c_p \|f\|_{L^p}\|g\|_{L^q}. \tag{3.3.12}$$

In particular, if $p = q = 2$,

$$\|\mathcal{V}_g f\|_{L^2} = \|f\|_{L^2}\|g\|_{L^2} \tag{3.3.13}$$

Finally, $\mathcal{V}_g f \in L^{p'}(\mathbb{R}^{2d})$ *for every* $p' \in [\max(p,q), \infty]$.

PROOF. Assume first $1 < p < 2$, then $\max(p,q) = q$ and, since $\mathcal{V}_g f(t,\omega) = \widehat{f(\cdot)\overline{g(\cdot - t)}}(\omega)$, where $\hat{\cdot}$ denotes the Fourier transform, the sharp Hausdorff-Young inequality (see [Babenko, 1962, Beckner, 1975]) yields

$$\|V_g f\|_{L^q}^q = \int_{\mathbb{R}^d} \|\widehat{f(\cdot)\overline{g(\cdot - t)}}\|_{L^q}^q \, dt \leq c_p \int_{\mathbb{R}^d} dt \Big(\int_{\mathbb{R}^d} |f(s)|^p |g(s-t)|^p \, ds \Big)^{\frac{q}{p}}, \tag{3.3.14}$$

and, using the Hölder inequality with exponents q/p and $q/(q-p)$,

$$\|V_g f\|_{L^q}^q \leq \|f\|_{L^p}^{q-p} \int_{\mathbb{R}^d} dt \int_{\mathbb{R}^d} |f(s)|^p |g(s-t)|^q \, ds \leq \|g\|_{L^q}^q \|f\|_{L^p}^q \tag{3.3.15}$$

If $2 < p < \infty$ the inequality follows from the previous computations and the fact that $\mathcal{V}_g f(t,\omega) = \mathrm{e}^{-2\pi \mathrm{i}\omega t} \overline{\mathcal{V}_f g(-t,-\omega)}$. If $p = 1$ or $p = \infty$, the inequality is straightforward, and if $p = q = 2$ the identity follows easily from Plancherel identity (see [Gröchenig, 2001b], Theorem 1.1.2). The last statement of the lemma is an immediate consequence of interpolation and the previous lemma. □

Example 3.3.5. When $p < 2$, the hypotheses of Lemma 3.3.4 are not sufficient to guarantee $V_g f \in L^p(\mathbb{R}^{2d})$. As an example, take $f(x) = \mathbb{1}_{[0,1]}(x)$ the indicator function of the interval $[0,1] \subset \mathbb{R}$, and $g(x) = \mathbf{1}(x)$ the constant one function. Then, $f \in L^1(\mathbb{R})$ and $g \in L^\infty(\mathbb{R})$, but $\mathcal{V}_g f(t,\omega) = -\mathrm{e}^{-2\pi \mathrm{i}\omega}$, and for any fixed $t \in \mathbb{R}$ the function $F(\omega) = \mathcal{V}_g f(t,\omega)$ is not in $L^1(\mathbb{R})$. Thus we conclude that $\mathcal{V}_g f \notin L^1(\mathbb{R}^2)$. ■

3.3.2. Convergence of the sampled entropies. Given $\alpha > 0$, with $\alpha \neq 1$, we have defined in equation (3.2.6) the Rényi entropy of a spectrogram,

$$\mathrm{H}_\alpha(\mathrm{PS}f) = \frac{1}{1-\alpha} \log \iint \Big(\frac{|\mathcal{V}_g f(t,\omega)|^2}{\iint |\mathcal{V}_g f(t',\omega')|^2 \, dt' \, d\omega'} \Big)^\alpha \, dt \, d\omega$$

and we write as well the Shannon entropy of a spectrogram,

$$\mathrm{H}_1(\mathrm{PS}f) = \iint \phi\Big(\frac{|\mathcal{V}_g f(t,\omega)|^2}{\iint |\mathcal{V}_g f(t',\omega')|^2 \, dt' \, d\omega'} \Big) \, dt \, d\omega, \tag{3.3.16}$$

where $\phi(t) = -t \log t$. Fix a (discrete) lattice $\Lambda \subset \mathbb{R}^{2d}$ and consider the sampled entropy

$$\mathrm{H}_\alpha^\Lambda[\mathrm{PS}f] = \frac{1}{1-\alpha} \log \sum_{(t,\omega)\in\Lambda} \Big(\frac{|\mathcal{V}_g f(t,\omega)|^2}{\sum_{(t',\omega')\in\Lambda} |\mathcal{V}_g f(t',\omega')|^2} \Big)^\alpha + d \log(ab) \, , \tag{3.3.17}$$

for $\alpha \neq 1$, and

$$\mathrm{H}_1^\Lambda[\mathrm{PS}f] = \sum_{(t,\omega)\in\Lambda} \phi\Big(\frac{|\mathcal{V}_g f(t,\omega)|^2}{\sum_{(t',\omega')\in\Lambda} |\mathcal{V}_g f(t',\omega')|^2} \Big), \tag{3.3.18}$$

for $\alpha = 1$, where ϕ is defined as above. The entropies H_α, $\mathrm{H}_\alpha^\Lambda$ can be reformulated, at least for $\alpha \neq 1$, in terms of functional norms. To this end, given $a, b > 0$, define for a function $h : \mathbb{R}^{2d} \to \mathbb{R}^d$ the function $(h)_{a,b}$ as

$$(h)_{a,b}(t,\omega) = \sum_{k,n\in\mathbb{Z}^d} h(ka, nb)\mathbb{1}_{k,n}^{a,b}(t,\omega), \tag{3.3.19}$$

where $\mathbb{1}_{k,n}^{a,b} = \mathbb{1}_{E_{k,n}(a,b)}$ is the indicator function of the cube

$$E_{k,n}(a,b) = \big\{(t,\omega) \in \mathbb{R}^d \times \mathbb{R}^d : |t_i - k_i a| \leq \frac{a}{2},\ |\omega_i - n_i b| \leq \frac{b}{2},\ i = 1,\dots,d\big\}.$$

It is easy to see that for $p \geq 1$,

$$\sum_{k,n\in\mathbb{Z}^d} |h(ak, bn)|^p = (ab)^{-d}\|(h)_{a,b}\|_{L^p}^p\,. \tag{3.3.20}$$

For $\alpha \neq 1$, it is easy to verify that

$$\mathrm{H}_\alpha(\mathrm{PS}f) = \frac{1}{1-\alpha}\log\Big(\frac{\|\mathcal{V}_g f\|_{L^{2\alpha}}}{\|\mathcal{V}_g f\|_{L^2}}\Big)^{2\alpha},$$

and

$$\mathrm{H}_\alpha^{\Lambda_{a,b}}[\mathrm{PS}f] = \frac{1}{1-\alpha}\log\Big((ab)^{d(1-\alpha)}\frac{\|(\mathcal{V}_g f)_{a,b}\|_{L^{2\alpha}}^{2\alpha}}{\|(\mathcal{V}_g f)_{a,b}\|_{L^2}^{2\alpha}}\Big)\,,$$

where $\mathrm{H}_\alpha^{a,b}[\mathrm{PS}f] = \mathrm{H}_\alpha^{\Lambda_{a,b}}[\mathrm{PS}f]$. It turns out that, in order to prove that the discrete entropy $\mathrm{H}_\alpha^{a,b}[\mathrm{PS}f]$ is convergent to $\mathrm{H}_\alpha(\mathrm{PS}f)$, it is sufficient to show that $(\mathcal{V}_g f)_{a,b}$ converges to $\mathcal{V}_g f$ in $L^2(\mathbb{R}^{2d})$ and in $L^{2\alpha}(\mathbb{R}^{2d})$. We investigate first, in the following lemma, the convergence in $L^2(\mathbb{R}^{2d})$ and we postpone to the following sections the convergence in $L^{2\alpha}(\mathbb{R}^{2d})$.

Lemma 3.3.6. *If $g \in W$ and $(a,b) \in \Psi_g$, then for every $f \in L^2(\mathbb{R}^d)$, $\|(\mathcal{V}_g f)_{a,b}\|_{L^2} \to \|\mathcal{V}_g f\|_{L^2}$ as $(a,b) \to (0,0)$ within Ψ_g .*

PROOF. By proceeding as in [Sun, 2010, Lemma 2.5], we get that if $\varphi \in C_c^\infty(\mathbb{R}^d)$, then for every $(a,b) \in \Psi_g$,

$$\|\mathcal{V}_g\varphi - (\mathcal{V}_g\varphi)_{a,b}\|_{L^2} \leq c_1(a+b)\|\varphi\|_\star,$$

where $\|\varphi\|_\star^2 = \sum_{\alpha,\beta\in\{0,1\}^d}\|x^\alpha D^\beta\varphi\|_{L^2}$. Let $f \in L^2(\mathbb{R}^d)$ and fix $\epsilon > 0$. Let $\varphi \in C_c^\infty(\mathbb{R}^d)$ be such that $\|f - \varphi\|_{L^2} \leq \epsilon$, then $\|\mathcal{V}_g f - \mathcal{V}_g\varphi\|_{L^2} = \|g\|_{L^2}\|f - \varphi\|_{L^2} \leq \|g\|_{L^2}\epsilon$ and, by (3.3.20), the frame inequality (3.2.2) and (3.3.6),

$$\|(\mathcal{V}_g f)_{a,b} - (\mathcal{V}_g\varphi)_{a,b}\|_{L^2} = \|(\mathcal{V}_g(f-\varphi))_{a,b}\|_{L^2} \leq (ab)^{\frac{d}{2}}B_{a,b}^{\frac{1}{2}}\|f - \varphi\|_{L^2} \leq c_2\epsilon.$$

In conclusion

$$\begin{aligned}\|\mathcal{V}_g f - (\mathcal{V}_g f)_{a,b}\|_{L^2} \leq \\ \leq \|\mathcal{V}_g f - \mathcal{V}_g\varphi\|_{L^2} + \|\mathcal{V}_g\varphi - (\mathcal{V}_g\varphi)_{a,b}\|_{L^2} + \|(\mathcal{V}_g\varphi)_{a,b} - (\mathcal{V}_g f)_{a,b}\|_{L^2} \leq \\ \leq \big(c_2 + \|g\|_{L^2}\big)\epsilon + c_1(a+b)\|\varphi\|_\star,\end{aligned}$$

and in the limit $a, b \to 0$, $\limsup\|\mathcal{V}_g f - (\mathcal{V}_g f)_{a,b}\|_{L^2} \leq (c_2 + \|g\|_{L^2})\epsilon$. By choosing ϵ arbitrarily small, the lemma follows. □

3.3.3. The case $\alpha > 1$.

Proposition 3.3.7. *If $\alpha > 1$, $g \in W$ and $(a,b) \in \Psi_g$, then for every $f \in L^2(\mathbb{R}^d)$ the Rényi entropy $\mathrm{H}_\alpha(\mathrm{PS}f)$ and its discrete version $\mathrm{H}_\alpha^{a,b}[\mathrm{PS}f]$ on the lattice $\Lambda_{a,b} = a\mathbb{Z}^d \times b\mathbb{Z}^d$ are finite. Moreover, $\mathrm{H}_\alpha^{a,b}[\mathrm{PS}f] \to \mathrm{H}_\alpha(\mathrm{PS}f)$ as $(a,b) \to (0,0)$ within Ψ_g .*

PROOF. The continuous entropy $\mathrm{H}_\alpha(\mathrm{PS}f)$ is well–defined by Lemma 3.3.4. The discrete entropy $\mathrm{H}_\alpha^{\Lambda_{a,b}}[\mathrm{PS}f]$ is well–defined by the frame inequality (3.2.2), indeed, for $p = 2\alpha > 2$,

$$\sum_{k,n\in\mathbb{Z}^d} |\mathcal{V}_g f(ka,nb)|^p \le \Big(\sum_{k,n\in\mathbb{Z}^d} |\mathcal{V}_g f(ka,nb)|^2 \Big)^{\frac{p}{2}} \le B_{a,b}^{\frac{p}{2}} \|\mathcal{V}_g f\|_{L^2}^p,$$

where $B_{a,b} = B_{\Lambda_{a,b}}$.

We prove the convergence of $(\mathcal{V}_g f)_{a,b}$ to $\mathcal{V}_g f$ in $L^p(\mathbb{R}^d)$. We know by Lemma 3.3.3 that $\mathcal{V}_g f \in UC_b(\mathbb{R}^{2d})$, hence $|\mathcal{V}_g f(t_1,\omega_1) - \mathcal{V}_g f(t_2,\omega_2)| \le w(a+b)$ if $|t_1 - t_2| \le a$ and $|\omega_1 - \omega_2| \le b$, where w is the (uniform) modulus of continuity of $\mathcal{V}_g f$. With this position,

$$\begin{aligned}
\|(\mathcal{V}_g f)_{a,b} - \mathcal{V}_g f\|_{L^p}^p &\le \iint \sum_{k,n\in\mathbb{Z}^d} |\mathcal{V}_g f(ka,nb) - \mathcal{V}_g f(t,\omega)|^p \mathbb{1}_{k,n}^{a,b}(t,\omega)\, dt\, d\omega \\
&\le w^{p-2}(a+b)^{p-2} \iint \sum_{k,n\in\mathbb{Z}^d} |\mathcal{V}_g f(ka,nb) - \mathcal{V}_g f(t,\omega)|^2 \mathbb{1}_{k,n}^{a,b}(t,\omega)\, dt\, d\omega \\
&\le 2w^{p-2}(a+b)^{p-2} \Big(\sum_{k,n\in\mathbb{Z}^d} (ab)^d |\mathcal{V}_g f(ka,nb)|^2 + \|\mathcal{V}_g f\|_{L^2}^2 \Big) \\
&\le c\|f\|_{L^2}^2 \|g\|_{L^2}^2 w^{p-2}(a+b)^{p-2},
\end{aligned}$$

where we have used Lemma 3.3.4, the frame inequality (3.2.2) and the fact that $(ab)^d B_{a,b}$ is uniformly bounded (see the inequality (3.3.6)). □

3.3.4. The case $\frac{1}{2} \le \alpha < 1$. We now recall the definitions of two fundamental function spaces; considering the Schwartz space $\mathcal{S}$ and its dual $\mathcal{S}'$, if $g \in \mathcal{S}$, the *modulation space* $M_m^{p,q}(\mathbb{R}^d)$ is given by

$$M_m^{p,q}(\mathbb{R}^d) = \{f \in \mathcal{S}'(\mathbb{R}^d) : \mathcal{V}_g f \in L_m^{p,q}(\mathbb{R}^{2d})\};$$

the *amalgam space* $W(L_m^{p,q}(\mathbb{R}^{2d}))$ is defined as

$$W(L_m^{p,q}(\mathbb{R}^{2d})) = \{F \in L_{\mathrm{loc}}^\infty(\mathbb{R}^{2d}) : \sum_{k\in\mathbb{Z}^d} \Big(\sum_{n\in\mathbb{Z}^d} a_{nk}(F)^p m(n,k)^p \Big)^{\frac{q}{p}} < \infty\},$$

where $a_{nk}(F) = \operatorname{ess\,sup}_{(t,\omega)\in[-\frac{1}{2},\frac{1}{2})^{2d}+(n,k)} |F(t,\omega)|$ (see [Gröchenig, 2001a] for further details); in both the definitions, we write only p if $p = q$.

Proposition 3.3.8. *If $\alpha \in [\frac{1}{2}, 1)$, if $g \in S_0$ and $(a,b) \in \Psi_g$, then for every $f \in M_1^{2\alpha}$ the Rényi entropy $\mathrm{H}_\alpha(f)$ and its discrete version $\mathrm{H}_\alpha^{a,b}$ on the lattice $\Lambda_{a,b} = a\mathbb{Z}^d \times b\mathbb{Z}^d$ are finite. Moreover, $\mathrm{H}_\alpha^{a,b} \to \mathrm{H}_\alpha(f)$ as $(a,b) \to (0,0)$ within Ψ_g .*

Before proving the theorem, we give an elementary result which, together with the regularity properties of $\mathcal{V}_g$, will prove the above theorem.

Lemma 3.3.9. *Let $p \geq 1$ and $h \in W(L^p_1(\mathbb{R}^{2d})) \cap UC(\mathbb{R}^{2d})$, then*

$$\|(h)_{a,b} - h\|_{L^p(\mathbb{R}^{2d})} \longrightarrow 0,$$

as $(a,b) \to (0,0)$.

PROOF. We prove the lemma in two steps.

Step 1. Assume that $h \in UC(\mathbb{R}^{2d})$ has compact support, then

$$|(h)_{a,b}(t,\omega) - h(t,\omega)| \leq \sum_{n,k \in \mathbb{Z}^d} |h(t_k,\omega_n) - h(t,\omega)| \mathbb{1}_{[k,k+1)\times[n,n+1)}(t/a, \omega/b),$$

and by uniform continuity the above term is small when $a \vee b$ is small. The compact support ensures that the sum is over a finite number of indices. In conclusion $(h)_{a,b} \to h$ uniformly on $\mathbb{R}^{2d}$ and in particular in every L^p.

Step 2. Assume now that $h \in W(L^p_1(\mathbb{R}^{2d})) \cap UC(\mathbb{R}^{2d})$ and consider the family of truncations $(\eta_\epsilon)_{\epsilon>0}$ defined as $\eta_\epsilon(x) = \eta(\epsilon|x|)$ for $x \in \mathbb{R}^{2d}$, where $\eta \in C^\infty([0,\infty))$ with $0 \leq \eta \leq 1$, $\eta \equiv 1$ on $[0,1]$ and $\eta \equiv 0$ on $[2,\infty)$. We have that

$$\begin{aligned}\|h - (h)_{a,b}\|_{L^p} &\leq \|h - h\eta_\epsilon\|_{L^p} + \|h\eta_\epsilon - (h\eta_\epsilon)_{a,b}\|_{L^p} + \|(h\eta_\epsilon)_{a,b} - (h)_{a,b}\|_{L^p} \\ &= \boxed{1} + \boxed{2} + \boxed{3}.\end{aligned}$$

By the definition of η_ϵ it follows that

$$\boxed{1} \leq \|h\mathbb{1}_{\{\epsilon|x|\geq 1\}}\|_{L^p}$$

which converges to 0 as $\epsilon \to 0$, since $h \in W(L^p_1(\mathbb{R}^{2d}))$, hence $h \in L^p(\mathbb{R}^{2d})$. Likewise,

$$\boxed{3} = \|(h - h\eta_\epsilon)_{a,b}\|_{L^p} \leq \|h - h\eta_\epsilon\|_{W(L^p_1)} \leq \|h\mathbb{1}_{\{\epsilon|x|\geq 1\}}\|_{W(L^p_1)},$$

which again converges to 0 (uniformly in a, b) as $\epsilon \to 0$. Hence, by the first step,

$$\limsup_{(a,b)\to(0,0)} \|h - (h)_{a,b}\|_{L^p} \leq \|h\mathbb{1}_{\{\epsilon|x|\geq 1\}}\|_{L^p} + \|h\mathbb{1}_{\{\epsilon|x|\geq 1\}}\|_{W(L^p_1)},$$

and the statement of the lemma follows by taking the limit $\epsilon \to 0$. □

Remark 3.3.10. Lemma 3.3.9 actually holds for $p > 0$: when $0 < p < 1$, the inequalities for $\boxed{1}$ and $\boxed{3}$ are the same, while

$$\limsup_{(a,b)\to(0,0)} \|h - (h)_{a,b}\|_{L^p} \leq K(\|h\mathbb{1}_{\{\epsilon|x|\geq 1\}}\|_{L^p} + \|h\mathbb{1}_{\{\epsilon|x|\geq 1\}}\|_{W(L^p_1)}),$$

for some $K > 1$, as $\|\cdot\|_{L^p}$ is not a norm but a quasi-norm. ■

PROOF OF PROPOSITION 3.3.8. By [Gröchenig, 2001a, Theorem 12.2.1] we have that if $g \in M^1_v(\mathbb{R}^d)$ and $f \in M^{p,q}_m(\mathbb{R}^d)$, where m and v are two weights such that $m(t_1+t_2) \leq Cv(t_1)m(t_2)$ for $C > 0$ and all t_1, t_2 in $\mathbb{R}^{2d}$, then $V_g f \in W(L^{p,q}_m(\mathbb{R}^{2d}))$. Hence under the assumptions of the proposition, it follows that $V_g f \in W(L^{2\alpha}_1)$. It is easy to check that $W(L^{2\alpha}_1(\mathbb{R}^{2d})) \subset L^{2\alpha}(\mathbb{R}^{2d})$, therefore the Rényi entropy $\mathrm{H}_\alpha(\mathrm{PS}f)$ is finite. The fact that the discrete entropy $\mathrm{H}^{a,b}_\alpha[\mathrm{PS}f]$ is finite follows almost immediately from the definition of the space $W(L^{2\alpha}_1(\mathbb{R}^{2d}))$. Indeed, the summands in equation (3.3.17) can be

grouped with respect to the larger cells of the lattice $\mathbb{Z}^d \times \mathbb{Z}^d$ (here we think a, b small) and each big cell contains at most $(2 + [\frac{1}{a}])(2 + [\frac{1}{b}])$ evaluation points from the smaller cells, where $[\cdot]$ denotes the integer part. It follows that

$$\sum_{k,n} |V_g f(t_k, \omega_n)|^{2\alpha} \leq (2 + [\tfrac{1}{a}])(2 + [\tfrac{1}{b}]) \sum_{n,k} a_{kn} (V_g f)^{2\alpha} \leq (2 + [\tfrac{1}{a}])(2 + [\tfrac{1}{b}]) \|V_g f\|^{2\alpha}_{W^{2\alpha}_1},$$

which is finite.

Finally, by Lemma 3.3.3 we also know that $V_g f \in UC(\mathbb{R}^{2d})$, hence the previous lemma (applied with $p = 2\alpha$ and $h = V_g f$) ensures the convergence. □

3.4. Biasing spectral coefficients through the α parameter

The α parameter in equation (3.2.6) introduces a biasing on the spectral coefficients, which gives them a different relevance in the entropy evaluation of the representation; this means that different values of α determine different concepts of sparsity: in this section we propose two tests to give a qualitative description of this biasing.

We first consider a collection of vectors composed by a variable amount of large and small coefficients. We realize a vector D of length $N = 100$ generating numbers between 0 and 1 with a normal random distribution; then we consider the vectors D_M, $1 \leq M \leq N$ such that

$$D_M[k] = \begin{cases} D[k] & \text{if } k \leq M \\ \frac{D[k]}{20} & \text{if } k > M \end{cases} \tag{3.4.1}$$

and then normalize to obtain a unitary sum. We then apply Rényi entropy measures with α varying between 0 and 3: as detailed in Chapter 4, these are the values we use in our adaptive framework for the entropy evaluation. These vectors are a simplified model of spectrogram frames, with coefficients whose amplitudes vary around two main values: the vectors are not necessarily frames of the spectrogram of a real signal, the scope of the test being to represent a limit case; actually, such a configuration may represent more general situations, as the entropy measure is permutation-invariant, so that the order of the coefficients in a vector does not modify its entropy value.

As we see from Figure 3.1, there is a relation between the number of large coefficients M and the slope of the entropy curves for the different values of α. For $\alpha = 0$, $\mathrm{H}_0[D_M]$ is the logarithm of the number of non-zero coefficients and it is therefore constant; when α increases, we see that densities with a small amount of large coefficients gradually decrease their entropy, faster than the almost flat vectors corresponding to larger values of M. This means that by increasing α we emphasize the difference between the entropy values of a peaky distribution and that of a nearly flat one. The sparsity measure we consider, privileges analyses with minimal entropy, so reducing α rises the probability of less peaky distributions to be chosen as sparsest: in principle, this may be desirable, as considering only large peaks lower the importance of weaker signal components, such as partials, which have to be taken into account in the sparsity

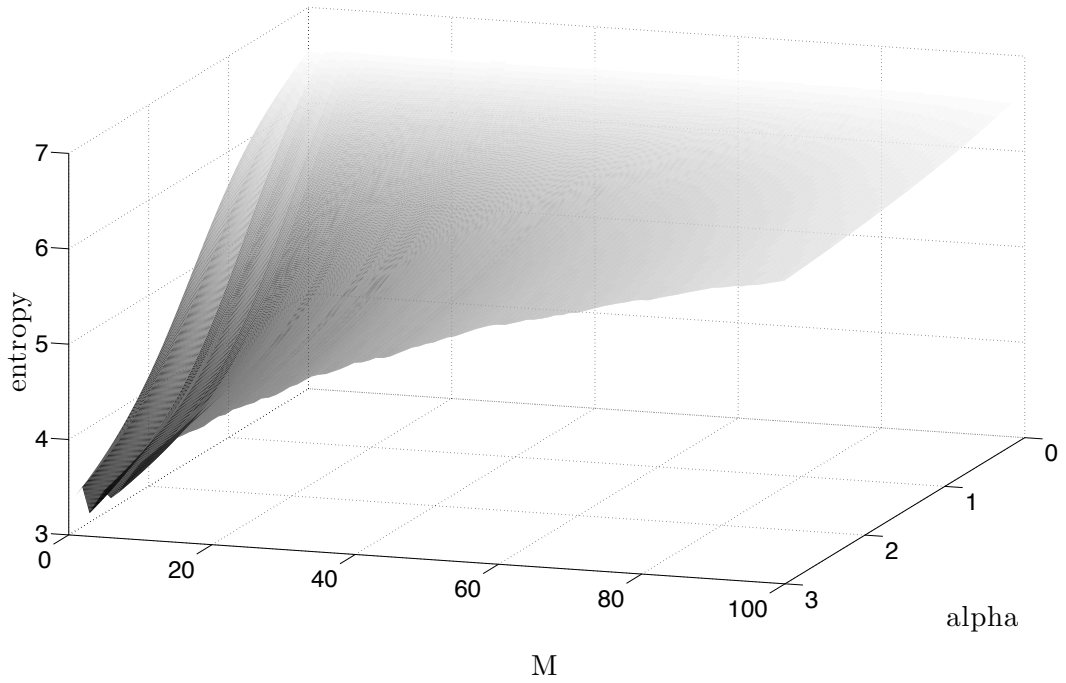

FIGURE 3.1. Rényi entropy evaluations of the D_M vectors with varying α; the distribution becomes flatter as M increases. Therefore increasing α favors a representation with a few large peaks(see text).

evaluation.

The second example we consider shows that the just mentioned principle should be applied with care, as a small coefficient in a spectrogram could be determined by a partial as well as by noise; with an extremely small α, the best window selected could vary without a reliable relation with spectral concentration, depending on the noise level within the sound. We illustrate how noise has to be taken in account when tuning the α parameter by means of another model of spectrogram: taking the same vector D considered previously, and two integers $1 \leq N_{part}$, $1 \leq R_{part}$, we define D_L like follows:

$$D_L[k] = \begin{cases} 1 & \text{if } k = 1 \\ \frac{D[k]}{R_{part}} & \text{if } 1 < k \leq N_{part} \\ \frac{D[k]}{R_{noise}} & \text{if } k > N_{part} \, . \end{cases} \tag{3.4.2}$$

where $R_{noise} = \frac{R_{part}}{L}$, $L \in [\frac{1}{16}, 1]$; then we normalize to obtain a unitary sum. This vectors are a simplified model of the spectrogram frames, whose coefficients correspond to one main peak, N_{part} partials with amplitude reduced by R_{part}, and some noise, whose amplitude varies proportionally to the L parameter, from a negligible level to the same one of the partials. Applying Rényi entropy measures with α varying between 0 and 3,

we obtain the figure 3.2, which shows the impact of the noise level L on the evaluations with different values of α.

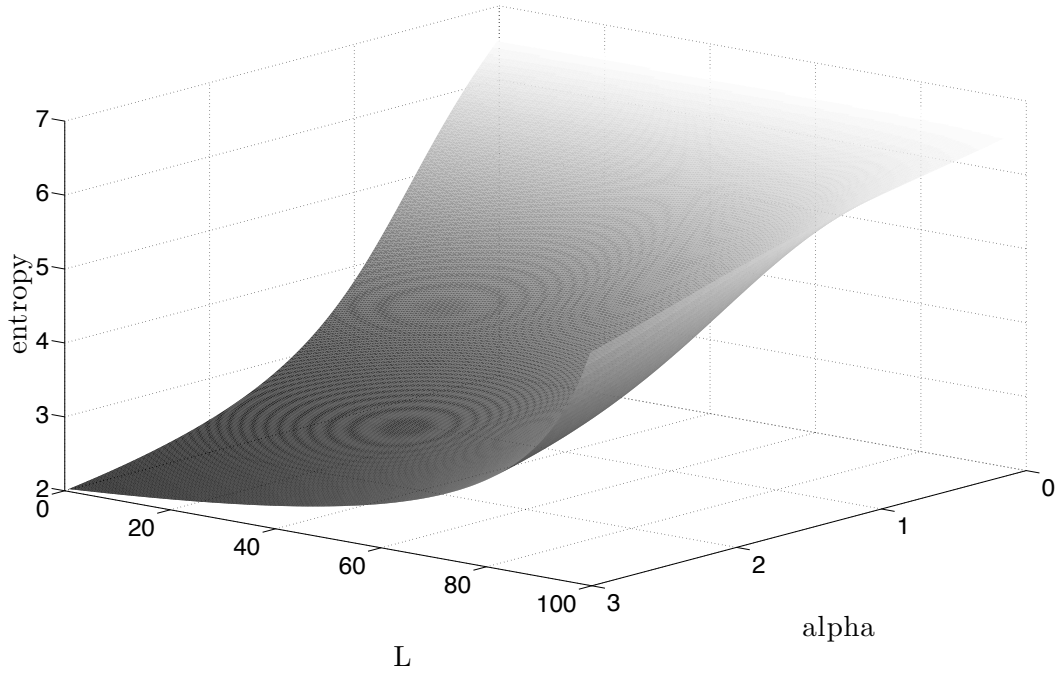

FIGURE 3.2. Rényi entropy evaluations of the D_L vectors with varying α, $N_{part} = 5$ and $R_{part} = 2$; the entropy values rise differently as L increases, depending on α: this shows that the impact of the noise level on the entropy evaluation depends on the entropy order (see text).

The increment of L corresponds to a strengthening of the noise coefficients, causing the rise of the entropy values for any α. The key point is the observation of how they rise, depending on the α value: the convexity of the surface in figure 3.2 increases as α becomes larger, and it describes the impact of the noise level on the evaluation; the stronger convexity when α is around 3 denotes an higher robustness, as the noise level needs to be high to determine a significant entropy variation. Our tests show that, as a drawback, in this way we lower the sensitivity of the evaluation to the partials, and the measure keeps almost the same profile for every $R_{part} > 1$.
On the other hand, when α tends to 0 the entropy growth is almost linear in L, showing the significant impact of noise on the evaluation, as well as a finer response to the variation of the partials amplitude. As a consequence, the tuning of the α parameter has to be performed according to the desired tradeoff between the sensitivity of the measure to the weak signal components to be observed, and the robustness to noise.

3.5. Rényi entropy evaluation of weighted spectrograms

The representation we take into account is the spectrogram of a signal f, as defined in equation (3.2.1), so $\mathrm{PS}_g f(t,\omega) = |\mathcal{V}_g f(t,\omega)|^2$. Given a Gabor frame $\{g_{k,l}\}$ we obtain a sampling of the spectrogram coefficients considering $z_{k,l} = |\langle f, g_{k,l}\rangle|^2$. We

have seen that, with an appropriate normalization, both the continuous and sampled spectrogram can be interpreted as probability densities. The new idea introduced in [**Liuni et al., 2011a**] is to use Rényi entropies as sparsity measures for weighted time-frequency distributions: if we consider a weight function $0 \leq w(k,l) \leq \infty$, we can weight here the discrete spectrogram obtaining a new distribution $z^*_{k,l} = w(k,l)z_{k,l}$ which is not necessarily the spectrogram of a signal: nevertheless, by the definition of $w(k,l)$, its entropy can still be evaluated from (3.6.7). This value gives an information of the concentration of the distribution within the time-frequency area emphasized by the specific weight function: as we show in the Subsection 5.2, this can be useful for the customization of the adaptation procedure, mainly for sound analysis purposes.

3.6. Spectral change detection in audio streams by means of Rényi entropies

In this section, we exploit the possibility of modeling spectral measures by means of the Rényi entropy, appropriately varying the α parameter; together with another basic property of entropies, which we discuss in Subsection 3.6.2, the possibility to bias spectral coefficients is used here to detect changes within an audio stream: this is an important task in several domains, such as classification and segmentation of a sound or of a music piece, as well as indexing of broadcast news or surveillance applications (we give several references fot this topic later in this section). The two novel methods for spectral change detection introduced in [**Liuni et al., 2011c**], that we detail here, are both based on the evaluation of information measures deduced by the Rényi entropy, and applied to the spectrogram: we show that they allow refined results compared to those obtained with standard divergences. These methods provide a low computational cost and are well-suited as a support for higher level analysis, segmentation and classification algorithms.

The detection of spectral changes within an audio signal can be performed according to many different criteria, depending on the applications; the key point is what kind of spectral change has to be considered significant. A typical problem in audio classification is to identify signal segments with different contents, for example when analyzing a radio stream to separate speech, music or mix of them; another type of problem is speaker change detection, which typically occurs when indexing audio recording of conferences, interviews or lectures. In either case we have to perform a segmentation and a classification, but the interesting spectral changes are completely different. The point of view we consider is at the signal level, without any assumption about the input sound and its content.

The use of information measures to evaluate the features of a time-frequency representation of a signal is frequent in the literature: Shannon entropy is applied to evaluate the concentration of the representation seen as a probability distribution, and the derived divergence measures [**Lin, 2002**] are employed to identify variations within the representation. The representation we consider is the spectrogram of the signal: through a normalization which gives a unitary sum, we consider the discrete spectrogram in a finite time interval as a probability distribution, and we can apply typical

information measures to evaluate its concentration in the time-frequency plane. Fixing the signal f, we write

$$\mathrm{PS}^m = \{\mathrm{PS}f(m,k),\ k = 1, ..., N\} \tag{3.6.1}$$

to indicate the m-th normalized analysis frame in the discrete spectrogram $\mathrm{PS}f$, where the *FFT size* N is the finite number of sample frequencies considered.

The *Kullback* I and J divergence measures are derived from the Shannon entropy [**Rényi, 1961**] as follows, where we assume $0 \log 0 = 0$ and $\log \frac{0}{0} = 0$. The I *directed divergence* is

$$I(\mathrm{PS}^1, \mathrm{PS}^2) = \sum_{k=1}^{N} \mathrm{PS}^1[k] \log \frac{\mathrm{PS}^1[k]}{\mathrm{PS}^2[k]}\ , \tag{3.6.2}$$

which is nonnegative and additive but not symmetric; a symmetric extension of this measure is given by

$$J(\mathrm{PS}^1, \mathrm{PS}^2) = \sum_{k=1}^{N} (\mathrm{PS}^1[k] - \mathrm{PS}^2[k]) \log \frac{\mathrm{PS}^1[k]}{\mathrm{PS}^2[k]}\ , \tag{3.6.3}$$

where PS^1 has to be absolutely continuous with respect to PS^2 for I to be defined while PS^1 and PS^2 have to be equivalent in the definition of J. The K *directed divergence* [**Lin, 2002**] is an alternative well suited for difference measures; it is defined as

$$K(\mathrm{PS}^1, \mathrm{PS}^2) = \sum_{k=1}^{N} \mathrm{PS}^1[k] \log \frac{\mathrm{PS}^1[k]}{\frac{1}{2}\mathrm{PS}^1[k] + \frac{1}{2}\mathrm{PS}^2[k]}\ , \tag{3.6.4}$$

so we have $K(\mathrm{PS}^1, \mathrm{PS}^2) \geq 0$ and the equality is attained only if $\mathrm{PS}^1 = \mathrm{PS}^2$. The last two measures are both derived from the I one, as $J(p,q) = I(p,q) + I(q,p)$ and $K(p,q) = I(p, \frac{1}{2}p + \frac{1}{2}q)$.

Given two normalized analysis frames PS^1 and PS^2, the K divergence is usually employed to have a measure of their difference: a spectral change is detected whenever $K(\mathrm{PS}^1, \mathrm{PS}^2)$ is larger than a chosen threshold. A refinement of this method (see for example [**Basu, 2003**]) provides a better robustness to false alarms defining a *mean spectrum* PS_{mean} and comparing its divergence with the new analysis frame.

The first method we introduce is a straight extension of the one just described: we consider the divergence measure derived from the Rényi entropy [**Rényi, 1961**] instead of the K directed divergence, allowing a tuning of the detection criteria thanks to the dependance of the measure on a parameter. The second method is not based on divergence but on Rényi entropy itself, exploiting one of its fundamental property: the entropy of a union of probability distributions can be evaluated considering the entropy values of the individual distributions. Since we do consider analysis frames as

probability distributions, this property can be used to establish the expected entropy value of a certain signal segment when the following frame is added: if the actual value differs significantly from the expected one, the last frame is considered to contain a spectral change.

This kind of algorithm does not need acoustic models to refer to, nor data training: a certain metric is evaluated in a given space [**Kemp et al., 2000**]. The information measures we take into account can be applied on several different representation of the signal: in [**Siegler et al., 1997**] the K divergence is used in a GMM framework instead of on the spectrogram. In several approach, for example in [**Foote, 2002**], difference measures are calculated as a first step which gives a suitable analysis for segmentation and classification purposes: for all these algorithms, the class of measures we introduce could ameliorate the detection performances as they allow a further parameter of choice, while still including the K divergence for a given value of the parameter.

In this section we discuss the methods introduced, while a technical description of the algorithms and examples are given in Section 5.4

3.6.1. Rényi information measures. We have seen that given a finite probability density P and a rational number $\alpha \geq 0$, the Rényi entropy of P is defined as in equation (3.2.4),

$$\mathrm{H}_\alpha[P] = \frac{1}{1-\alpha} \log_2 \sum_{k=1}^{N} P_k^\alpha \; ; \tag{3.6.5}$$

given a second finite probability density Q of the same length, if P and Q have exactly the same zeros the *Rényi information* [**Rényi, 1961**] is defined as follows,

$$I_\alpha(Q,P) = \frac{1}{\alpha-1} \log_2 \sum_{k=1}^{N} \frac{Q_k^\alpha}{P_k^{\alpha-1}} \; , \tag{3.6.6}$$

and it tends to the Kullback I divergence [**Lin, 2002**] as α tends to one. We can thus consider this class of measures to obtain different divergences as for the Kullback I one, and apply them to the spectrogram frames: as long as we can give an interpretation to the α parameter, this class of measures offers a more detailed information about the time-frequency representation of the signal.
The biasing introduced on the spectral coefficients by the α parameter have been interpreted by means of the simplified models in (3.4.1) and (3.4.2): looking at the first one, for $\alpha = 0$, $\mathrm{H}_0[D_M]$ is the logarithm of the number of non-zero coefficients and it is therefore constant; when α increases, we see that densities with a small amount of large coefficients gradually decrease their entropy. This means that increasing α we emphasize the difference between the entropy values of a peaky distribution and that of a nearly flat one. In the next subsection we give an example of the exploiting of this property, still considering that the smaller is the α parameter, the less the change detection is robust to noise level.

3.6.2. The entropy prediction method. The second method we introduce is not based on a divergence criterium, but on entropy itself. We first give the definition of Rényi entropy for the case of distribution obtained with a discretization of their continuous version [**Baraniuk et al., 2001**]: let PS_f be a normalization with unitary sum of a discrete spectrogram, then the Rényi entropy of PS_f is

$$\mathrm{H}_\alpha[\mathrm{PS}_f] = \frac{1}{1-\alpha}\log_2 \sum_{n,k}(\mathrm{PS}_f[n,k])^\alpha + \log_2(ab)\ , \tag{3.6.7}$$

where k varies between 1 and the FFT size N while n varies in the time interval where the evaluation has to be performed, according to the time grid. The term $\log_2(ab)$ takes into account the time and frequency steps a and b of the lattice Λ used to sample the continuous spectrogram: this guarantees the stability of the discrete entropy when changing the hop and the FFT sizes, as long as the sampling grid is dense enough in the time-frequency plane. For the entropy of a single analysis frame we write $\mathrm{H}_\alpha[\mathrm{PS}_f] = \mathrm{H}_\alpha[\mathrm{PS}_m]$ as above, where m is the time index of the analysis frame considered; for L different analyses frames, we write $\mathrm{H}_\alpha[\mathrm{PS}_f] = \mathrm{H}_\alpha[\mathrm{PS}_m, ..., \mathrm{PS}_{m+L}]$ to focus on the individual vectors. The following properties are straightforward by the definitions.

Proposition 3.6.1 (Rényi entropy prediction)**.** *Consider a spectrogram* PS_f *and a rational number* $\alpha \geq 0$.

(i) *Let* PS_m *be an analysis frame in* PS_f*; if* PS_k *is obtained rearranging the elements of* PS_m*, then*

$$\mathrm{H}_\alpha[\mathrm{PS}_m] = \mathrm{H}_\alpha[\mathrm{PS}_k] = \mathrm{H}\ , \tag{3.6.8}$$

$$\mathrm{H}_\alpha[\mathrm{PS}_m, \mathrm{PS}_k] = \mathrm{H} + 1\ . \tag{3.6.9}$$

(ii) *In general, if* $\mathrm{PS}_{m+1}, ..., \mathrm{PS}_{m+L}$ *are obtained rearranging the elements of* PS_m*, than*

$$\mathrm{H}_\alpha[\mathrm{PS}_m, ..., \mathrm{PS}_{m+L}] = \mathrm{H} + \log_2(L+1)\ . \tag{3.6.10}$$

As a rearrangement we mean a reordering of the frame coefficients, thus including the case of equality between frames. The idea of our method is that given the entropy of a certain signal segment $\mathrm{H}_\alpha(\mathrm{PS}_m, ..., \mathrm{PS}_{m+L})$ composed by L contiguous frames, we can predict $\mathrm{H}_\alpha(\mathrm{PS}_m, ..., \mathrm{PS}_{m+L+1})$ supposing the new frame to be spectrally coherent and thus iso-entropic with the previous ones. If on the other hand the entropy value of the new segment largely differs from the predicted value, we assume the new frame to be incoherent with the previous and so a spectral change is detected. There is here a strong assumption concerning the equivalence between the concept of spectral coherence and the fact that two frames are obtained with a rearrangement of their elements; according to the specific needs in the applications, the detection criteria can be based on variations of the property (3.6.10) to take into account different definitions of spectral coherence: for example, considering a set of admissible operations on the analysis coefficients in relation with the entropy variation that they provide.

3.7. A sparsity measure based on sinusoidal peaks

In Section 1.3 we have introduced the deterministic plus stochastic decomposition (1.3.8) of a signal, which requires the separation of sinusoidal and non-sinusoidal components in its Fourier spectrum. This model serves as a base for several parameter estimation and signal manipulation techniques, where the two parts are treated with different methods: their quality largely relies on the accuracy of the separation. Several approaches have been proposed for the classification of spectral coefficients, whose purpose is to establish which values of the representations are related to sinusoidal, or non-sinusoidal parts of the signal (see [**Wells and Murphy, 2010**] for a comparative survey).

We aim to deduce an optimization problem based on the separation task, which would provide our adaptive framework with a sparsity measure related to the sinusoidal modeling of the signal. Consider as before a set of indexes S and window functions $g_s,\ s \in S$, which are scalings of a same window function g; a discrete STFT $\mathcal{V}_s f$ is calculated with each window g_s, determining a different decomposition in sinusoidal and noise components for each s in S,

$$f = f^s_{sin} + f^s_{noise}\ . \tag{3.7.1}$$

We need now a criterium to privilege a certain decomposition among the ones obtained. For applications based on sinusoidal modeling, most of the manipulations is performed on the sinusoidal components: once fixed the classification algorithm and given the different decompositions in (3.7.1), a natural choice for an optimal representation is to look for the maximum of the ratio between the energies of the sinusoidal and noise detected parts,

$$\max_{s\in S} \frac{\|f^s_{sin}\|_2}{\|f\|_2}\ . \tag{3.7.2}$$

The problem is well-posed, in the sense that the maximum is in general unique, and varies depending on the analyzed signal; indeed, the classification changes depending on the frequency resolution of the analysis, and on the amplification of the noise level determined by the window function: both of these quantities change when varying the window size.

The criterium will favor sparse representations, in the sense that a maximum of energy is represented by the sinusoidal components, which are the ones that we consider significative, while the minimum is in the residual. The advantage of such a model-based measure is that it provides, by its definition, the best application-oriented criterium for the adaptation of the STFT window. On the other hand, a characterization of the measure heavily relies on the classification algorithm, thus determining a certain lack of generality. In Section 4.2, we discuss the tests about the local adaptation of the STFT with this criterium, by comparing them with the ones obtained with the entropy-based measure.

For the classification algorithm, we consider the approach in [**Röbel et al., 2004**]; the signal spectra are decomposed by means of classifying individual spectral peaks. Different descriptors are defined on individual peaks of a discrete STFT, considering the following quantities in relation with the individual bins of an analysis frame: mean time, duration, mean frequency, bandwidth (these quantities correspond to $\langle t \rangle$, σ_t , $\langle \omega \rangle$, σ_ω, as defined in Section 1.3), group delay (see [**Cohen, 1995**]) and reassigned frequency (see [**Auger and Flandrin, 1995**]); then a decision tree is established, and each peak is associated to the sinusoidal, sidelobe, or noise class.

The choice of another classification algorithm would lead to a possibly different measure, and this makes the whole framework application-dependent. Our choice is motivated by the aim to conceive a system that can be used in quasi real-time: that is, the computational complexity of the sparsity evaluation has to to be of the same order than the one of the spectral analysis system. The algorithm in [**Röbel et al., 2004**] (as well as the others analyzed in [**Wells and Murphy, 2010**]) is a frame-by-frame system, where the classification is performed on the audio for a single analysis frame, without the need for subsequent frames to be acquired; therefore, considering its accuracy and computational complexity, it constitutes an appropriate candidate.

CHAPTER 4

Algorithms and tests

We have realized several algorithms based on the theoretical framework detailed in the previous chapters: our first focus has been on the design of analyses, with a variable and automatically adapted time-frequency resolution, to privilege the readability of the sound representation. We have then considered the problem of defining synthesis operators associated to the analyses introduced: we have conceived different reconstruction methods extending the classical FFT-based approach, looking for algorithms which guarantee a perfect reconstruction of the original sound, as long as it is theoretically achievable, or an approximation within a predictable small error.

4.1. Automatic selection of the window size

We first consider the problem of choosing an optimal window size out of a given finite set; for the tests we take into account in the following, the window type is fixed: the choice of a specific family of windows is principally a matter of the envisaged application and of the desired representation features. Nevertheless, the theoretical framework developed in the previous chapters includes all of the common window types adopted in the applications: thus we expect that the results obtained for a given window type still hold, with the appropriate adaptations, for all the compactly supported, symmetric and sufficiently regular windows. Because of its good properties (see Subsection 1.5.1), the spectrograms we use are obtained from different scalings of a *Hanning window*

$$g(t) = \cos^2(\pi t)\chi_{[-\frac{1}{2},\frac{1}{2}]}\,, \tag{4.1.1}$$

with χ the indicator function of the specified interval.

Each size we consider is obtained with a scaling of the same original window: therefore we have a finite set S of positive scaling factors, and different scaled version of a compactly supported, symmetric window g,

$$g_s(t) = \frac{1}{\sqrt{s}}\, g\left(\frac{t}{s}\right)\,. \tag{4.1.2}$$

Given a signal f, a discrete spectrogram $\mathrm{PS}_s f$ is calculated for every window g_s. Each $\mathrm{PS}_s f$ is a time-frequency representation of the signal, whose sparsity can be evaluated according to a chosen measure, thus defining an optimization problem whose solutions indicate the windows we consider as *best*. The problem thus takes the following form,

$$\min_{s\in S} \mathrm{H}^R_\alpha[\mathrm{PS}_s f] \tag{4.1.3}$$

where R is a certain rectangle in the time-frequency plane, $\alpha \geq 0,\ \alpha \neq 1$ and H_α^R is the Rényi entropy defined in equation (3.2.6), whose discrete form is deduced in equation (3.3.17). We use a modified version of the measure, better suited for our problem: in the discrete finite case, each $\mathrm{PS}_s f$ is a matrix in $\mathbb{R}^m \times \mathbb{R}^n$, where m and n vary for each s; to compare the different entropy values, we thus need to investigate the dependance of the measure on m and n. We impose the basic requirement that, given a spectrogram with identical rows or columns (even if this is not necessarily the spectrogram of any signal), the measure must not depend on m and n, respectively: this implies, for instance, that for any window g_s and any stationary sound f, the entropy value $\mathrm{H}_\alpha[\mathrm{PS}_s f]$ does not depend on the number of analysis frames considered. From equation (3.6.10) and its analogous in the frequency dimension, we deduce a normalizing term that we add in equation (3.3.17), as follows,

$$\mathrm{H}_\alpha^{a,b}[\mathrm{PS}_s f] = \frac{1}{1-\alpha} \log \sum_{(l,k)} \Big(\frac{\mathrm{PS}_s f(l,k)}{\sum_{(l',k')} \mathrm{PS}_s f(l',k')} \Big)^\alpha + \log \frac{ab}{mn}\ , \tag{4.1.4}$$

where l and k are the row and column indexes of the matrix $\mathrm{PS}_s f$, while a and b are the time and frequency steps used for the discrete spectrogram. Throughout the work, we refer to the measure in (4.1.4) as the *normalized* discrete Rényi entropy, to distinguish it from the discrete Rényi entropy in (3.3.17): the two measures show different properties of stability, that we detail in the following subsection.

By fixing a value of α, the sparsest local analysis is defined to be the one with minimum Rényi entropy: thus the optimization is performed on the scaling factor s, and the best window is defined consequently, with a similar approach to the one developed in [**Jaillet and Torrésani, 2007**].

4.1.1. Entropy evaluation for basic signals. We give here some examples of solutions for the problem (4.1.3) with f a basic signal of finite duration: a random noise, the sum of stationary sinusoids, a sinusoidal burst, and a sinusoid with sinusoidal frequency modulation; we perform here a global evaluation of the entropy, that is R coincides with the whole time-frequency support of f. This will give an insight of the selection realized by the sparsity measure we have defined, when the evaluation is taken on a subset of the signal support: indeed, even if music and instrumental sounds are much more complicated than the signals we consider here, their local behavior can show similarities with these elementary examples. For all of the tests in this subsection, we use the normalized entropy measure (4.1.4).

We first consider the case of a L-point random noise, whose samples respect a standard normal distribution, obtained with the function *randn* in Matlab: here, the solution for the problem (4.1.3) is random, regardless of L and the scaling factors in S. This is a first property that holds for the normalized discrete entropy, but not for the standard one: in our problem, this is an advantage, because if a best analysis window were established for random noise, this would lower the interpretability of the best window selection for a deterministic signal embedded in noise. On the other hand, the standard discrete Rényi entropy (3.3.17) would rather privilege the largest window size.

Test 4.1.1. In this test, f is a single stationary sinusoid embedded in random noise with variable SNR (signal to noise ratio); as we have seen in Subsection 3.2.1, without noise it is not hard to provide an analytical solution, which is that the largest window size is the best. We can thus interpret the following measures as the robustness of the solution to noise. Consider the following setup:

- f is a 5000 points sinusoid with normalized frequency 0.01;
- the sizes len_s of the windows g_s are the even numbers from 512 to 4096;
- the spectrograms $\mathrm{PS}_s f$ are taken with FFT sizes equal to len_s, and the time step is $\frac{len_s}{4}$.

Figure 4.1 contains the Rényi entropy measures of $\mathrm{PS}_s f$ as a function of len_s; the α parameter is set to 3, and we see that the solution is definitively robust, as it becomes aleatory when the noise level is much higher than the sinusoid one. The cyclic oscillation we see for all the SNR values is due to the discrete finite dimension of the signals, and thus of $\mathrm{PS}_s f$: as we fixed a constant signal length for all the different windows g_s, there are several different spectrogram matrices with a same number of columns (analysis frames), given by

$$n_s = \left\lfloor \frac{L - len_s + a_s}{a_s} \right\rfloor , \tag{4.1.5}$$

where $\lfloor \cdot \rfloor$ is the integer part function; the oscillation in the curves thus depends on the discontinuities of n_s when s varies. Moreover, this implies that different spectrograms are obtained analyzing different portions of the signal: in Section 4.3 we detail the solution we adopt to reduce this oscillation.

Figure 4.2 is obtained with the same setup, but α is set to 0.1; as discussed in Section 3.4, small values of the α parameter raise the influence of the noise component in the entropy evaluation: as the best window for random noise, in the entropy sense, is random, the measure starts taking random values at a higher SNR than in the previous case.

We have extended the test to the sum of stationary sinusoids with different frequencies and same constant amplitudes: as the analytical results suggest (see [**Baraniuk et al., 2001**], the measure is stable as long as the sinusoids are *well* separated: that is, the difference between any two of their frequencies is larger than the maximum frequency step of the windows g_s. When this separation does not hold, the measure is hard to predict, and the solution of problem (4.1.3) is not necessarily given by the largest len_s. ■

Test 4.1.2. We consider now a sinusoidal burst, that is a short-duration signal obtained with an exponentially-decaying amplitude modulation of a sinusoid: this is a simplified model of a percussive sound, where the tone component is given by the vibrating membrane. For a readable analysis of such a signal, the time precision has to be privileged, as the information is strongly localized in time and relatively spread in frequency; the

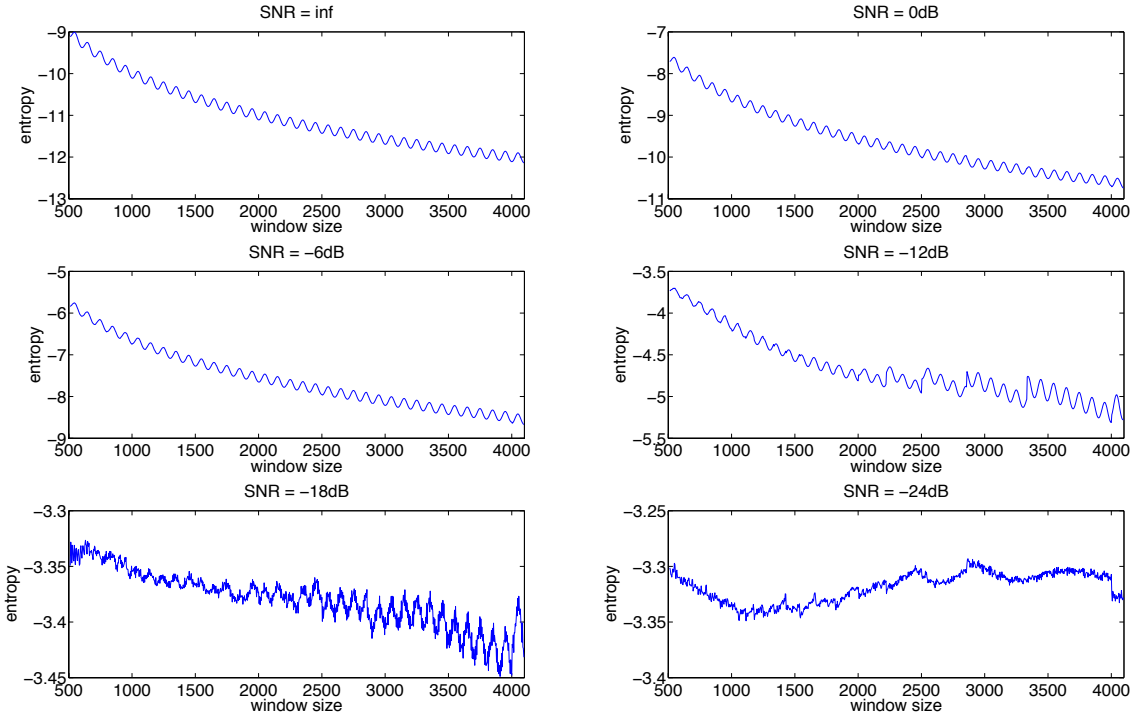

FIGURE 4.1. *Entropy values for different spectrograms of a single stationary sinusoid embedded in noise, as detailed in Test 4.1.1: the signal to noise ratio SNR is indicated for each figure, the entropy order is $\alpha = 3$ and the abscissa represents the length len_s of the windows g_s used for the spectrogram.*

shortest window size available is the one providing the best localization of the transient. We consider the same setup of Test 4.1.1, with the only difference about the input signal:

- f is a 5000 points signal containing a sinusoidal burst of 1500 points, the sinusoid normalized frequency is 0.25 (see the signal and its spectrum plots, on top of Figure 4.3).

As we see in Figure 4.3, without noise the Rényi entropy measure is lower for shorter window sizes. We would expect a concave curve, while we see steps and local convexities: they can be explained with similar considerations to those for the oscillation in the previous test. Nevertheless, the overall concavity ensures a satisfying stability of the solution of problem (4.1.3), in the noiseless case, given by the shortest window size. When noise is added, we see that even with $\alpha = 3$ the measure rapidly becomes unstable, which is normal considering the short duration of the signal compared to that of the introduced random noise. ■

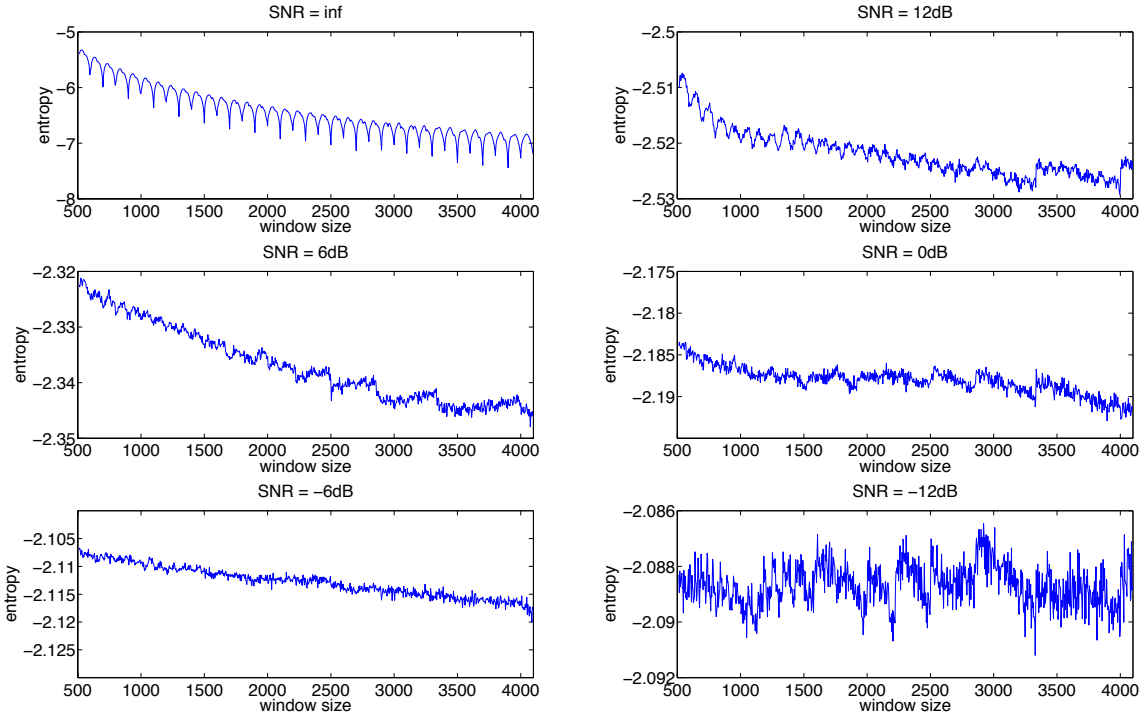

FIGURE 4.2. *Entropy values for different spectrograms of a single stationary sinusoid embedded in noise, as detailed in Test 4.1.1: the signal to noise ratio SNR is indicated for each figure, the entropy order is* $\alpha = 0.1$ *and the abscissa represents the length* len_s *of the windows* g_s *used for the spectrogram.*

Test 4.1.3. The last test we consider here is a sinusoid with sinusoidal frequency modulation: this will give an insight of the values taken by the entropy measure on vibrato sounds. We consider again the setup of Test 4.1.1, with the following input signal:

- f is a 10000 points sinusoid with sinusoidal modulation: the normalized frequency starts at 0.034, varying between 0.036 and 0.032.

If we suppose a sample rate of 44.1kHz, this corresponds to a 1500Hz sinusoid modulated of about a semitone, and could thus represent a partial of an instrumental note played with a vibrato effect. Different modulation periods are taken, along with the common perception of a vibrato effect, which is between 4 and 8 Hertz: if the frequency variation is fast, compared to the window length, this gives a frequency smearing within the analysis frame; we expect the measure to prevent this case, choosing a shorter window as best. As we see from Figure 4.4, this requirement is conveniently fulfilled by the entropy measure. ■

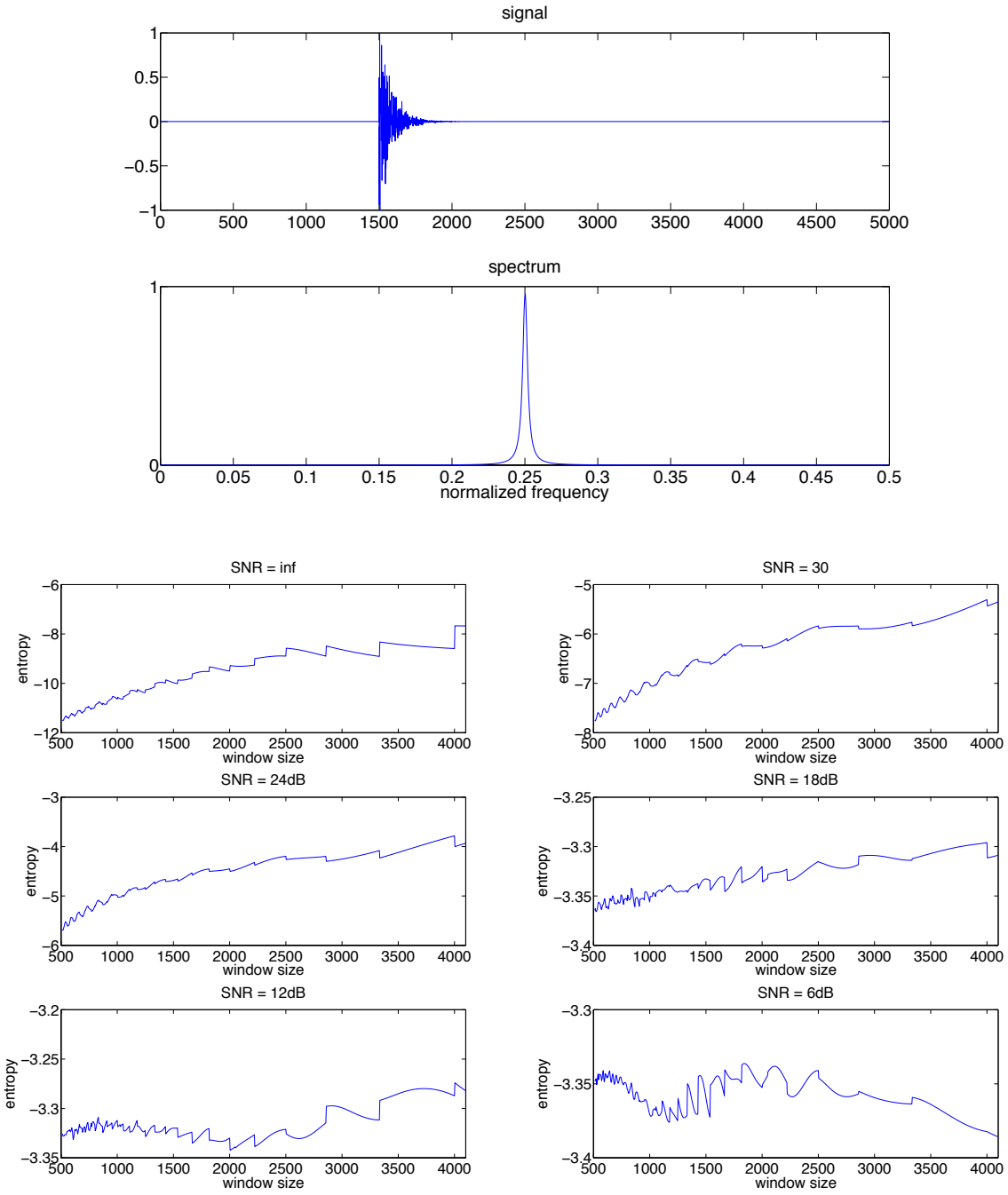

FIGURE 4.3. *Entropy values for different spectrograms of a sinusoidal burst embedded in noise, see Test 4.1.2: on top, the signal plot and its frequency spectrum one; in the other six plots, the entropy values are shown: the signal to noise ratio SNR is indicated for each figure, the entropy order is $\alpha = 3$ and the abscissa represents the length len_s of the windows g_s used for the spectrogram.*

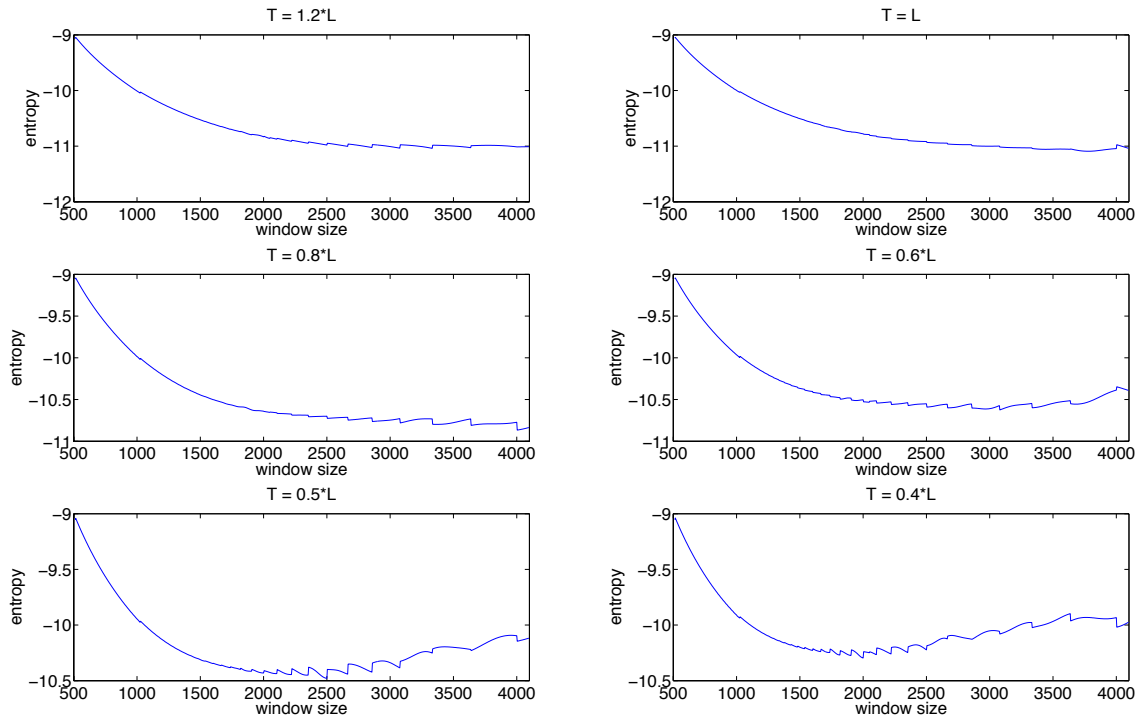

FIGURE 4.4. *Entropy values for different spectrograms of a sinusoid with sinusoidal frequency modulation, see Test 4.1.3: the signal length is $L = 10000$ points, the modulation period T is indicated for each figure as a multiple of L; the entropy order is $\alpha = 3$ and the abscissa represents the length len_s of the windows g_s used for the spectrogram.*

4.1.2. Time-frequency sampling steps. We now consider the solutions of the problem (4.1.3) when varying the time and frequency step a and b used for the discretization of the spectrogram. For discrete finite signals, the Hilbert vector space we consider is $\mathbb{C}^L$, where L is the signal length. Given the set S with the scaling factors, the window functions g_s have lengths len_s much smaller than L, thus determining the number of non-zero values of the FFT input vector, at each time step of the STFT. The FFT size is the number of points in the output vector: indicate with F_s the FFT size of the STFT with window g_s, which is the number of frequency coefficients of the signal transform, then the frequency step is given by $b_s = \frac{L}{F_s}$.

In the discrete finite case, the painless condition is verified when $len_s \leq F_s$: when equality holds, all of the information available in the windowed signal is exploited, as the output vector is given by an expansion whose coefficients are the non-zero input entries. If a larger number of input points is considered, then some zeros are included, and further frequency values are obtained as an interpolation between the ones in the

equality case (this technique is known as *zero-padding*, or *frequency-oversampling*). We consider that the frequency redundancy of the s-th analysis is given by $redf_s = \frac{L}{b_s} = F_s$; together with the one in time, that we are going to define as $redt_s$, this quantity allows to measure the overall redundancy of the analysis.

With regard to the time step, the results proved in Chapter 3 show that as long as the sampling lattice becomes denser, the discrete entropy (3.3.17) of a spectrogram converges to its continuos version: we would hence expect that over a certain redundancy, discrete entropy measures are robust to lattice variations. As a bound for the frequency step is implicitly fixed by the window size and the painless condition, which we require to be verified in all the analyses, the critical redundancy concerns the time step a. If we write the time redundancy of the s-th analysis as $redt_s = \frac{L}{a_s}$, then the global redundancy red_s of the spectrogram $\mathrm{PS}_s f$ is given by the following expression,

$$red_s = \frac{redt_s \cdot redf_s}{L} = \frac{\frac{L}{a_s} \cdot \frac{L}{b_s}}{L} = \frac{\frac{L}{a_s} \cdot F_s}{L}\ ; \tag{4.1.6}$$

in particular, $red_s = 1$ when the standard FFT is calculated, without windowing: in this case, the hop step is $a_s = L$, while the output vector has L points, so that $b_s = 1$.

We have conducted several experiments to verify the stability of the normalized and standard entropy measures to frequency oversampling and time step, that are resumed in the following test.

Test 4.1.4. We consider here a single window g, and verify how the discrete entropy measures change depending on the frequency oversampling and the time step; thus we focalize on a single spectrogram, and check the stability of the entropy values when varying its redundancy. Consider the following setup:

- three different signals: a 5000 points random noise, and the signals defined in the Tests 4.1.1 and 4.1.2 (sinusoid, sinusoidal burst, without noise);
- a single window function g of 1024 points;

(a) in a first configuration, the three signals are analyzed with a constant hop size of 256 points and variable FFT size, from 1024 to 8192 points;

(b) in a second configuration, the three signals are analyzed with a constant FFT size of 1024 points and variable hop size, from 64 to 512 points.

Figure 4.5 shows the results of the test, considering the standard discrete entropy measure (3.3.17) with $\alpha = 3$: the first configuration corresponds to the plots in the left column; the redundancy grows as the FFT size increases, because the frequency step b is inversely proportional to the frequency oversampling. The second configuration corresponds to the plots in the right column; the redundancy grows as the hop size decreases, as it coincides with the time step. In both cases, we see that the discrete entropy measure (3.3.17) is definitively stable when redundancy grows, given that the painless conditions are satisfied.

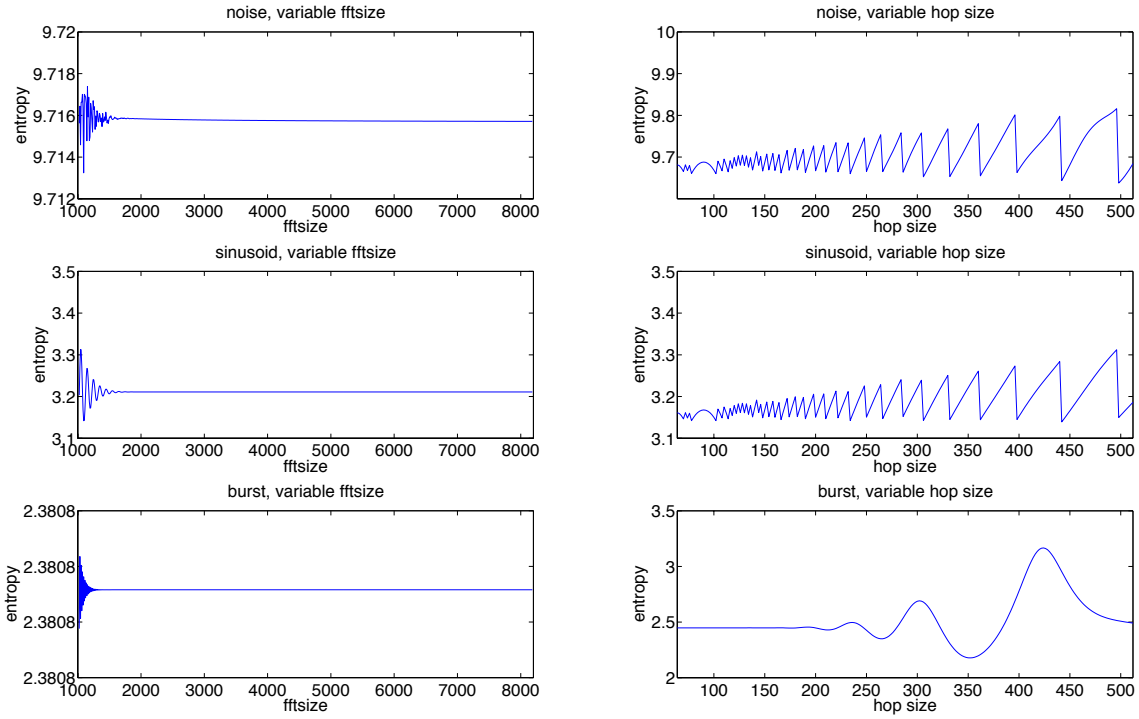

FIGURE 4.5. *Entropy values for the spectrograms of different signals, obtained varying their hop size and FFT size (see Test 4.1.4).*

Concerning the normalized discrete entropy measure (4.1.4), we deduce that it is not stable to redundancy variations, as it is obtained from the standard one, subtracting the logarithm of the spectrogram matrix dimensions, $\log mn$. Therefore, when comparing analyses with different windows and redundancies, the choice of the measure should take into account the type of stability required. ■

As a consequence of the tests shown in this section, the redundancies of the different spectrograms influence the interpretation of the comparison between their entropy values: therefore, we limit our investigation to the cases where analyses share the same redundancy, defining two different strategies.

In a first version of our algorithm, the different spectrograms are calculated with the same time step a and frequency step b; this implies that, for each signal segment analyzed, the different frames have Heisenberg boxes whose centers lay on a same lattice on the time-frequency plane, as illustrated in Figure 4.6. Given N the finite number of scaling factors in the set S, the window lengths are ordered such that $len_1 \leq ... \leq len_N$, and the same holds for the FFT sizes F_s. To guarantee that all the scaled windows constitute a frame when translated and modulated according to this global lattice, the time

step a must be set with the hop size assigned to the smallest window, which is a_1. On the other hand, to guarantee the painless condition for all of the analyses, the frequency step b has to be determined by the FFT size of the largest window, that is F_N: for the smaller ones, zero-padding is performed. In these hypotheses, all the spectrograms have the same redundancy, given by

$$red = \frac{L}{a_1 b_N} ;$$

moreover, the spectrogram matrices have the same number of rows and columns m and n: therefore, the discrete measure we use for the entropy evaluation is the standard one (3.3.17), as the normalized version would simply subtract a constant quantity to each evaluation.

As seen, a first strategy to obtain analyses with a common redundancy is to take the same time and frequency steps, defining a common lattice which is the narrowest in time and frequency among the ones of the individual analyses. A second strategy is to take the same time and frequency oversampling: in a further version of our algorithm, the window sizes len_s and the lattices Λ_s are automatically defined, given the smallest and largest lengths $len_1,\ len_N$, and the overall redundancy required for the analyses; in particular, the time steps are $a_s = c \cdot len_s$, while the FFT sizes are $F_s = d \cdot len_s$, where the parameters c and d, with $0 < c < 1$ and $d \geq 1$, define the redundancy shared by all the spectrograms,

$$red = \frac{d}{c} .$$

Figure 4.7 shows the configuration of the time centers for a given choice of S and c. In this case, the spectrogram matrices have different numbers of rows and columns m and n, so the discrete measure we use for the entropy evaluation is the normalized one (4.1.4). The redundancy of the analyses considered in the first version of the algorithm, with the same analysis parameters and windows, would be

$$red = \frac{d}{c} \cdot \frac{len_N}{len_1} ,$$

which is in general much higher than $\frac{d}{c}$.

The analyses in the first algorithm are oversampled versions of the ones in the second, and we refer to the latter as standard: depending on the window length len_s, the time and frequency steps of a standard analysis are reduced of a certain factor, adding rows and columns to the corresponding spectrogram matrix. For each s, the number of rows added is proportional to $\frac{len_s}{len_1}$, while the number of columns added is proportional to $\frac{len_N}{len_s}$; therefore, the ratio between the matrix dimension in the oversampled and standard sampling, given by

$$r_s = \frac{m_N n_1}{m_s n_s} ,$$

is approximatively the same for all the values of s, up to a discretization term. This implies that the solutions of the problem (4.1.3), considering the normalized

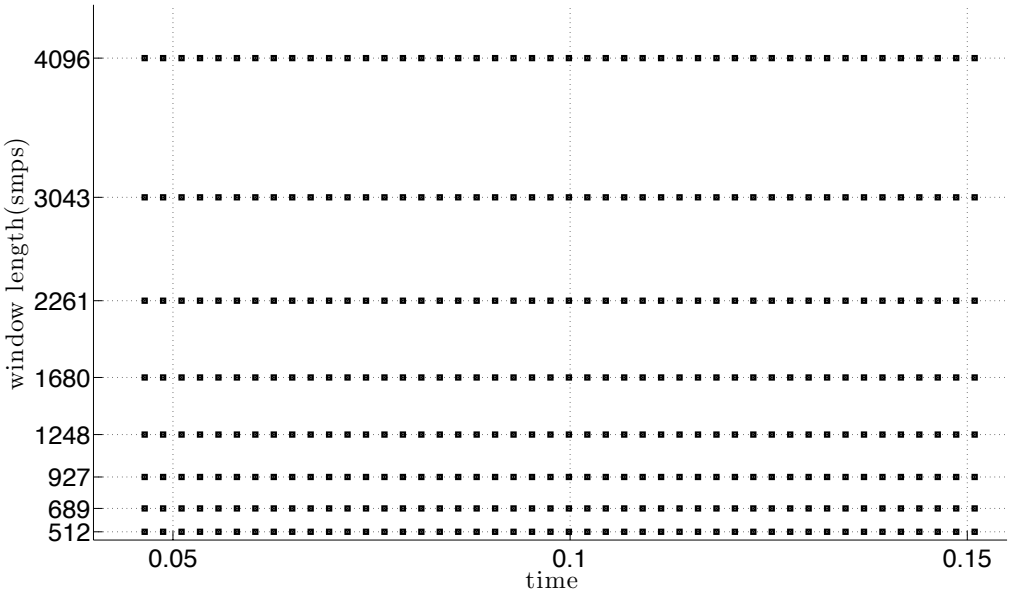

FIGURE 4.6. *Time centers in the spectrograms calculated by the first version of the algorithm, for a given choice of window lengths: the time step a_1 is determined by the time redundancy of the smallest window (see Subsection 4.1.2).*

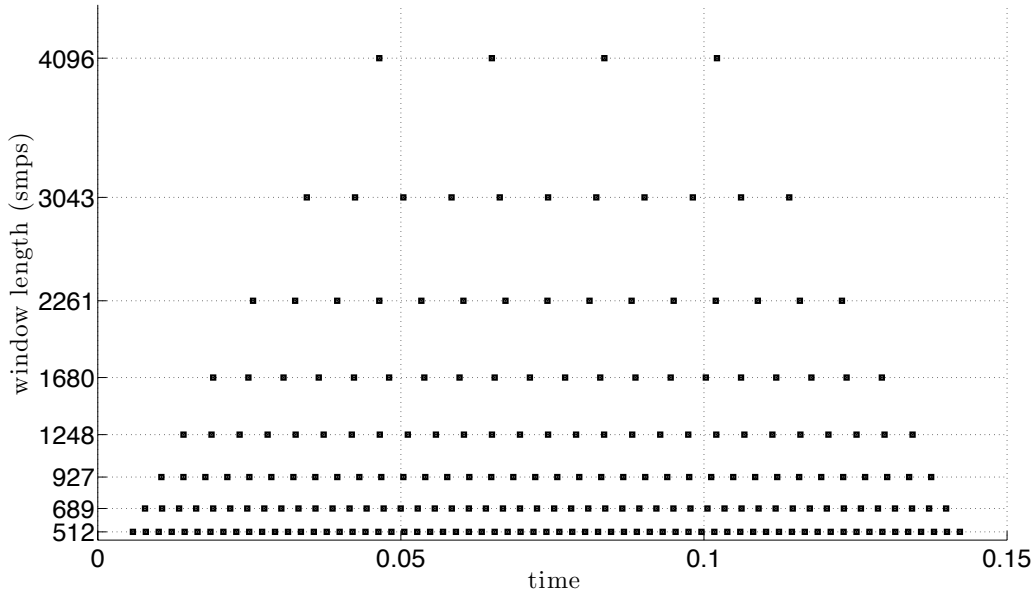

FIGURE 4.7. *Time centers in the spectrograms calculated by the second version of the algorithm, for a given choice of window lengths and redundancy: the time step $a_s = c \cdot len_s$ is a function of the window length len_s (see Subsection 4.1.2).*

entropy measure, are essentially the same with both the standard and oversampled spectrograms. This is a key point, as it guarantees that the sparsity measure obtained allows a total independence between the hop and FFT sizes of the different analyses, provided that the redundancy is the same: with the implementation of proper structures to handle multi-hop STFTs, we have obtained a more efficient algorithm in comparison with those imposing a fixed hop and FFT size, as the one proposed in [**Lukin and Todd, 2006**]; the computation saved is determined by the smaller number of FFTs required for the larger windows, which are the more expensive. For the experiments we show in the following sections and in the next chapter, this second version of the algorithm is used.

4.2. Adaptation of the STFT based on sinusoidal modeling

In the previous section we have conducted several tests to characterize the solutions of problem (4.1.3); here, a similar characterization for the solutions of the maximization problem (3.7.2) is given: as expected, when working with sums of stationary sinusoids, the best window size is chosen according to the frequency resolution guaranteeing the separation of the individual sinusoids. On the other hand, the results with an individual stationary sinusoid and a random noise show unexpected solutions, which are discussed in the following tests.

Test 4.2.1. Consider the following setup:

- f is a 20000 points random noise;
- the sizes len_s of the windows g_s are the powers of 2 between 256 and 8192, both included;
- the spectrograms $\mathrm{PS}_s f$ are taken with FFT sizes equal to $4len_s$, and the time step is $\frac{len_s}{4}$.

For every spectrogram, the classification between sinusoidal and non-sinusoidal peaks is performed, as detailed in Section 3.7; then the best window chosen is the solution of problem (3.7.2), which maximizes the energy of the sinusoidal component. The test is repeated 100 times with different realization of the random noise, and Figure 4.8 (on top) shows the occurrences of the different windows as optimal solution: as no sinusoidal peaks are present in the signal, we would expect the best solution to be randomly chosen: we see that the obtained distribution does not reveal any significant concentration, as expected.

We obtain similar results with a variation of the test; with the same setup, consider the case where f is a stationary sinusoid: for each one of the 100 repetitions, the normalized frequency randomly varies between 0 and 0.5: here, a single sinusoidal peak is present, which is classified as sinusoidal using any one the different windows. Therefore, we would again expect the best solution to be randomly chosen, which is confirmed by Figure 4.8 (at the bottom). ■

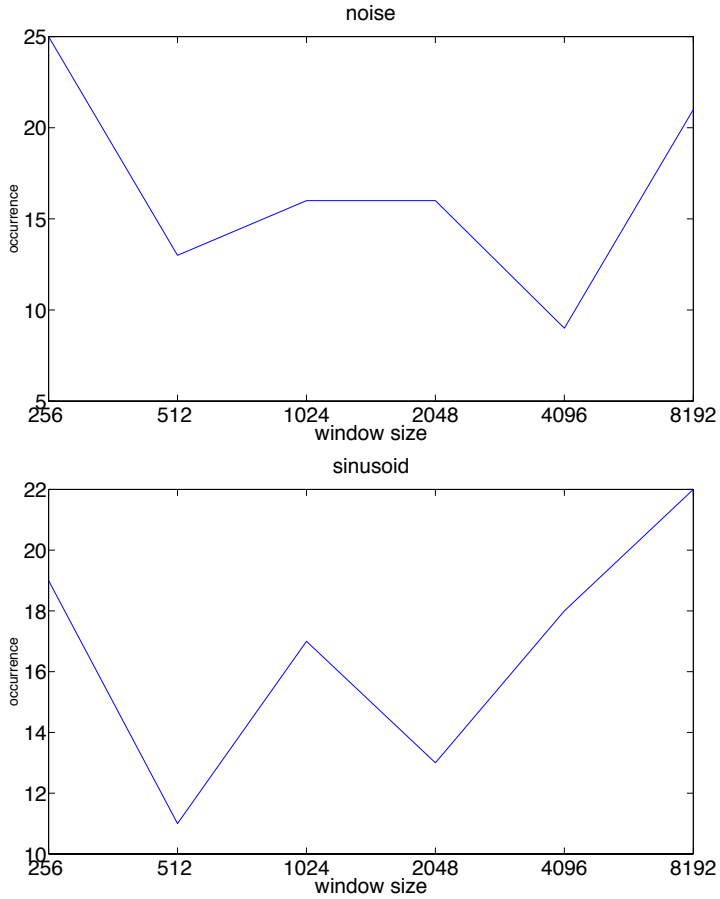

FIGURE 4.8. *Best window choice made by the classification-based sparsity measure, 100 different evaluations are realized for each signal type (see Test 4.1.2): the number of occurrences of a given window size as the best one is plotted.*

Even if, theoretically, the optimal resolution for a stationary sinusoid should be the highest one, from the point of view of the classification all the windows are good, as they all behave correctly with a unique sinusoidal peak. A further test is provided, detailing the solutions to the maximization problem (3.7.2) when f is a sum of two stationary sinusoids.

Test 4.2.2. The basic property required by the classification-based measure we have introduced is to privilege window sizes which resolve sinusoids; in this test we discuss the optimum defined by the measure when f is a sum of two stationary sinusoids with close frequencies. Consider the analysis setup of Test 4.2.1; choosing a frequency oversampling, a limit on the resolution of the windows is imposed: two sinusoids whose normalized frequencies difference is lower than $fres(s) = \frac{2}{len_s}$, are associated to a same frequency coefficient in the analysis with window g_s; the value $fres(s)$ indicates

the frequency resolution determined by g_s, as it is the lower frequency distance at which two sinusoids are properly resolved.

The signal f is a sum of stationary sinusoids, whose frequency separation is dif: a first normalized frequency is randomly chosen. We realize $|S| = 6$ repetitions of the evaluation, fixing the second normalized frequency such that $dif < fres(s)$ at the repetition s. Therefore, as s increases, the sinusoids are closer: at each step s, only the windows with index larger than s resolve the sinusoids, so we expect the maximum to hold at values higher than the current index. In Figure 4.9, we see that at each step a discontinuity of the measures occurs, at the index corresponding to the analysis window which does not resolve the sinusoids anymore, where the measure has its absolute minimum: at each step, the new window which does not separate the sinusoids is penalized. On the other hand, we see that reducing the frequency difference, the left part of the curves presents a decreasing slope, and smallest windows are not always penalized, even in some cases where they do not separate the two sinusoids: this effect is due to the fact that, when the frequency difference is below its frequency resolution, the sum of two stationary sinusoids is analyzed as a single sinusoid with sinusoidal amplitude modulation. In the classification algorithm, such a peak is classified as sinusoidal depending on the modulation rate and the window size: in particular, for smaller windows the energy of the signal is as well classified as sinusoidal.■

These two tests show that the classification-based measure allows a precise interpretation of the variation induced by unresolved sinusoids: the solutions to problem (3.7.2) are thus, in general, more easy to motivate in terms of signal processing considerations, compared to the ones obtained from the entropy minimization. The different solutions for the two problems, for instance in the limit case of single stationary sinusoids, show that two distinct sparsity criteria are defined, as expected. As a future perspective for this research, further experiments on appropriate classes of signals should show the characteristics of the different solutions in real music contexts. This would clarify the applications where one of the two measures should be preferred, as a criterium for the local adaptation of the analysis resolution.

4.3. Adaptive analysis

We have seen in Chapter 2 that it is possible to define time-frequency representations of a signal with variable resolution, moving from the classic Fourier-based approach. Then, in Chapter 3, we have established a rule to determine an optimal resolution out of a finite set of choices, and provided two methods of global adaptation in Sections 4.1 and 4.2. Here, we merge these concepts in a procedure for the local adaptation of the time-frequency resolution of the spectrogram. The sparsity criterium we adopt is the entropy-based one.

Given a finite set S of positive scaling factors, we consider different scaled versions g_s of a window g, as in equation (4.1.2). We know that each g_s, together with an appropriate lattice Λ_s defined by a time step and a frequency step, forms a frame

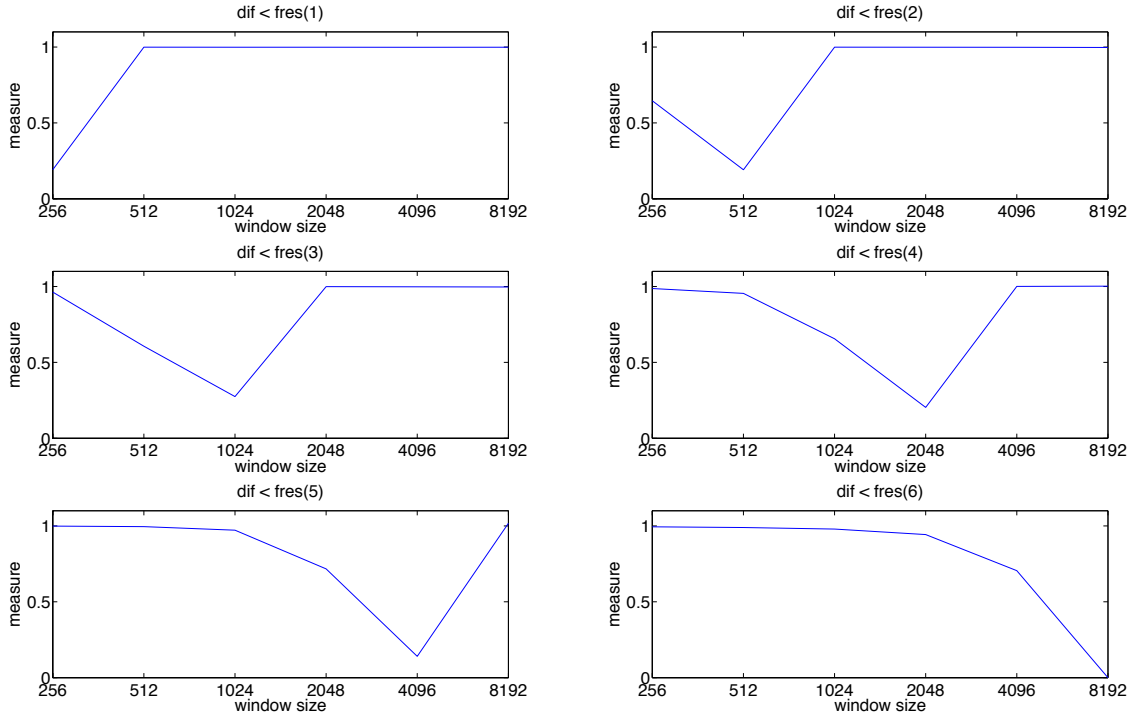

FIGURE 4.9. *Best window choice made by the classification-based sparsity measure: the signal is a sum of 2 stationary sinusoids whose normalized frequencies difference is progressively reduced, depending on the frequency resolutions of the analysis windows, as indicated in the titles of the subplots (see Test 4.2.2).*

for our space of signals. Associating the analysis coefficients to the points of the lattice, we can represent the discrete spectrogram $\mathrm{PS}_s f$ by means of the lattice Λ_s. A local evaluation of a sparsity measure takes into account a certain subset of the analysis coefficients, depending on the envisaged localization: when f is a sound, its time frequency support can be inscribed in a rectangle $\overline{R}$, whose horizontal and vertical sides are the time support of f and its essential frequency support, respectively.

The localization we are interested in, is realized by choosing a rectangle R in the time-frequency plane, whose time-frequency shifts cover $\overline{R}$; the area inside R corresponds to the analysis coefficients considered for the sparsity evaluation, and thus for the adaptation procedure: for each shift of R a best resolution is chosen and assigned to that portion of plane (see Figure 4.10).

For a better understanding of the type of localization obtained, we have to consider the relation between the analyzed coefficients and the signal segment they correspond

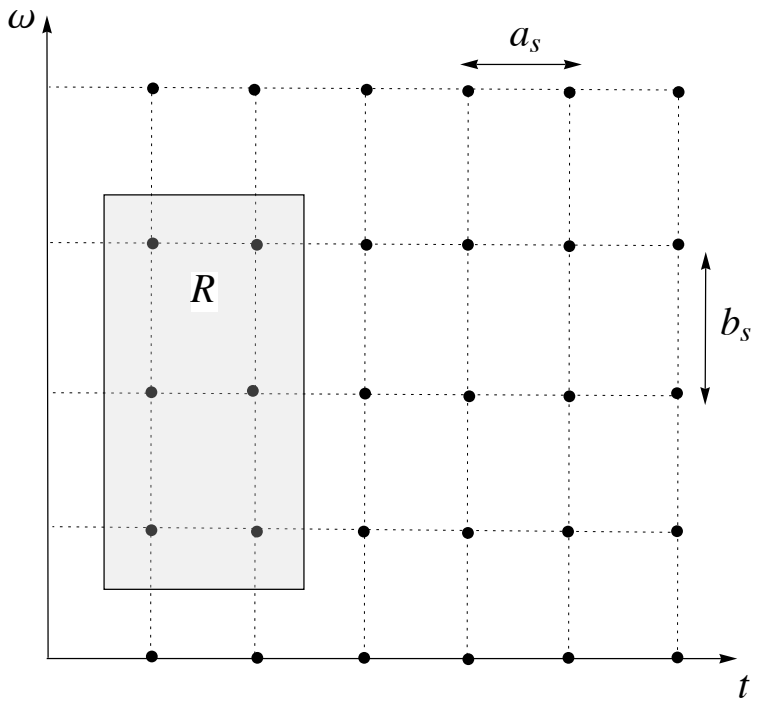

FIGURE 4.10. *Graphic representation of a step of the evaluation procedure: for a given spectrogram* $\mathrm{PS}_s f$ *and a time-frequency shift of the rectangle* R*, the Rényi entropy of the coefficients within* R *is calculated (see Section 4.3).*

to: because of the windowing, each coefficient in $\mathrm{PS}_s f$ is referred to a $\mathrm{supp}(g_s)$-long segment of f. Therefore, for the evaluation to be performed on rectangles containing at least one coefficient of each $\mathrm{PS}_s f$, we take R such that the temporal supports of the scaled windows g_s is inside R, i.e. $\mathrm{supp}(g_s) \subseteq R$ for any $s \in S$.

At each step of our algorithm, the rectangle R is shifted in the time-frequency plane with a certain overlap with the previous position. Within the area of the shifted R, the best coefficients are defined as the ones which belong to the spectrogram $\mathrm{PS}_s f$ providing the lowest Rényi entropy; in the overlapping regions, the decision is updated at each step of the algorithm. The adaptive global analysis is thus obtained as an union of the best local analyses selected by the algorithm. The parameters and the essential steps performed by the algorithm are represented in Figure 4.11, and will be detailed in the following subsections: examples of its application are provided in Chapter 5.

4.3.1. Time adaptation. The entropy evaluation is recursively performed on segments of the signal, taking into account the whole frequency spectrum. Each signal segment identifies a time-frequency rectangle R for the entropy evaluation: the horizontal side is the time interval of the considered segment, while in this case

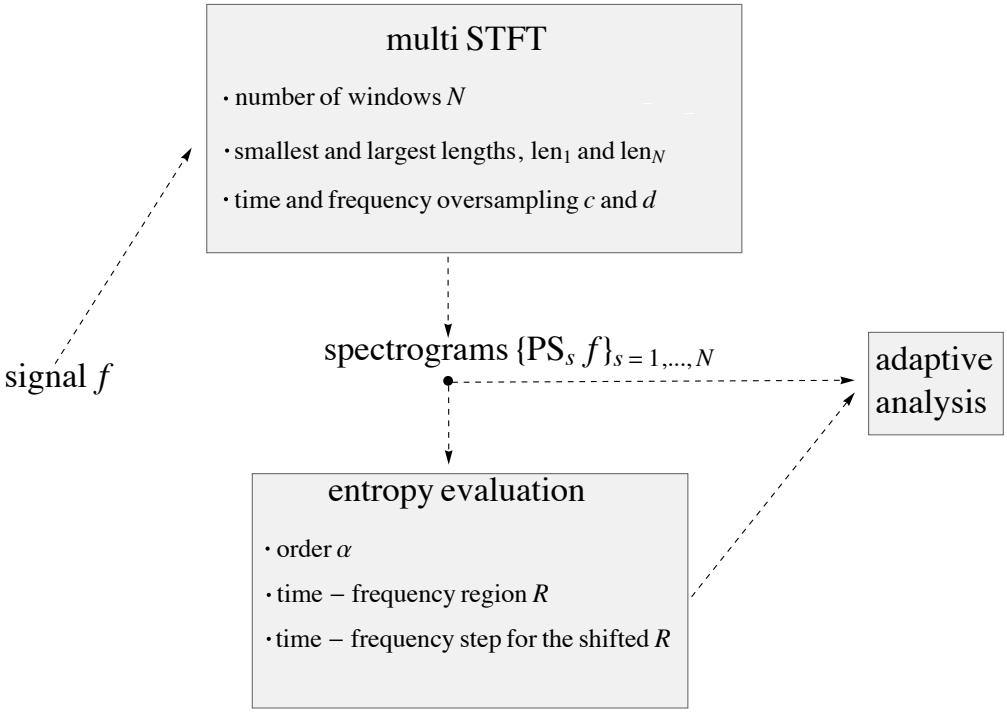

FIGURE 4.11. *Graphic representation of the main steps performed by the algorithm for the automatic local adaptation of the spectrogram window size (see Section 4.3).*

the vertical one is the whole frequency lattice. For each spectrogram, the rectangle R defines a subset of coefficients belonging to R itself. Because of the discrete setting, the coefficients identified by R over the different lattices Λ_s do not correspond in general to the same part of signal, as window lengths len_s and hop sizes $c \cdot len_s$ have to be integers (see Figure 4.12); moreover, different windows determine different amplification of the signal segment extremes. Therefore, we perform a preliminary weighting before the calculations of the local spectrograms, consisting of a multiplication of the extreme left and right sides of the signal segment, by the left half and right half of the largest window, respectively: this step reduces the variation of the entropy calculus coming from the signal segment extremes.

After the pre-weighting, we calculate the normalized discrete entropy of every spectrogram $\mathrm{PS}_s f$ as in (4.1.4). Having the $|S|$ entropy values corresponding to the different local spectrograms, the sparsest local analysis is defined as the one with minimum Rényi entropy: the window associated to the sparsest local analysis is chosen as best window for all the time points contained in R.
The global time adapted analysis of the signal is finally realized with a further spectrogram calculation of the unweighted signal, employing the best windows selected at

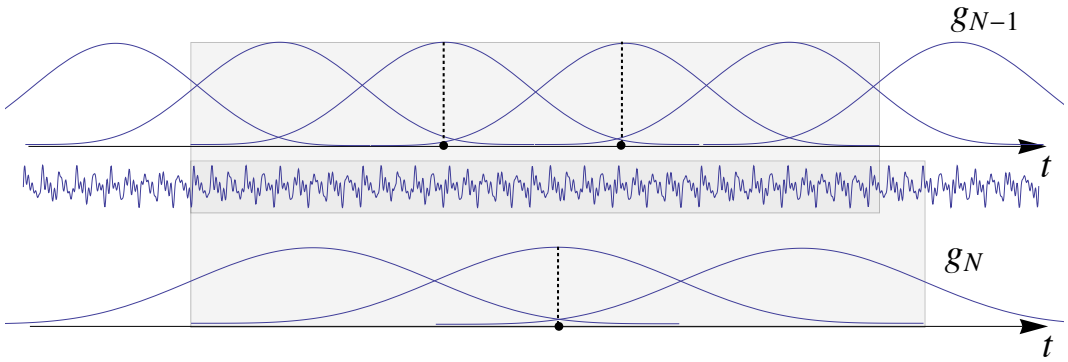

FIGURE 4.12. *Graphic representation of the signal segments analyzed by the different windows, at a given step of the algorithm. We choose a time interval corresponding to one frame of the window g_N (rectangle at the bottom); then we take the coefficients associated to time-shifts of the window g_{N-1}, whose supports are included in the time interval: as we see, the signal segment analyzed with the window g_{N-1} (rectangle on top) is not in general the same of the one analyzed with the window g_N.*

each time location.

4.3.2. Time adaptation with different masks. This variation of the algorithm is particularly useful for analysis purposes: music signals have often dense spectra, for which a global frequency adaptation is not meaningful. With the weight functions introduced in Section 2.5, we are able to limit the frequency range of the rectangle R at each time location: consider a weight function $0 \leq w^{an}(\omega) \leq 1$, and weighted versions of the spectrograms $w^{an}(\omega) \cdot \mathrm{PS}_s f$; if we evaluate the entropy of these distributions, this corresponds to the selection, or the biasing, of certain frequency bands of the signal, before the entropy evaluation step of the algorithm. We use the notation w^{an} for the analysis weights, to distinguish them from the weighting functions w for the reconstruction step, which may in general be different (see Section 4.4).

As done in the time case, the local best window at each time location is assigned to all the frequency points individuated by the time range of R. With this technique, we can, for instance, adapt the global analysis at each time location, according to the best local resolution required by a certain instrument.

4.3.3. Time-frequency adaptation. Extending the approach based on frequency masks, we introduce a further method for the time-frequency adaptation of the analysis resolution: chosen a certain number P of frequency bands, we perform a time adaptation with a mask for each one of them. For this purpose, P different binary masks $w_p^{an}(\omega),\ p = 1, ..., P$ are considered, matching the different bands, and such that for every ω we have $\sum_p w_p^{an}(\omega) = 1$. At the step p of the algorithm, the spectrograms

$w_p^{an}(\omega) \cdot \mathrm{PS}_s f$ are used for the entropy evaluation.

This means that the rectangle R is iteratively modified: at each entropy evaluation, its frequency range is given by the band considered. In this way, the resolution at each point of the adapted time-frequency lattice is determined by its time location and the frequency band which it belongs to. At the end of the algorithm, P different time-adapted spectrograms are calculated, corresponding to the same number of nonstationary Gabor frames in the time painless case.

4.4. Re-synthesis from adaptive analyses

In Sections 2.2.1, 2.5.1 and 2.6 we have treated three different reconstruction methods based on adaptive analyses; the first is used if the analysis resolution is uniquely time-dependent: in this case, the decomposing atoms form a nonstationary Gabor frame whose dual gives a re-synthesis formula with perfect reconstruction. Here, we focus on the two new approximation methods introduced, considering two frequency bands, so $P = 2$: both of the methods are defined from weighted analyses, the difference concerning the approximated dual frame used by the synthesis operator.

Given the signal f and its reconstruction f_{rec}, we measure their precision by means of the maximum of the absolute value of the error $peak = \|f - f_{rec}\|_\infty$, and the *RMS* (Root Mean Square) of the error, that is

$$rms = \sqrt{\frac{\sum_{n=1}^{L}(f[n] - f_{rec}[n])^2}{L}}\,, \tag{4.4.1}$$

where L is the signal length. We first consider our implementation of the nonstationary Gabor frames reconstruction formula: for the music signals in Section 5.1, with the time-adapted analyses realized by our algorithm, we obtain $peak \simeq 10^{-15}$ and $rms \simeq 10^{-16}$: the sound files are in standard cd format, stored with .wav extension 16 bit PCM, with amplitude range between -1 and 1.

We refer now to the approximation methods based on weighted analyses. We first detail our approach in terms of stationary Gabor frames, which is also the case which the estimates in Subsection 2.6.2 refer to. Then, we will extend the methods to the nonstationary case, which is used in our framework.

Using the notation introduced in Chapter 2, and Subsection 2.3.1 in particular, we consider two weight functions w_p, depending only on the frequency ω, such that $w_1(\omega) + w_2(\omega) = 1$ for every ω. Given two window functions g and h, we want to associate the Gabor frame $\mathbf{G}(g, a_1, b_1)$ to the first frequency band, and $\mathbf{G}(h, a_2, b_2)$ to the other. We do this by means of the weight functions, whose supports have to coincide with the two bands, eventually considering an overlap. In particular, the method we indicate as *analysis-weight* is given by the filter bank approach with Gabor multipliers (see Subection 2.6.2), and the reconstruction formula is given by

$$f_{rec} = \mathrm{D}_{\tilde{\mathrm{g}}}(\mathrm{C}_{\mathrm{g}}^{w_1} f) + \mathrm{D}_{\tilde{\mathrm{h}}}(\mathrm{C}_{\mathrm{h}}^{w_2} f)\ . \tag{4.4.2}$$

Therefore, each weighted analysis is used in the expansion with the original dual window, without calculating the exact dual of the global composed frame.

The second reconstruction method we use is formula (2.5.7) with stationary Gabor frames, that we indicate as *extended weight*, for the case $P = 2$. The reconstruction error given by the two methods is detailed here by means of two tests, and is applied on a music signal in Section 5.3.

Test 4.4.1. We first want to quantify the error obtained when the weight functions w_p are binary masks, and f is a basic signal whose spectral energy is concentrated at the cut frequency. Therefore, we consider a stationary sinusoid, with the following setup:

- f is a sinusoid sampled at 44.1kHz with frequency 11.025kHz;
- the functions w_p are two binary masks, the cut frequency ω_{cut} corresponds to the sinusoid frequency;
- the windows g and h are Hanning windows of size 512 and 4096 samples.

With the same setup, we then define different weights with a certain frequency overlap, and check if the obtained error is reduced. We define w_p as follows, given $\omega_1 = 10.05$kHz, $\omega_2 = 12$kHz and $Ny = 22.05$kHz the Nyquist frequency,

$$w_1(\omega) = \begin{cases} 0 & \text{if } 0 \leq \omega \leq \omega_1 \\ \frac{\omega - \omega_1}{\omega_2 - \omega_1} & \text{if } \omega_1 \leq \omega \leq \omega_2 \\ 1 & \text{if } \omega_2 \leq \omega \leq Ny \end{cases}$$

and $w_2 = 1 - w_1$; therefore, the two masks realize a linear cross fade, in frequency, between the two sets of analysis coefficients. The reconstruction error we obtain is resumed in the Table 4.1.

We see that, as expected, if the spectral energy of the signal is included within the overlap of the wighting functions, then the reconstruction error is definitively small, compared to the one obtained with a simple binary mask. ■

Test 4.4.2. We now consider a signal whose spectral energy oscillates: taken a sinusoid with sinusoidal frequency modulation, we want to measure the reconstruction error with binary masks, and the reduction obtained allowing the masks for an overlap. Therefore, f is a sinusoid sampled at 44.1kHz with frequency modulation: the start frequency is 350Hz and the modulation varies between 130Hz and 570Hz with a period of half a second; given $\omega_{cut} = 350$Hz, $\omega_1 = 200$Hz and $\omega_2 = 500$Hz, we consider the same masks used in Test 4.4.1. The aim is to show that the reconstruction error obtained with the two methods can be reduced, appropriately choosing the overlap of the two masks, according on the signal spectral energy.

TABLE 4.1. *Reconstruction error when f is a stationary sinusoid: the masks are indicated on the left, together with their frequency significant values (see Test 4.4.1).*

Weight method	Parameters	peak	rms
Binary mask	$\omega_{cut} = 11025$	0.3492	0.032
Linear cross	$\omega_1 = 10050$Hz $\omega_2 = 12000$Hz	0.0207	0.019
Linear cross	$\omega_1 = 5050$Hz $\omega_2 = 17000$Hz	0.0034	0.0032
Extended weight	$\omega_1 = 10050$Hz $\omega_2 = 12000$Hz	0.0078	$4.58 \cdot 10^{-4}$
Extended weight	$\omega_1 = 5050$Hz $\omega_2 = 17000$Hz	0.0014	$8.4771 \cdot 10^{-5}$

In Table 4.2 we have the errors obtained: as we can see, with the overlap 200-500Hz, which does not include the whole modulation range, then the reduction we obtain in the rms error is limited with the analysis-weight: this is due to the fact that, as a consequence of the weighting, many coefficients are set to 0, and therefore are not considered in the expansion (2.6.13); as we are considering too few coefficients for the reconstruction of the two individual bands, then the error on the global reconstruction is still considerable. For the extended weight method, the rms error increases; as we see from the expansion (2.5.1), for this method the non-zero weights cancel, and a large overlap is the only possibility to reduce the error: on the other hand, for limited overlap the error obtained is comparable to the one with binary weights.

TABLE 4.2. *Reconstruction error when f is a sinusoid with sinusoidal modulation: the masks are indicated on the left, together with their frequency significant values (see Test 4.4.2).*

Weight method	Parameters	peak	rms
Binary mask	$\omega_{cut} = 350$Hz	0.5102	0.0967
Linear cross	$\omega_1 = 200$Hz $\omega_2 = 500$Hz	0.1856	0.0725
Extended weight	$\omega_1 = 200$Hz $\omega_2 = 500$Hz	0.4708	0.1445
Linear cross	$\omega_1 = 50$Hz $\omega_2 = 650$Hz	0.0576	0.0262
Extended weight	$\omega_1 = 50$Hz $\omega_2 = 650$Hz	0.0392	0.0104

In the last two lines of Table 4.2, we see that, as expected, increasing the overlap of the weights we get a considerable reduction of the error. In particular, if the weights are positive (overlap over the all frequency dimension), then the extended weight give perfect reconstruction (as shown in Section 2.5), while the analysis-weight with linear cross gives approximations with rms error lower than 10^{-5}, depending on the absolute maxima and minima of the weights. The drawback is that analyses with such an overlap are hard to be interpreted, as all the different atoms employed give contributions at every time frequency point. In particular, it would be extremely hard to conceive sound processing techniques dealing with all of the different resolutions at the same time.

Finally, in Figure 4.14 we see the composed spectrogram obtained with the binary masks, and the consequent reconstruction error. The same, in Figure 4.13, for the analysis-weight method with linear cross, and overlap 50-650Hz. We thus see that the spectral energy of the error with overlapping weights is lower, and more uniformly distributed. ■

The tests we have shown are obtained with two stationary Gabor frames, each one associated to a frequency band. In our framework, we extend this methods to non-stationary Gabor frames. With the different scalings g_s of a same window function, and appropriate lattices Λ_s, we realize the analyses $\mathcal{V}_{g_s} f$ and their weighted versions $\mathcal{V}_{g_s} f(t,\omega) w_p(\omega)$. These weighted analyses are used for the reconstruction, after the automatic adaptation detailed in Subsection 4.3.3: at the end of the optimization procedure, the frequency band p is associated to the nonstationary Gabor frame $\{g^p_{k,l}\}$ of the best windows at the corresponding time-frequency points: if we indicate with $\mathrm{C_p}$ and $\mathrm{D_p}$ the analysis and synthesis operators associated to the p-th frame and its canonical dual, then the analysis-weight method implemented in our framework takes the following form,

$$f_{rec} = \mathrm{D}_1(\mathrm{C}_1^{w_1} f) + \mathrm{D}_2(\mathrm{C}_2^{w_2} f)\ , \tag{4.4.3}$$

while the extended weight reconstruction is given in formula (2.5.7).

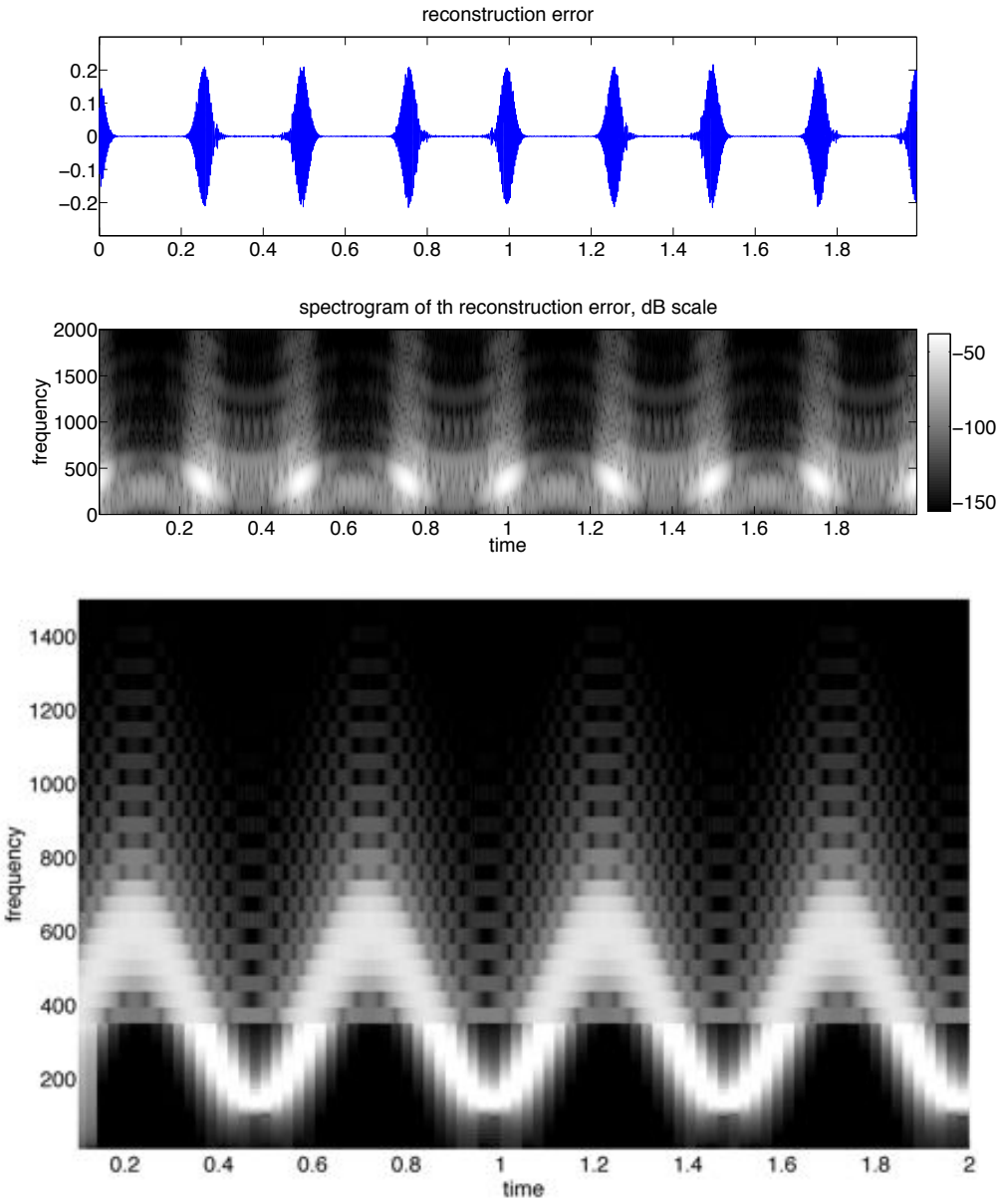

FIGURE 4.13. *Composed spectrogram of a sinusoid with sinusoidal frequency modulation: the signal is analyzed with two different windows, the spectrogram coefficients are weighted with two binary masks and then summed together (see Test 4.4.2). On top, the reconstruction error obtained with the analysis-weight method, and its spectrogram.*

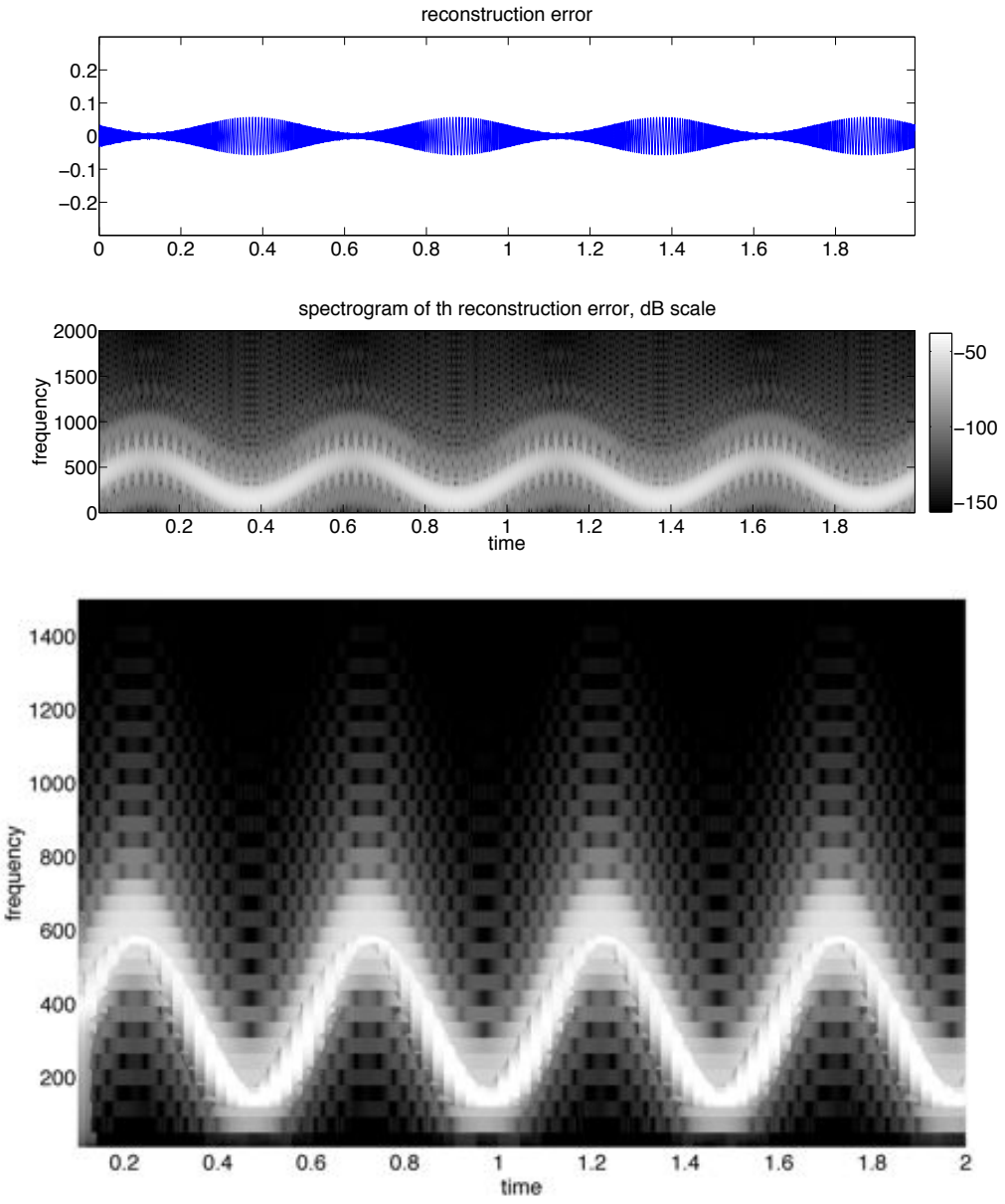

FIGURE 4.14. *Composed spectrogram of a sinusoid with sinusoidal frequency modulation: the signal is analyzed with two different windows, the spectrogram coefficients are weighted with the w_p masks, with overlap 50-650Hz, and then summed together (see Test 4.4.2). On top, the reconstruction error obtained with the analysis-weight method, and its spectrogram.*

CHAPTER 5

Applications and examples

One of the main interest of this work is to establish a concrete bridge, between adaptive techniques for time-frequency analysis and state-of-the-art sound processing methods. In this chapter, we show several examples of applications on real music signals, detailing the improvements we get in their viewing and manipulations.

5.1. Time adaptation

When a music sample contains fast transients together with dense harmonic parts, the choice of a fixed resolution determines a loss in time or frequency precision within the analysis (see Example 5.1.1). In most cases, we can obtain a higher precision by varying the resolution along the time dimension: we would thus choose a smaller window for the analysis of fast transients, and a larger one for dense harmonic parts.

This type of adaptation is achievable with our algorithm, by means of an automatic procedure, with no need of supervision by the user. In music signals, we find several situation where fast transients are alternated with dense harmonic parts: a typical case is when percussions are played together with solo instruments, but even others, as shown in the following example.

Example 5.1.1. Here, we take a guitar solo excerpt of the Flamenco song *Sera Tu Misma Conciencia* by Antonio Fernandez Diaz, played by Paco De Lucia (sound file **ex_DeLucia_2.wav**, standard cd format). Consider the following setup:

- $N = 8$, and the window lengths varies between $len_1 = 1024$ and $len_N = 4096$ points;
- $c = 0.15$ and $d = 2$, and for every window g_s the analysis is calculated with hop size $c \cdot len_s$ and FFT size $d \cdot len_s$;
- the Rényi entropy order considered is $\alpha = 0.3$;
- R covers all the frequency support, and includes 3 time shifts of the largest window g_N; this corresponds to 6144 points and about 0.139 seconds, as the sampling rate SR is 44.1kHz;
- at each step of the algorithm, R is shifted in time, the overlap with the previous position including 2 time shifts of the window g_N; that is, 5120 points and about 0.116 seconds.

In Figure 5.1 we see two spectrograms of the sound considered, calculated with different fixed resolutions, corresponding to the window g_1 and g_N together with their

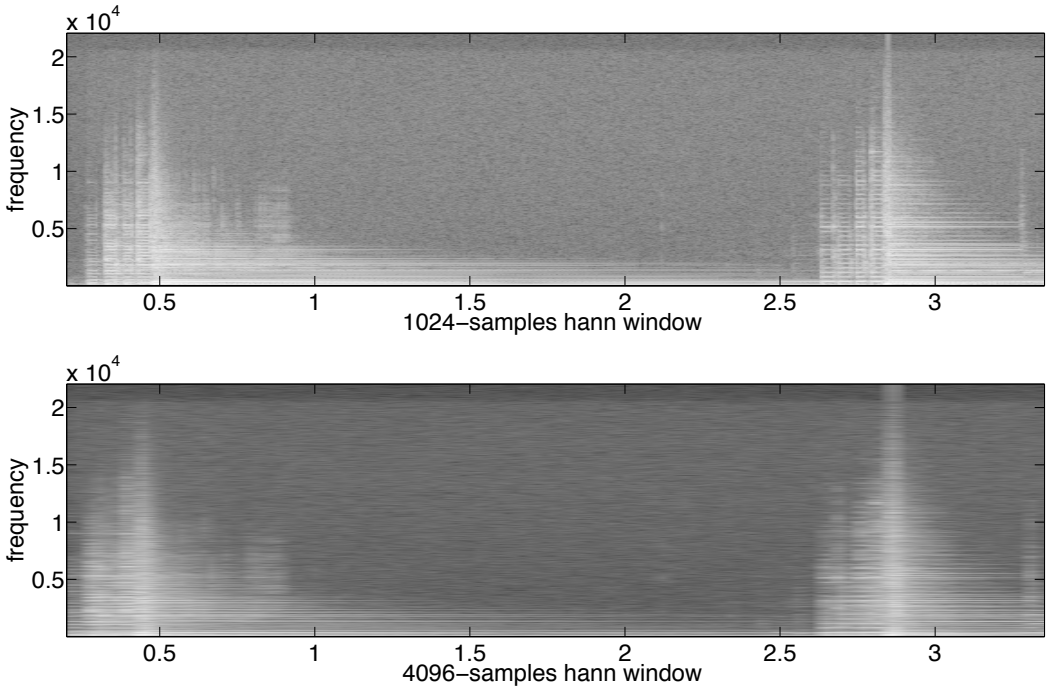

FIGURE 5.1. *Spectrograms of a solo guitar sample (see Example 5.1.1): the two Hanning windows used have lengths 1024 and 4096 points; in both cases the hop size is a quarter of the window length and the FFT size is twice the window length.*

analysis parameters listed above. With the largest window, we see that the fast note repetitions before the two chords are not properly separated: together with the visual blur, this problem affects all the spectral manipulations of the sound. In this and the following examples, we consider in particular time dilatations of the analyzed sounds, performed with a state of the art library, the extended phase vocoder SuperVP [1]. In this case, the different notes are not individually perceivable in the dilated sound, originating an artifact (sound file **ex_DeLucia_2_stand _4096_3.5.wav**). On the other hand, the smallest window has a low frequency resolution, which determines the fusion of close sinusoids in zones with a dense spectrum, like chords; the phase vocoder treats two unresolved sinusoids as a single sinusoid with sinusoidal amplitude modulation: in the dilated sound, the modulation is slowed and a new modulated tone is perceived, thus introducing harmonic distortions (sound file **ex_DeLucia_2_stand_1024_3.5.wav**).

Figure 5.2 shows the adaptive spectrogram automatically calculated by our algorithm, with the parameters specified above: as we see, the window choice at each time location fulfills the need of time or frequency precision, according to zones with transients or harmonic content. The matrix with the best windows selected corresponding to their time location is then used for the same sound dilatation performed with fixed

[1] see http://anasynth.ircam.fr/home/english/software/supervp

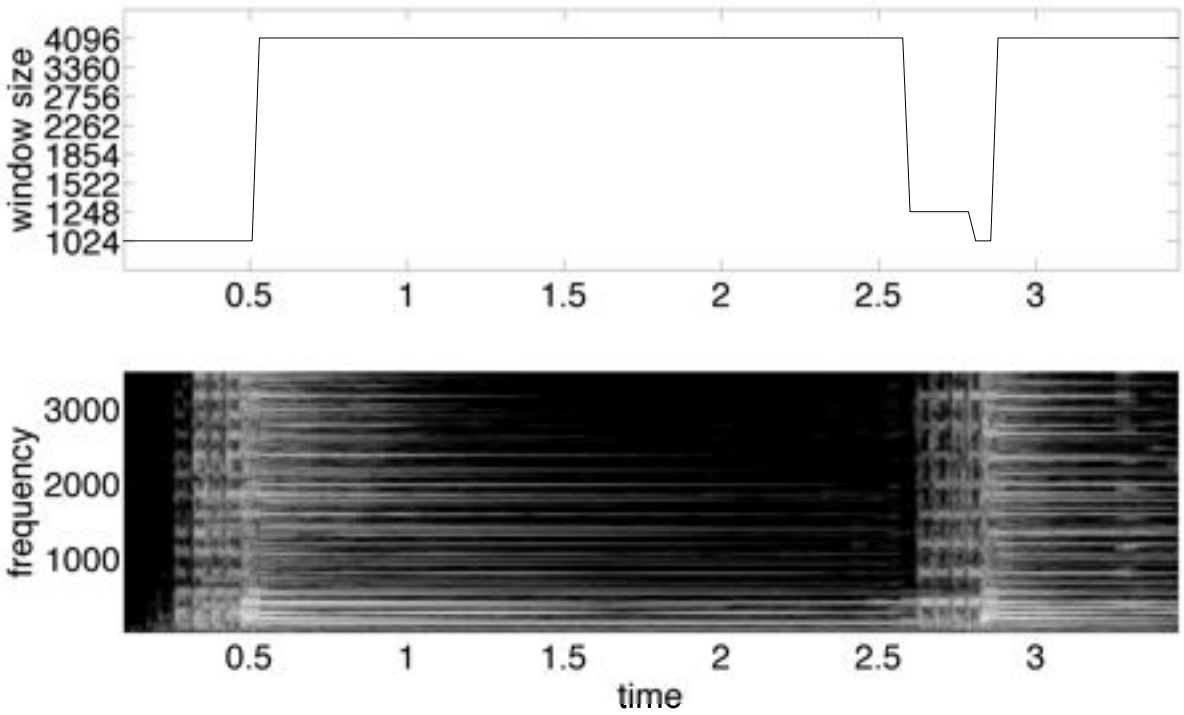

FIGURE 5.2. *Adaptive analysis of a solo guitar sample (see Example 5.1.1); the frequency range is limited to enhance readability: on top, the best window selected by the automatic algorithm is shown for each time location, in correspondence to the part of adaptive analysis that it determines (at the bottom).*

window, which is possible thanks to the advanced options in SuperVP (sound file **ex_DeLucia_2_adapt_ts3.5.wav**). ■

A further situation, where a music sample requires a time-varying resolution, is given by an ensemble with a solo instrument, playing with vibrato in the high frequency range. Here, the density of the ensemble texture requires in general a large window, to provide a good frequency resolution; nevertheless, a window which is too large discretizes the fast frequency modulation, *breaking* the vibrato.

Example 5.1.2. Here, we take an excerpt from the work by Gerard Grisey *Quatre chants pour franchir le seuil*, for soprano and ensemble, sung by Catherine Dubosc: the excerpt starts at 5'25" of the first track, *Prélude I. La mort de l'ange* (sound file **ex_Grisey_2.wav**, standard cd format). Consider the same setup of Example 5.1.1: as before, in Figure 5.3 we see two spectrograms with the smallest and largest fixed resolutions considered. With the smallest window, we can properly view the frequency modulation for all the partials of the singing voice; in the spectrogram with the largest window, this information is nearly lost for frequencies higher than 10kHz. On the other hand, the

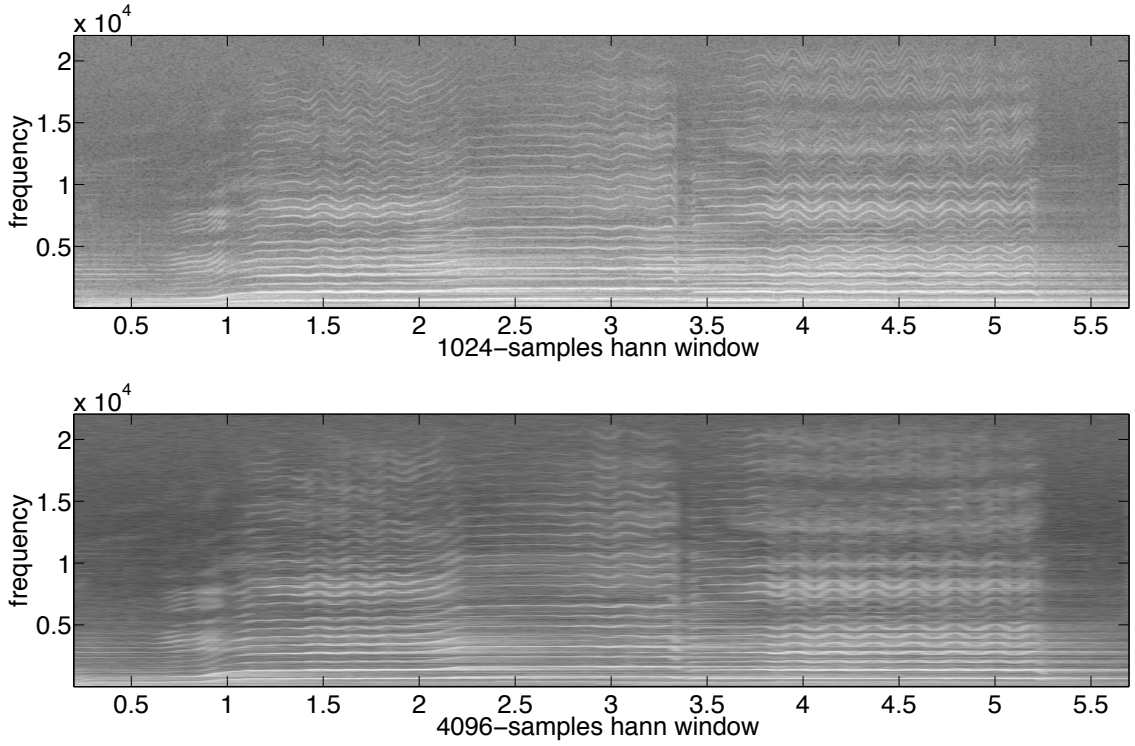

FIGURE 5.3. *Spectrograms of an excerpt of the work for soprano and ensemble (see Example 5.1.1): the two Hanning windows used have lengths 1024 and 4096 points; in both cases the hop size is a quarter of the window length and the FFT size is twice the window length.*

smallest window does not provide a sufficient frequency resolution on the instrumental part. To verify the quality of these analyses, we consider again dilatations of the analyzed sound: with the largest window, the vibrato is separated into alternating close notes (sound file **ex_Grisey_2_stand _4096_3.5.wav**); on the other hand, the smallest window introduces harmonic distortions, in particular on the initial percussive sound (sound file **ex_Grisey_2_stand_1024_3.5.wav**). A further treatment with an intermediate window size is proposed (sound file **ex_Grisey_2_stand_2048_3.5.wav**), which represents a compromise to partially reduce the two different artifacts.

Figure 5.4 shows the adaptive spectrogram automatically calculated by our algorithm: the window choice is chosen at each time location depending on the frequency modulation rate, either on the glissando and the vibrato; but when the voice is not modulated, then the highest frequency resolution is privileged. The improvements in the analysis quality can also be heard in the dilatation with adapted window size (sound file **ex_Grisey_2_mod1_adapt_ts3.5.wav**). ■

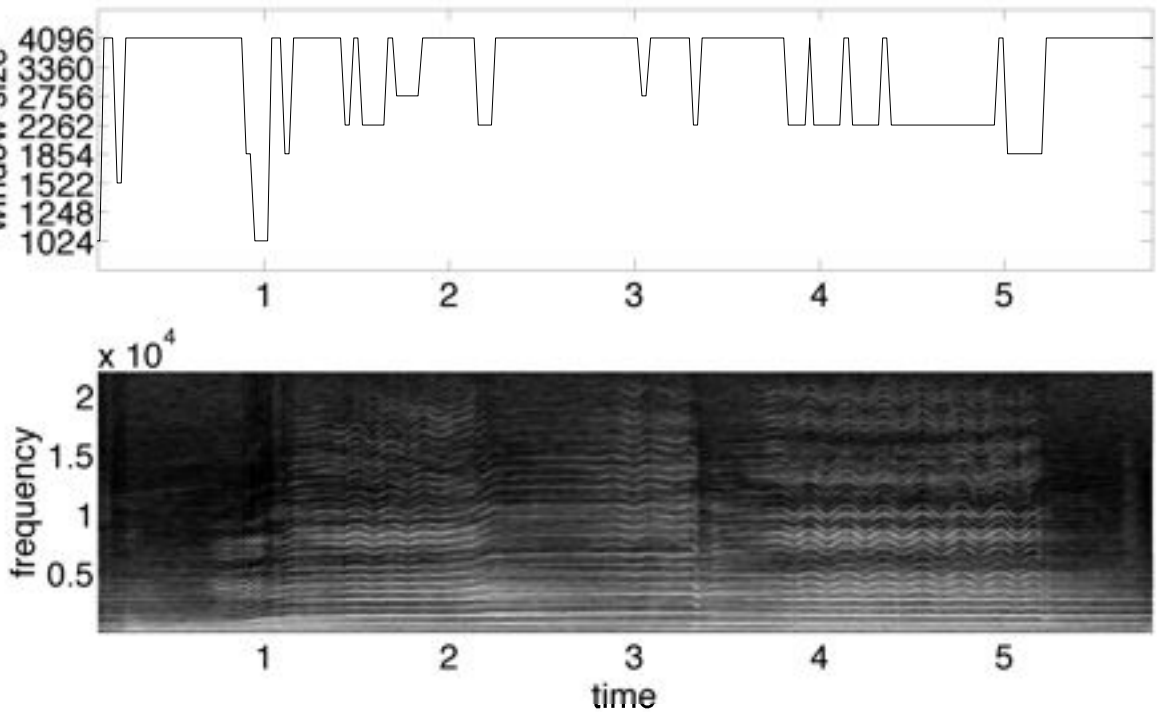

FIGURE 5.4. *Adaptive analysis of an excerpt of the work for soprano and ensemble (see Example 5.1.2): on top, the best window selected by the automatic algorithm is shown for each time location, in correspondence to the part of adaptive analysis that it determines (at the bottom).*

5.2. Time adaptation with different masks

When a music signal presents different components at the same time, the choice of an individual resolution, only depending on the time position, may not be sufficient to resolve them. As an example, we consider a sound sample (sound file **ex_b66_1.wav**, standard cd format) where a tabla is playing, an Indian percussion instrument of the membranophone family; at time 2.22" a sitar also plays, a plucked stringed instrument. The tabla presents, at once, fast transients and long tones in the mid-low frequency range, even with fast frequency modulations played by the thumb on the larger drum. Together with the melody played on the metal strings of the sitar, the music which is originated has a highly heterogeneous spectrum: the best resolution to resolve individual components varies with both the time and the frequency localization of the analysis atom (see Figure 5.5 for the spectrograms with fixed resolutions).

For analysis purposes, a possible choice is to privilege the readability of the harmonic structures; with the procedure described in Sections 4.3.1 and 4.3.2, we can use a binary mask to select the frequency region considered in the adaptation routine. Figure 5.6 shows the adapted spectrogram we obtain analyzing this sound with the

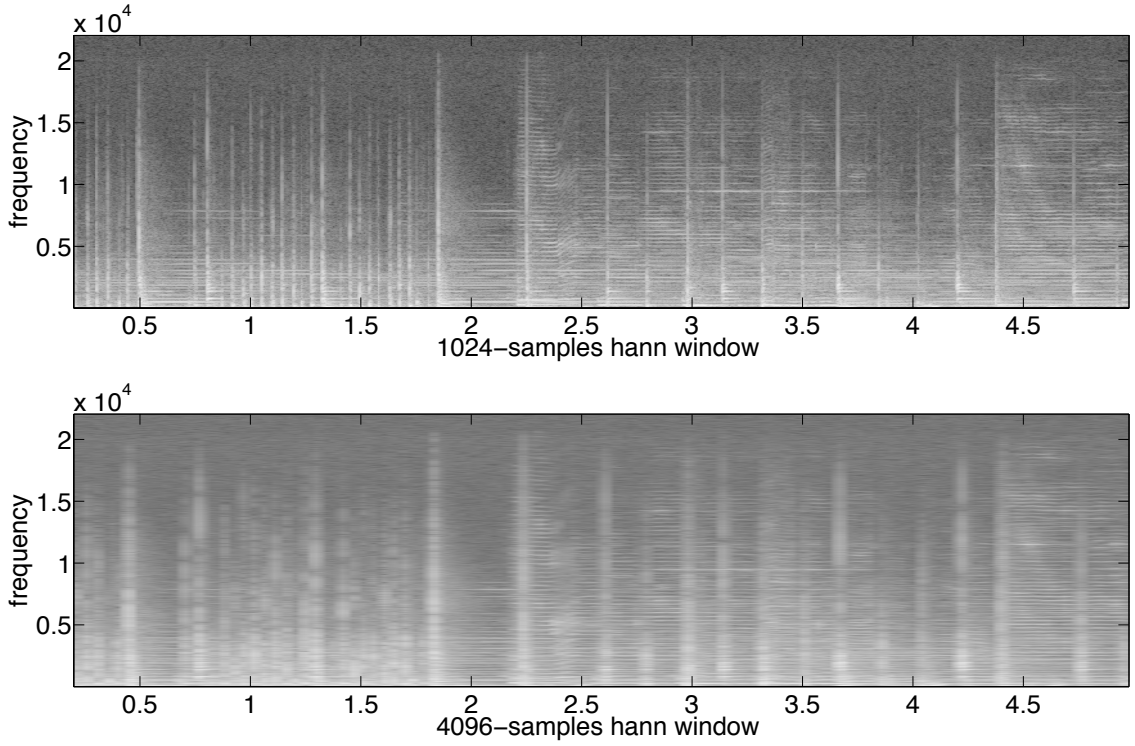

FIGURE 5.5. *Spectrograms of a sound sample with tabla and sitar (see Section 5.2): the two Hanning windows used have lengths 1024 and 4096 points; in both cases the hop size is a quarter of the window length and the FFT size is twice the window length.*

same setup of Example 5.6; the difference is that al the spectrograms are weighted with a binary mask setting to 0 the coefficients above 1kHz before the entropy evaluation. The chosen mask rises a window choice adapted to the frequency area where the first harmonics of the two instruments are predominant. Nevertheless, we see that within the parts where fast transients are predominant, or exclusive, the best window selected is still small, as required: this is a major advantage with respect to analysis methods where different windows are a priori associated to certain region depending on the frequency renge.

5.3. Time-frequency adaptation

The music sample considered in the previous section, where a tabla and a sitar play together, is an example of the need for spectral processing techniques with variable time-frequency resolution: a fixed resolution, or a time-dependent resolution like the one we have introduced, would not be appropriate in certain frequency regions. The

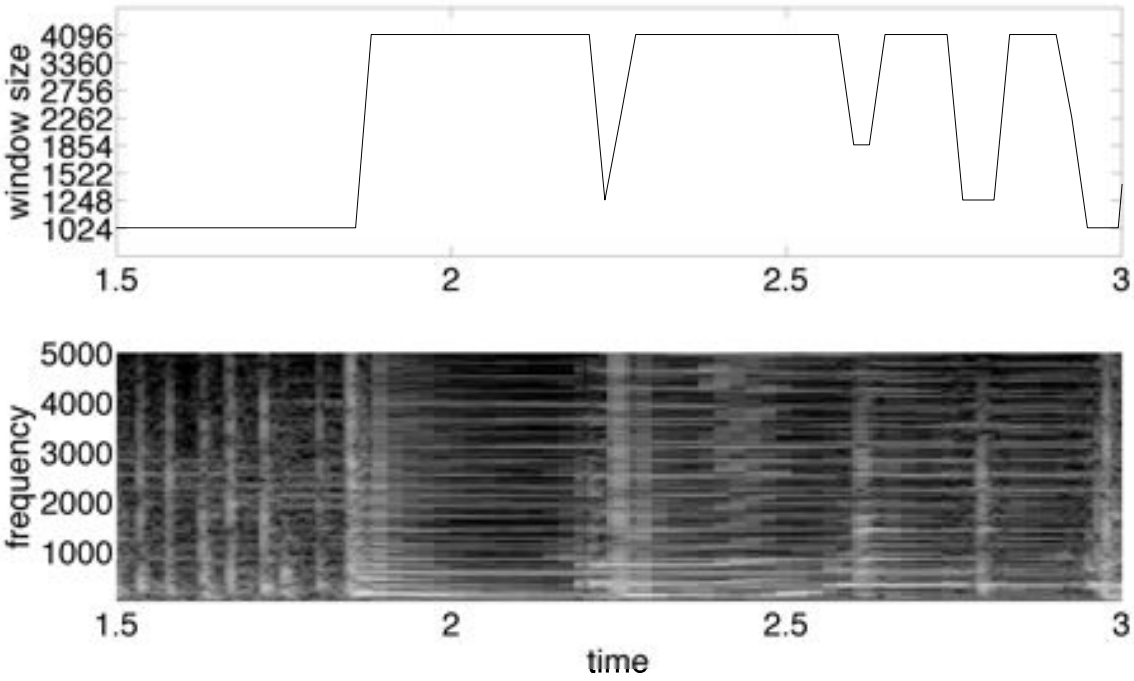

FIGURE 5.6. *Adaptive weighted analysis of a sound sample with tabla and sitar (see Section 5.2): the binary mask for the adaptation set to 0 the coefficients above 1kHz; the frequency range is limited to enhance readability: on top, the best window selected by the automatic algorithm is shown for each time location, in correspondence to the part of adaptive analysis that it determines (at the bottom).*

SuperVP library [2] provides an advanced automatic adaptation of the window size, which is based on a previous estimation of the fundamental frequency of the analyzed sound (see [**Vinet and al, 2011**]): in this case, such an estimation is not possible, as the sound is not monophonic.

The adaptation we introduce here is based on the method detailed in Subection 4.3.3: in particular, we consider, for the adaptation routine, two complementary binary masks with cut frequency at 1kHz. The adaptive analysis obtained on the low frequency band has been shown in Figure 5.6; the complementary analysis, where the window selection is adapted to the higher frequency band, is shown in Figure 5.7.

We see that the overall profile remains the same, in particular on the fast transients part at the beginning; but there are some important differences; in particular, the frequency modulation of the first sitar note, at time 2.5". When high frequencies are masked, the partials of the sitar taken into account are the first ones, for which the modulation range is limited: a large window is chosen, privileging the frequency precision, but still guaranteeing the continuity of the modulation below 1kHz; but

[2] see http://anasynth.ircam.fr/home/english/software/supervp

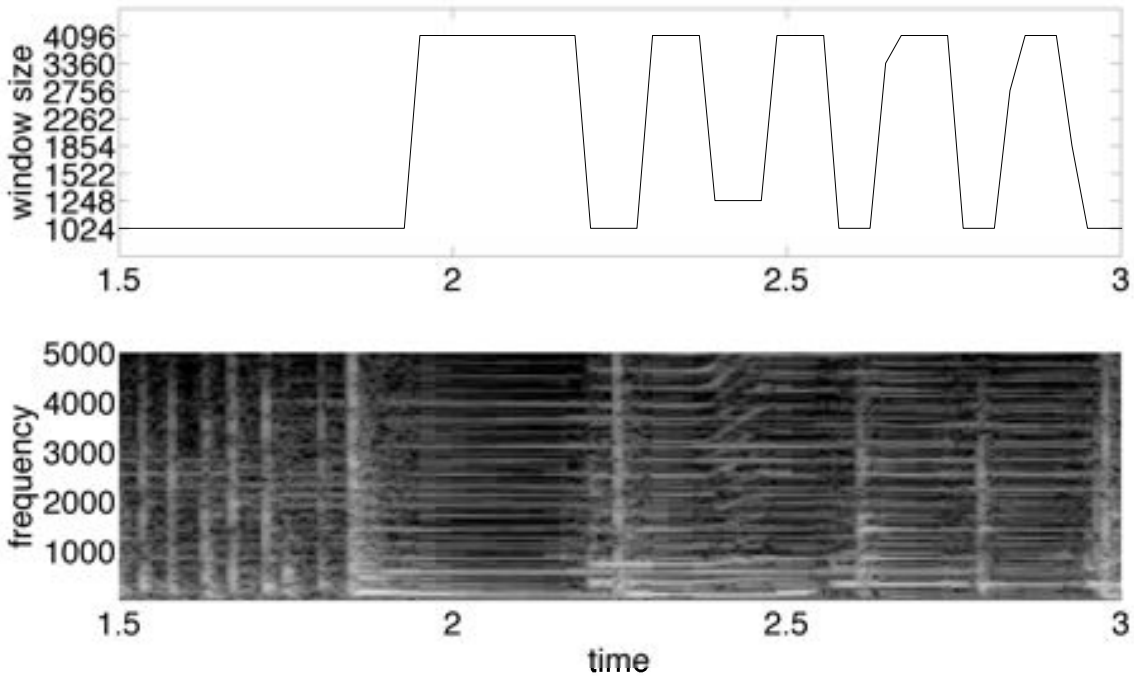

FIGURE 5.7. *Adaptive weighted analysis of a sound sample with tabla and sitar (see Sections 5.2 and 5.3): the binary mask for the adaptation set to 0 the coefficients below 1kHz; the frequency range is limited to enhance readability: on top, the best window selected by the automatic algorithm is shown for each time location, in correspondence to the part of adaptive analysis that it determines (at the bottom).*

as we see in the upper part of Figure 5.6, the modulation is highly blurred at the frequencies above. On the other hand, the continuity of the modulation is conveniently provided by the complementary analysis, where a small window is chosen, as seen in Figure 5.7. Other differences concern the way the transients are treated in the two cases, providing a higher time or frequency precision depending on the considered mask. The resulting composed analysis with variable time-frequency resolution is shown in Figure 5.8.

Table 5.1 shows the reconstruction error obtained on this music signal, with the analysis-weight and extended weight methods detailed in Section 4.4. Here, we see that even with a larger overlap the reduction of the error is soft, as the overlap is chosen regardless of the local spectral energy: further developments of this framework should aim to an efficient method to adaptively deal with overlaps; once individuated a desired frequency band, the optimal limits should be chosen, within a certain frequency range, in order to minimize the signal spectral energy where the first coefficients are set to 0.

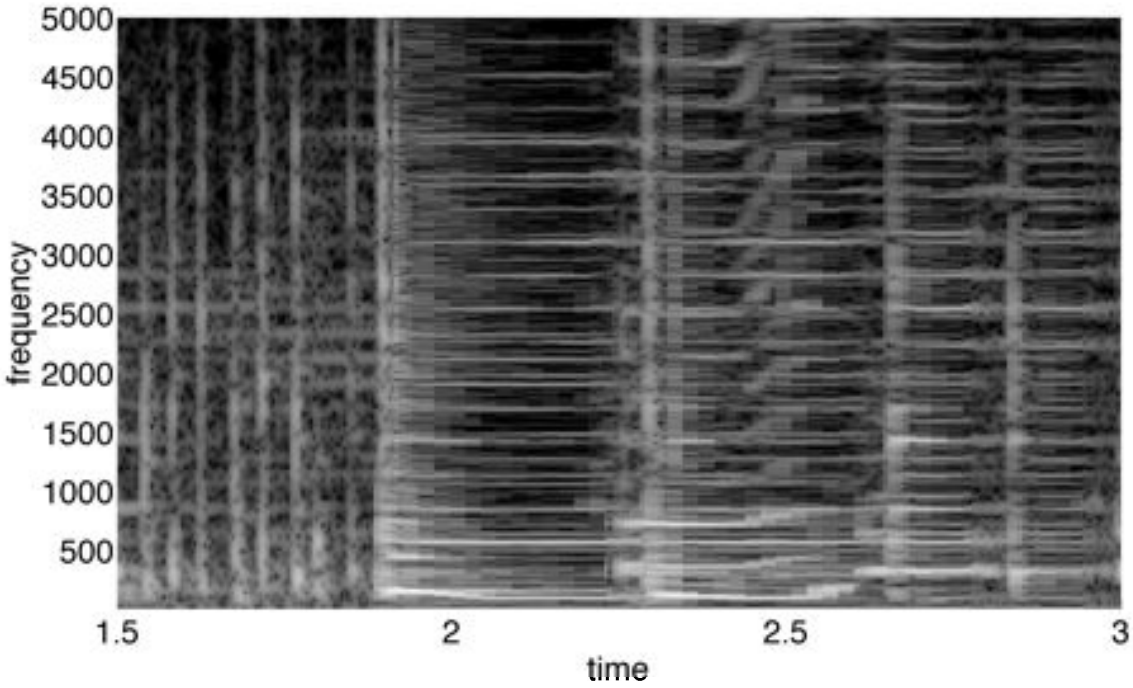

FIGURE 5.8. *Adaptive analysis of a sound sample with tabla and sitar (see Section 5.3); the frequency range is limited to enhance readability: the resolution is adapted in time and in two frequency bands, above and below 1kHz.*

The sound file **rec_ex_b66_1.wav** is reconstructed with the analysis-weight method, from weighted analysis with linear frequency cross fade at 750-1250Hz. The reconstruction error (sound file **er_ex_b66_1_42dB.wav**, amplified by 42dB) and its spectrogram are shown in Figure 5.9: comparing this figure with the ones of the adapted analyses for the two different bands (Figures 5.6 and 5.7), we see that the error energy is concentrated at the time location where the window choice differs within the two bands, and within a frequency range determined by the overlap of the two masks.

Even if the error is quite small, fast FFT-based methods, like the ones we define, cannot reduce it till the perfect reconstruction: but still, as the aim of these representations is to ameliorate sound processing algorithms, the perceived quality of the reconstruction is determinant, rather than an objective error measure. Therefore, further investigations should characterize the error from a perceptive point of view, performing tests on the perceived quality of the reconstruction.

At present, there are no common sound processing techniques dealing with time-frequency adapted analyses like the ones we introduce: therefore, for this case it is not possible to give examples based on sound manipulations. Nevertheless, our methods are conceived to allow for extensions of existing algorithms: the processing should simply be done iteratively on the different frequency bands, according to the weighted

TABLE 5.1. *Reconstruction error when f is a sound sample with tabla and sitar: the masks are indicated on the left, together with their frequency significant values.*

Weight method	Parameters	peak	rms
Binary mask	$\omega_{cut} = 1\text{kHz}$	0.0047	$4.0864 \cdot 10^{-04}$
Linear cross	$\omega_1 = 750\text{Hz}$ $\omega_2 = 1.25\text{kHz}$	0.0034	$2.3890 \cdot 10^{-04}$
Extended weight	$\omega_1 = 750\text{Hz}$ $\omega_2 = 1.25\text{kHz}$	0.0197	0.0018
Linear cross	$\omega_1 = 500\text{Hz}$ $\omega_2 = 1.5\text{kHz}$	0.0037	$1.9932 \cdot 10^{-04}$
Extended weight	$\omega_1 = 500\text{Hz}$ $\omega_2 = 1.5\text{kHz}$	0.0162	0.0013

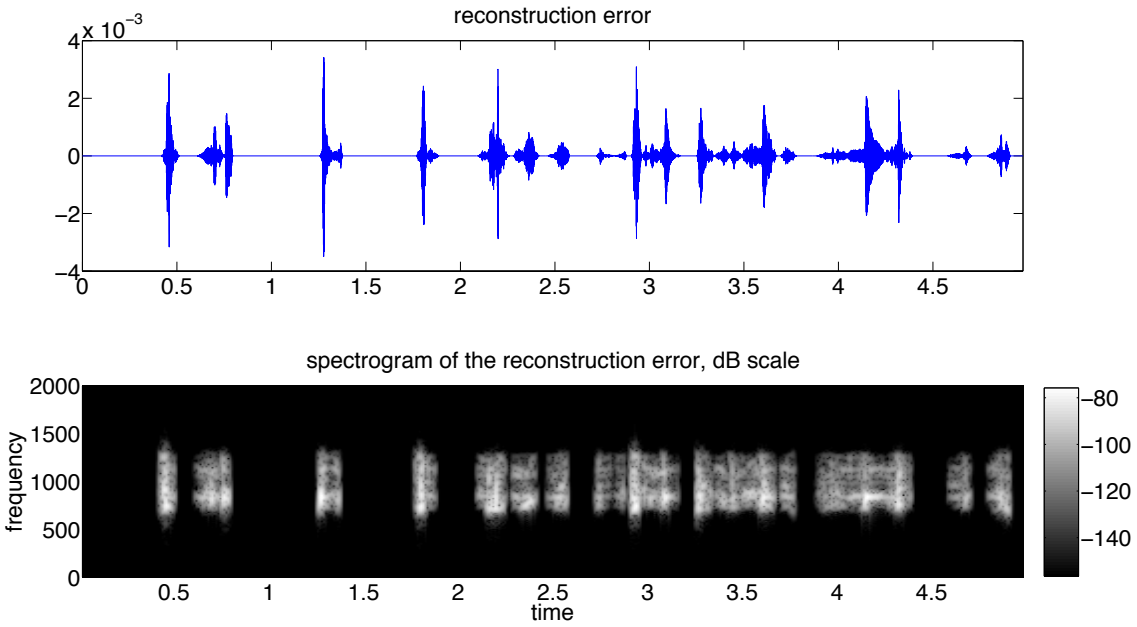

FIGURE 5.9. *Spectrogram of the reconstruction error given by the analysis-weight approach, on a sound sample with tabla and sitar (see Section 5.3); the frequency range is limited around the overlap of the weighting masks, from 750Hz to 1.25kHz.*

analyses; then, the fundamental task would be to conceive appropriate strategies to treat the overlapping zones, depending on the specific sound treatment. The interest

of these analyses is thus related to the improvements achievable with processing techniques fully exploiting them: the optimal local time-frequency resolution guarantees a solid ground to develop adaptive high-quality transformations.

5.4. Spectral change detection algorithm

We show here an application of the detection algorithms with the measures defined: the first algorithm we analyze has the same operations for the K divergence and Rényi information (3.6.6): we calculate the spectrogram of a signal with a 1024-samples Hamming window, 768-samples overlap and 2048-points FFT size; we obtain a mean spectrum taking the first 20 analysis frames, and calculate the divergence of the next frame with respect to the mean spectrum. Once we have the first divergence value, we shift the mean spectrum of one analysis frame and consider the following 20 frames, then calculate the divergence between the new mean spectrum and the following frame. At this point, if the ratio between the last divergence value and the previous exceeds a certain threshold, a change is detected at the incoming frame; otherwise the procedure goes on. The second algorithm is a variation of the first one based on entropy prediction: once obtained the spectrogram of the signal, we calculate the Rényi entropy of the vector composed of its first 6 analysis frames; then we consider the next frame and set the predicted entropy value according to (3.6.10). We calculate the actual entropy of the vector obtained adding the new frame to the previous ones, and if the ratio between this value and the predicted one exceeds a certain threshold, a change is detected. Then the procedure goes on as in the previous case.

The Rényi prediction shows a slightly better accuracy at the price of a higher computational cost; this is due to the larger dimensions of the vectors managed in the entropy calculus. The tuning of the α parameter gives interesting results: as seen in figure 3.1, higher values rise the difference between the entropies of a peaky distribution and a flat one; thus we expect in general a more refined detection increasing α, leaving the threshold unchanged. The signal we analyze is a speech fragment of a mail voice in French language, *Vénitienne et lui suce la bouche un quart d'heure*. We assume two references: an automatic phoneme segmentation for French language based on Hidden Markow Model [**Lanchantin et al., 2008**], and a voiced-unvoiced classification obtained with a PSOLA-based algorithm [**Mattheyses et al., 2006**]: they identify the major spectral changes in this kind of signal, so we expect our detection to confirm them. We are not interested in whether a marker belongs to one selection or the other, as this could be established in a later classification step. As we see at the top of figure 5.10, the Rényi prediction with $\alpha = 0.2$ identifies all the voiced-unvoiced transitions in both senses except at time 2.5, and a large part of phonemes. If we need a less refined detection, setting the α parameter to 0.05 (bottom of figure 5.10) preserves the detection of all the unvoiced-voiced transitions, while discarding all the phonemes and the voiced-unvoiced transitions. Both the measures provide a better detection with respect to the K divergence, which shows a higher number of unexpected markers.

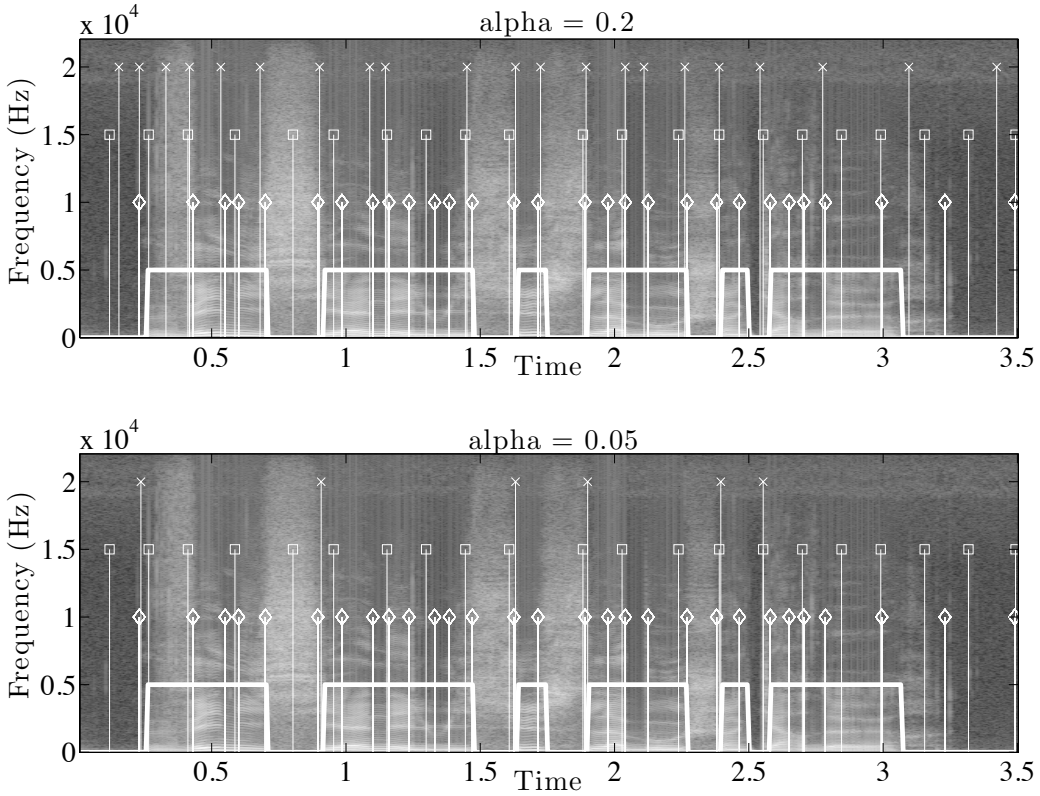

FIGURE 5.10. *Detections obtained with different methods on a speech fragment in French language;* ***cross markers****: Rényi entropy prediction method, on top with* $\alpha = 0.2$*, at the bottom with* $\alpha = 0.05$*;* ***square markers****: K divergence;* ***diamond markers****: HMM-based phoneme segmentation method;* ***bold line****: PSOLA voiced-unvoiced classification, 0 is unvoiced.*

CHAPTER 6

Conclusions and outlooks

In Section 1.7 we have listed the main original contributions of this work to the state of the art; here, we focus on each one of them, summarizing the major novelties we introduce and the perspectives they outline. In general terms, this work traces a straight line, starting from advanced results of Gabor frames theory and going to high-quality sound processing techniques: the automatic framework we realize, which is based on adaptive representations and the related reconstruction methods, is a concrete base to provide time-frequency sound transformations with adaptive strategies. Starting from the straightest algorithm we design, which is the automatic time-adaptation of the window size for the STFT, some of the introduced methods will be computationally optimized and integrated within AudioSculpt [1].

6.1. Automatic adaptation of the spectrogram

The adaptation we define is based on the choice of best local resolutions: the strategy we adopt is the same of [**Jaillet, 2005**, **Jaillet and Torrésani, 2007**], comparing different spectrograms and selecting the one which locally gives the minimal Rényi entropy. Concerning this measure, in Section 3.3 we give new results on the existence of Rényi entropy measures of spectrograms in the continuous case, extending the results of [**Baraniuk et al., 2001**]; we give also new results about the convergence of discrete versions of these measures to their continuous one, when the sampling grid becomes infinitely dense. The formulation of these results is given for Rényi entropies, but they apply more generally to all measures based on time-frequency integrals of real powers of the STFT.

The nodal point for the entropy-based adaptation criterium is the dependance on the α parameter: in Sections 3.4 and 3.6, we deduce some properties about the Rényi entropies and the parameter they depend on, which are useful for its interpretation in applicative contexts. For the applications we have shown in Chapter 5, different values between 0 and 1 are considered, which increase the importance of smaller spectral coefficients for the entropy evaluation. The characterization we give is mainly application-oriented, giving a useful insight on the tuning of α when entropy measures are applied to the spectrogram: a complete theoretical investigation on its role, when dealing with larger classes of TFRs, still needs to be established.

[1] see http://anasynth.ircam.fr/home/english/software/audiosculpt

As an alternative to the entropy-based criterium, in Section 3.7 we introduce a further measure to determine the best local resolution of a spectrogram, based on the classification algorithm in [**Röbel et al., 2004**]: its features are analyzed by means of some tests in Section 4.2, showing that it constitutes a valid and interpretable strategy for the spectrogram adaptation. An appropriate validation of this method requires a further experimentation stage: the efficient communication between the classification algorithm and our adaptive framework will need major implementation changes, and is thus left as a task for future research activities in this direction.

6.2. Reconstruction from adapted analyses

In Chapter 2 we propose two novel reconstruction methods, for adaptive analyses with resolution varying among both time and frequency: the two approaches are indicated as *extended weight* and *filter bank* (see Sections 2.5 and 2.6). For the latter, an upper bound for the reconstruction error is analytically determined, in both the cases of analyses based on stationary or nonstationary Gabor frames; in Subsection 2.6.2, we then define a further variation of this method, considering Gabor multipliers instead of filters. In this case, the estimate (2.6.16) on the reconstruction error needs to be further refined: most of all, the positive lower bound of the error should be determined, depending on the windows and lattices used. We envisage that the ongoing work (an article is in preparation in this sense [**Engelputzeder, 2011**, **Balazs et al., 2012**]), about the approximation of convolution operators by means of Gabor multipliers, could clarify the relation between the error due to the truncation expansion of the filter-bank approach, and the one introduced by the approximation of filters with Gabor multipliers.

The latter algorithm, that we indicate as *analysis-weight* approach, and the extended weight one are implemented in our adaptive framework: in Section 4.4, they are applied on several basic signals, while in Section 5.3 a real-world sound is treated.

We have implemented new Matlab code for the whole framework of analysis, automatic adaptation and reconstruction; the different FFT-based reconstruction functions, for the extended weight and analysis-weight cases, are new extensions of the existing ones (see [**Balazs et al., 2011**, **Søndergaard et al.,**]).

6.3. Spectral change detection

Our investigation of the Rényi entropies properties has lead to a further application, in the domain of spectral change detection in audio streams: in Section 3.6, we define a novel method with promising results in the automatic segmentation of a spoken voice (see Section 5.4). Like all the algorithms in this work, this method allows a fast implementation, whose main computational cost is due to the FFT of the windowed signal. This speed, which guarantees a segmentation in pseudo-real time, has the disadvantage of a low robustness to noise, speaker's timbre and audio quality; moreover, being based on Rényi entropies, the dependance of the segmentation on the α parameter has to be taken into account. An ongoing research stage within

the Analysis/Synthesis Team at IRCAM, focused on unsupervised real time syllabic segmentation of spoken voice, is extending this method in several directions:

- a refined segmentation could be obtained by imposing rules defined by the target: that is, a set of constraints on the possible segmentations, deduced by the characterization of syllables in the considered language;

- with several parallel calls of the algorithm, on a same signal with different parameters, we could analyze the different outputs obtained, and deduce a final output exploiting the information coming from the individual ones;

- Rényi entropies are defined for probability distributions: instead of applying them on a spectrogram, we could define distributions by an appropriate collection of audio descriptors, better suited for the speech; then, the change detection would take place at a descriptor level, and the single coefficients in the distributions would have a more readable relation with the analyzed signal.

Bibliography

[Antoine and Balazs, 2011] Antoine, J.-P. and Balazs, P. (2011). Frames and semi-frames. *Journal of Physics A Mathematical General*, 44(20):205201.

[Auger and Flandrin, 1995] Auger, F. and Flandrin, P. (1995). Improving the readability of time-frequency and time-scale representations by the method of reassignment. *IEEE Trans. Signal Processing*, 43(5):1068–1089.

[Babenko, 1962] Babenko, K. I. (1962). An inequality in the theory of fourier integrals. In *Amer. Math. Soc. Transl.*, 2, pages 115–128.

[Balan et al., 2006] Balan, R., Casazza, P., and Edidin, D. (2006). On signal reconstruction without phase. *Appl. Comput. Harmon. Anal.*, 20(3):345–356.

[Balazs et al., 2010] Balazs, P., Antoine, J.-P., and Gribośs, A. (2010). Weighted and controlled frames: mutual relationships and first numerical properties. *Int. J. Wav, Mult. Info. Proc.*, 8(1):109–132.

[Balazs et al., 2011] Balazs, P., Dörfler, M., Jaillet, F., Holighaus, N., and Velasco, G. (2011). Theory, implementation and applications of nonstationary gabor frames. *Journal of Computational and Applied Mathematics*, 236(6):1481 – 1496.

[Balazs et al., 2012] Balazs, P., Engelputzeder, N., and Liuni, M. (2012). Representation of linear time invariant filters by Gabor multipliers. *in preparation*, -:–.

[Baraniuk et al., 2001] Baraniuk, R., Flandrin, P., Janssen, A., and Michel, O. (2001). Measuring Time-Frequency Information Content Using the Rényi Entropies. *IEEE Trans. Info. Theory*, 47(4):1391–1409.

[Basu, 2003] Basu, S. (2003). A linked-HMM model for robust voicing and speech detection. In *Proc. of ICASSP03*, pages I–816 – I–819, Hong Kong, China.

[Beck and Schlögl, 1993] Beck, C. and Schlögl, F., editors (1993). *Thermodynamics of chaotic systems*. Cambridge University Press, Cambridge, Massachusetts, USA.

[Beckner, 1975] Beckner, W. (1975). Inequalities in fourier analysis. *The Annals of Mathematics*, 102(1):159–182.

[Brezis, 1983] Brezis, H., editor (1983). *Analyse fonctionnelle : théorie et applications*. Éditions Masson, Paris, France.

[Casazza, 1999] Casazza, P. G. (1999). The Art of Frame Theory. *ArXiv Mathematics e-prints*.

[Chadabe, 1997] Chadabe, J. (1997). *Electric Sound: The Past and Promise of Electronic Music*. Prentice-Hall, Upper Saddle River, New Jersey.

[Chen, 1998] Chen, S. S. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

[Christensen, 2003] Christensen, O., editor (2003). *An Introduction To Frames And Riesz Bases*. Birkhäuser, Boston, Massachussets, USA.

[Clarke, 2006] Clarke, J. (2006). Jonathan Harvey's 'Mortuos Plango, Vivos Voco'. In *Analytical Methods of Electroacoustic Music*, pages 111–143. Routledge, London, England.

[Cohen, 1989] Cohen, L. (1989). Time-frequency distributions - a review. *Proceedings of the IEEE*, 77(7):941–981.

[Cohen, 1995] Cohen, L., editor (1995). *Time-Frequency Analysis*. Prentice-Hall, Upper Saddle River, New Jersey, USA.

[Coifman et al., 1992] Coifman, R. R., Meyer, Y., and Wickerhauser, V. (1992). Wavelet analysis and signal processing. In *In Wavelets and their Applications*, pages 153–178.

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

[Daubechies, 1990] Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Info. Theory*, 36(5):961–1005.

[Daubechies, 1992] Daubechies, I., editor (1992). *Ten Lectures on Wavelets*. PA: Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[Daubechies et al., 1986] Daubechies, I., Grossmann, A., and Meyer, Y. (1986). Painless nonorthogonal expansions. *J. Math. Phys.*, 27:1271–1283.

[Davis et al., 1997] Davis, G., Mallat, S., and Avellaneda, M. (1997). Adaptive greedy approximations. *Constructive Approximation*, 13:57–98. 10.1007/BF02678430.

[Dörfler, 2002] Dörfler, M. (2002). *Gabor Analysis for a Class of Signals called Music*. PhD thesis, Institut für Mathematik der Universität Wien.

[Dörfler, 2011] Dörfler, M. (2011). Quilted Gabor frames - A new concept for adaptive time-frequency representation. *Advances in Applied Mathematics*, 47(4):668 – 687.

[Dörfler and Torrésani, 2007] Dörfler, M. and Torrésani, B. (2007). Spreading function representation of operators and Gabor multiplier approximation. In *Proceedings of SAMPTA07*. NuHAG;MOHAWI.

[Dörfler and Torresani, 2010] Dörfler, M. and Torresani, B. (2010). Representation of operators in the time-frequency domain and generalized Gabor multipliers. *J. Fourier Anal. Appl.*, 16(2):261–293.

[Engelputzeder, 2011] Engelputzeder, N. (2011). *Linear Time Variant Systems and Gabor Riesz Bases*. PhD thesis, University Vienna.

[Feichtinger and Strohmer, 1998] Feichtinger, H. and Strohmer, T., editors (1998). *Gabor analysis and algorithms*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston Inc., Boston, MA.

[Feichtinger and Nowak, 2002] Feichtinger, H. G. and Nowak, K. (2002). A first survey of gabor multipliers. In *Advances in Gabor Analysis. Birkhauser*, pages 99–128. Birkhäuser.

[Flanagan, 1966] Flanagan, J. L. (1966). Phase vocoder. *the Bell System Technical Journal*, pages 1493–1509.

[Flandrin, 1999] Flandrin, P., editor (1999). *Time-Frequency/ Time-Scale Analysis*. Academic Press, San Diego, California, USA.

[Foote, 2002] Foote, J. (2002). Automatic audio segmentation using a measure of audio novelty. In *Proc. of ICME2000*, volume 1, pages 452–455, New York, NY.

[Gabor, 1946] Gabor, D. (1946). Theory of Communication. *IEEE Jour. Inst. Electrical Engineers*, 93(26):429–441.

[Gribonval and Nielsen, 2007] Gribonval, R. and Nielsen, M. (2007). Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22(3):335 – 355.

[Griffin and Lim, 1984] Griffin, D. and Lim, J. (1984). Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Trans. Acoust. Speech Signal Process.*, 32(2):236–242.

[Gröchenig, 2001a] Gröchenig, K. (2001a). *Foundations of time–frequency analysis*. Birkhäuser, Boston.

[Gröchenig, 2001b] Gröchenig, K., editor (2001b). *Foundations of Time-Frequency Analysis*. Birkhäuser, Boston, Massachusetts, USA.

[Harris, 1978] Harris, F. (1978). On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51 – 83.

[Harvey, 1981] Harvey, J. (1981). "Mortuos Plango, Vivos Voco": a Realization at IRCAM. *Computer Music Journal*, 5(4):22–24.

[Jaillet, 2005] Jaillet, F. (2005). *Représentation et traitement temps-fréquence des signaux audionumériques pour des applications de design sonore*. PhD thesis, Université de la Méditerranée - Aix-Marseille II.

[Jaillet et al., 2009] Jaillet, F., Balazs, P., Dörfler, M., and Engelputzeder, N. (2009). Nonstationary Gabor Frames. In *Proc. of SAMPTA'09*, Marseille, France.

[Jaillet and Torrésani, 2007] Jaillet, F. and Torrésani, B. (2007). Time-frequency jigsaw puzzle: adaptive and multilayered Gabor expansions. *International Journal for Wavelets and Multiresolution Information Processing*, 1(5):1–23.

[Jones and Baraniuk, 1994] Jones, D. and Baraniuk, R. (1994). A simple scheme for adapting time-frequency representations. *IEEE Trans. Signal Processing*, 42(12):3530–3535.

[Jones and Baraniuk, 1995] Jones, D. and Baraniuk, R. (1995). An adaptive optimal-kernel time-frequency representation. *Signal Processing, IEEE Transactions on*, 43(10):2361 –2371.

[Kemp et al., 2000] Kemp, T., Schmidt, M., Westphal, M., and Waibel, A. (2000). Strategies for automatic segmentation of audio data. In *Proc. of ICASSP2000*, volume 3, pages 1423–1426, Istanbul, Turkey.

[Lanchantin et al., 2008] Lanchantin, P., Morris, A. C., Rodet, X., and Veaux, C. (2008). Automatic phoneme segmentation with relaxed textual constraints. In *Proc. of LREC08*, Marrakech, Maroc.

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

[Lin, 2002] Lin, J. (2002). Divergence measures based on the shannon entropy. *IEEE Trans. Info. Theory*, 37(1):145–151.

[Liuni et al., 2011a] Liuni, M., Balazs, P., and Röbel, A. (2011a). Sound analysis and synthesis adaptive in time and two frequency bands. In *Proc. of DAFx11*, Paris, France.

[Liuni et al., 2010] Liuni, M., Röbel, A., Romito, M., and Rodet, X. (2010). A reduced multiple Gabor frame for local time adaptation of the spectrogram. In *Proc. of DAFx10*, pages 338 – 343, Graz, Austria.

[Liuni et al., 2011b] Liuni, M., Röbel, A., Romito, M., and Rodet, X. (2011b). An entropy based method for local time-adaptation of the spectrogram. In Ystad, S., Aramaki, M., Kronland-Martinet, R., and Jensen, K., editors, *Exploring Music Contents*, volume 6684 of *Lecture Notes in Computer Science*, pages 60–75. Springer Berlin / Heidelberg.

[Liuni et al., 2011c] Liuni, M., Röbel, A., Romito, M., and Rodet, X. (2011c). Rényi information measures for spectral change detection. In *Proc. of ICASSP11*, Prague, Czech Republic.

[Lukin and Todd, 2006] Lukin, A. and Todd, J. (2006). Adaptive Time-Frequency Resolution for Analysis and Processing of Audio. 120th Audio Engineering Society Convention, Paris, France, May 2006. http://graphics.cs.msu.ru/en/publications/text/LukinTodd.pdf.

[Mallat, 1999] Mallat, S., editor (1999). *A wavelet tour on signal processing*. Academic Press, San Diego, California, USA.

[Martin and Flandrin, 1985] Martin, W. and Flandrin, P. (1985). Wigner-ville spectral analysis of nonstationary processes. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(6):1461 – 1470.

[Mattheyses et al., 2006] Mattheyses, W., Verhelst, W., and Verhoeve, P. (2006). Robust Pitch Marking for Prosodic Modification of Speech using TD-PSOLA. In *Proc. of SPS-DARTS 2006*, pages 43–46, Antwerp, Belgium.

[Matusiak and Eldar, 2010] Matusiak, E. and Eldar, Y. C. (2010). Sub-Nyquist sampling of short pulses: Part i. http://arxiv.org/abs/1010.3132v1.

[Matz and Hlawatsch, 1998] Matz, G. and Hlawatsch, F. (1998). Time-frequency transfer function calculus (symbolic calculus) of linear time-varying systems (linear operators) based on a generalized underspread theory. *Journal of Mathematical Physics*, 39(8):4041–4070.

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

[Moorer, 1975] Moorer, J. A. (1975). On the segmentation and analysis of continuous musical sound by digital computer. Master's thesis, Stanford University, Stanford, CA.

[Moorer, 1976] Moorer, J. A. (1976). The synthesis of complex audio spectra by means of discrete summation formulas. *J. Audio Eng. Soc*, 24(9):717–727.

[Moorer, 1978] Moorer, J. A. (1978). The use of the phase vocoder in computer music applications. *J. Audio Eng. Soc*, 26(1/2):42–45.

[Painter and Spanias, 2000] Painter, T. and Spanias, A. (2000). Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451 –515.

[Plumbley et al., 2010] Plumbley, M., Blumensath, T., Daudet, L., Gribonval, R., and Davies, M. (2010). Sparse representations in audio and music: From coding to source separation. *Proceedings of the IEEE*, 98(6):995 –1005.

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

[Rényi, 1961] Rényi, A. (1961). On Measures of Entropy and Information. In *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, pages 547–561, Berkeley, California.

[Roads et al., 1997] Roads, C., Pope, S. T., Picialli, A., and De Poli, G., editors (1997). *Musical Signal Processing*. Studies on New Music Research. Swets & Zeitlinger, Lisse, Netherlands.

[Röbel, 2003] Röbel, A. (2003). A new approach to transient processing in the phase vocoder. In *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, pages 344–349.

[Röbel et al., 2004] Röbel, A., Zivanovic, M., and Rodet, X. (2004). Signal decomposition by means of classification of spectral peaks. In *International Computer Music Conference (ICMC)*, pages 446–449, Miami, USA.

[Rudoy et al., 2010] Rudoy, D., Basu, P., and Wolfe, P. (2010). Superposition frames for adaptive time-frequency analysis and fast reconstruction. In *IEEE Trans. Sig. Proc.*, pages 2581–2596, Cambridge, Massachussets.

[Serra and Smith, 1990] Serra, X. J. and Smith, J. O. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24.

[Siegler et al., 1997] Siegler, M. A., Jain, U., Raj, B., and M. Stern, R. (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. of DARPA Speech Recognition workshop*, Harriman, New York.

[Søndergaard et al.,] Søndergaard, P. L., Torrésani, B., and Balazs, P. The Linear Time Frequency Analysis Toolbox. http://www.univie.ac.at/nuhag-php/ltfat/toolboxref.pdf.

[Sun, 2010] Sun, W. (2010). Asymptotic properties of gabor frame operators as sampling density tends to infinity. *J. Funct. Anal.*, 258(3):913–932.

[Tibshirani, 1994] Tibshirani, R. (1994). Regression shrinkage and selection via the lasso.

[Tošić and and Frossard, 2011] Tošić and, I. and Frossard, P. (2011). Dictionary learning. *Signal Processing Magazine, IEEE*, 28(2):27 –38.

[Tropp, 2004] Tropp, J. A. (2004). Greedy is good. *IEEE Trans. Info. Theory*, 50(10):2231–2242.

[Vinet and al, 2011] Vinet, H. and al (2011). Sample Orchestrator : gestion par le contenu d'échantillons sonores. *Traitement du signal*, -. accepted for publication.

[Walnut, 1992] Walnut, D. F. (1992). Continuity properties of the gabor frame operator. *Journal of Mathematical Analysis and Applications*, 165(2):479 – 504.

[Wells and Murphy, 2010] Wells, J. and Murphy, D. (2010). A comparative evaluation of techniques for single-frame discrimination of nonstationary sinusoids. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):498 –508.

[Wolfe et al., 2001] Wolfe, P. J., Godsill, S. J., and Dörfler, M. (2001). Multi-gabor dictionaries for audio time-frequency analysis. In *Proc. of the IEEE WASPAA*, pages 43–46, Mohonk, New York.

[Zibulski and Zeevi, 1997] Zibulski, M. and Zeevi, Y. Y. (1997). Analysis of multiwindow gabor-type schemes by frame methods. *Applied and Computational Harmonic Analysis*, 4(2):188–221.

[Zyczkowski, 2003] Zyczkowski, K. (2003). Rényi Extrapolation of Shannon Entropy. *Open Systems & Information Dynamics*, 10(3):297–310.