Automatic Adaptation of the Time-Frequency Resolution for Sound Analysis and Re-Synthesis

Marco Liuni, *Member, IEEE,* Axel Röbel, *Member, IEEE,* Ewa Matusiak, *Fellow, University of Vienna,* Marco Romito, *Associate Professor, University of Pisa,* and Xavier Rodet, *Member, IEEE*

***Abstract*—We present an algorithm for sound analysis and re-synthesis with local automatic adaptation of time-frequency resolution. The reconstruction formula we propose is highly efficient, and gives a good approximation of the original signal from analyses with different time-varying resolutions within complementary frequency bands: this is a typical case where perfect reconstruction cannot in general be achieved with fast algorithms, which provides an error to be minimized. We provide a theoretical upper bound for the reconstruction error of our method, and an example of automatic adaptive analysis and re-synthesis of a music sound.**

***Index Terms*—automatic adaptivity, Gabor frames, sparsity, spectral processing.**

I. Introduction

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.

Traditional sound analysis methods, based on single sets of atomic functions like Gabor window 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 other; 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.

Our main assumption is 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. In particular, the need to provide manual low level configuration is a major limitation for the use of advanced signal processing methods by large communities. The possibility to dispose of an automatic time-frequency resolution drastically limits the parameters to set, without affecting, and even ameliorating, the treatment quality: the result is an improvement of the user experience with high-quality sound processing techniques, like transposition and time-stretch.

The paper is organized as follows: Section II introduces some related works, and summarize the contribution of this article, while Section III gathers some definitions and symbols used. Then, our first and fundamental objective (see Section IV) is the formal definition of mathematical models whose interpretation leads to theoretical and algorithmic methods for adaptive analysis. The further objective is to make this adaptation automatic (see Section V): we choose the best local time-frequency resolution with the optimization of entropy-based *sparsity measures*. Rényi entropies (see [1], [2] and [3] for their properties) are considered, as they constitute a class of different sparsity measures because of their dependence on a parameter; a particular concept of sparsity is determined for each value of the parameter, whose choice can be refined depending on the specific application requirements, keeping the framework unaltered. Then, spectral 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 (see Section VI). Section VII presents applications and properties of the algorithms we have realized, and a perceptive test for the validation of time-adapted sound dilatations is discussed in Section VIII.

II. Related work

In time-frequency analysis, adaptivity is the possibility to conceive representations and operators whose characteristics can be modeled according to their input. In this work, we look for methods providing a local variation of the time-frequency resolution for sound analysis and re-synthesis. For instance, 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, as this is the context where this work is included; a further main point of view concerns the direct design of adaptive adapted time-frequency representation (see [4], [5] 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 for the adapted representation; and finally, the possibility to define a reconstruction method from the adapted analysis. The concept of adaptivity is closely related with the one of sparsity: an adaptive analysis must give a sparse representation of the signal, according to specific measures to be optimized, the optimal resolution being signal- and application-dependent. This is a highly prolific approach, largely exploited in various signal processing applications (see [6], [7]). The idea of gathering a sparsity measure from information measures, and Rényi entropies in particular, is detailed in [8]. In [9] 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 [10], provides Gabor frames with analysis techniques with multiple resolutions.The *nonstationary* Gabor frames (see [11], [12] 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 [13]: if the analysis respect certain conditions, they provide for 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 [14] 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 [15] 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 [16], is a promising effort to establish a unified mathematical model for all the various frameworks cited above.

A. 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 [9], exploiting new theoretical results that we have proven in [17], extending the ones in [8]: in particular, they concern the existence of Rényi entropy measures of spectrograms in the continuous case, and the convergence of discrete versions of these measures to their continuous one, when the sampling grid becomes infinitely dense; we adopt a particular normalization of the Rényi entropy (see Section V-A and [17]), which is appropriate for the comparison of the entropy of discrete finite time-frequency representations with different dimensions.

Concerning the reconstruction from adapted analyses, in Section IV we introduce a novel method allowing for an efficient (in the sense of Remark 4.2) FFT-based implementation; the resolution of the adapted analyses changes depending on time **and** frequency, and our method gives an approximation of the original signal whose reconstruction error is analyzed by means of tests, and of a theoretical upper bound.

III. Notation

We will be working throughout the paper with the Hilbert space of complex square integrable functions $L^2(\mathbb{R})$, with inner product

$$\langle f, g\rangle = \int_{\mathbb{R}} f(t)\overline{g(t)}\,\mathrm{d}t \quad \text{for all} \quad f, g \in L^2(\mathbb{R})$$

where $\overline{g}$ denotes the complex conjugate of g. The norm induced by this inner product is given by $\|f\|_2^2 = \langle f, f\rangle$; given $p > 0$, the L^p-norm (or pseudo-norm if $0 < p < 1$) of f is given by $\|f\|_p^p = \int_{\mathbb{R}} |f(t)|^p\,\mathrm{d}t$.

The support of a function $f \in L^2(\mathbb{R})$ is the closure of the set where f is non-zero, $\mathrm{supp}(f) = \overline{\{t \in \mathbb{R} : f(t) \neq 0\}}$. We say that f is ϵ-concentrated within an interval $T \subset \mathbb{R}$ if the following integral condition over the complementary of T holds for a sufficiently small ϵ,

$$\left(\int_{T^c} |f(t)|^2 \mathrm{d}t\right)^{\frac{1}{2}} \leq \epsilon \|f\|_2\,,$$

and we say that T is the *essential support* of f.

The Fourier transform of $f \in L^2(\mathbb{R})$ is defined as

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

and is also square integrable with $\|\widehat{f}\|_2 = \|f\|_2$. The consideration about the essential support still holds for $\widehat{f}$, in this case we say that $\widehat{f}$ is ϵ-bandlimited within a certain frequency interval.

A main tool in our derivations are Gabor frames, which we review in Section IV, (see [18] for a detailed survey). Two important operators that play a central role in Gabor theory, are the translation and modulation operators defined for $x, \omega \in \mathbb{R}$ as

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

respectively. The composition $M_\omega\, T_x f(t) = e^{2\pi i \omega t} f(t-x)$ is called a time-frequency shift operator and gives rise to the short-time Fourier transform. For a fixed window $g \in L^2(\mathbb{R})$, the short time Fourier transform (STFT) of $f \in L^2(\mathbb{R})$ with respect to g is defined as

$$\mathcal{V}_g f(x,\omega) := \langle f, M_\omega T_x g \rangle\,.$$

IV. Gabor theory

We resume in this section the basics of Gabor frames theory, and two generalizations of the stationary case in Subsection IV-A and IV-B, which are exploited by our algorithm.

A collection $\mathcal{G}(g,a,b) = \{g_{k,l}(t) = M_{bk}\, T_{al} g(t)\,;\, k,l \in \mathbb{Z}\}$ is a Gabor frame for $L^2(\mathbb{R})$ if there exist constants $0 < A \le B < \infty$ such that

$$A\|f\|^2 \le \sum_{k,l\in\mathbb{Z}} |\langle f, g_{k,l}\rangle|^2 \le B\|f\|^2$$

for all $f \in L^2(\mathbb{R})$. We will indicate such a frame as *stationary*, since the window used for time-frequency shifts does not change and the time-frequency shifts form a lattice $\Lambda = a\mathbb{Z} \times b\mathbb{Z}$ (see Figure 1). To every collection $\mathcal{G}(g,a,b)$ we associate the *analysis operator* $\mathrm{C_g}$ given by $\mathrm{C_g} f = \{\langle f, g_{k,l}\rangle\,;\, k,l \in \mathbb{Z}\}$, *synthesis operator* $\mathrm{D_g}$, where $\mathrm{D_g} c = \sum_{k,l\in\mathbb{Z}} c_{k,l} g_{k,l}$ and $c \in \ell^2$ and the *frame operator* S given by $\mathrm{S} f = \mathrm{D_g}\,\mathrm{C_g} f$.

If $\mathcal{G}(g,a,b)$ constitutes a frame for $L^2(\mathbb{R})$, then there exist a function $\widetilde{g} \in L^2(\mathbb{R})$, such that every function $f \in L^2(\mathbb{R})$ can be represented as

$$f = \sum_{k,l\in\mathbb{Z}} \langle f, g_{k,l}\rangle \widetilde{g}_{k,l} = \mathrm{D}_{\widetilde{\mathrm{g}}}\mathrm{C_g} f\,. \tag{1}$$

The Gabor system $\mathcal{G}(\widetilde{g},a,b)$ is the dual frame to $\mathcal{G}(g,a,b)$. Consequently, the window $\widetilde{g}$ is referred to as the dual of g. Generally, there is more than one dual window $\widetilde{g}$. The canonical dual is given by $\widetilde{g} = \mathrm{S}^{-1} g$. The *spectogram* $\mathrm{PS}_g f = |\mathcal{V}_g f|^2$ is the time-frequency representation associated with STFT.

We consider windows g that are members of so-called Feichtinger algebra, denoted by S_0 (see [19]). Such windows guarantee that the synthesis and analysis mappings are bounded and consequently result in stable reconstructions, and that the dual window is in S_0. Formally,

$$S_0 := \{f \in L^2(\mathbb{R})\,;\, \|\mathcal{V}_f\varphi\|_1 = \iint_{\mathbb{R}^2} |\mathcal{V}_f\varphi(x,\omega)|\,\mathrm{d}x\,\mathrm{d}\omega < \infty\}\,,$$

where $\varphi(t) = e^{-\pi t^2}$. The norm in S_0 is defined as $\|f\|_{S_0} := \|\mathcal{V}_f\varphi\|_1$. Examples of functions in S_0 are the Gaussian, B-splines of positive order, raised cosine, and any $L^1(\mathbb{R})$ function that is bandlimited or any $L^2(\mathbb{R})$ function that is compactly supported in time with Fourier transform in $L^1(\mathbb{R})$. Note, that the rectangular window is not a member of S_0 since its Fourier transform is not in $L^1(\mathbb{R})$.

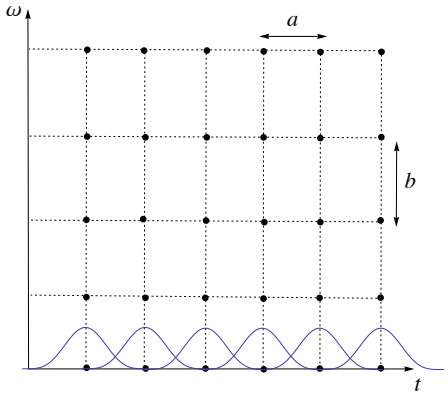

Fig. 1. *Time-frequency centers for a stationary Gabor frame* $\mathbf{G}(g,a,b)$ *with time and frequency steps a and b, respectively.*

We will be working with Gabor frames whose windows are supported on some compact interval. The support of g is denoted by $\mathrm{supp}(g)$. Such frames are often used in many standard applications.

Theorem 4.1: Consider $g \in L^\infty(\mathbb{R})$ with $\mathrm{supp}(g) \subset [-\frac{L}{2}, \frac{L}{2}]$; if $a \le L$, $b \le \frac{1}{L}$, then $\mathbf{G}(g,a,b)$ is a Gabor frame, and the frame operator S is the following multiplication operator,

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

The hypotheses of Theorem 4.1 define the *painless case*, where the dual window $\widetilde{g}$ is easy to calculate by means of a multiplication of the original one,

$$\widetilde{g}(t) = \mathrm{S}^{-1} g(t) = \frac{g(t)}{b^{-1}\sum_{l\in\mathbb{Z}} |g(t-al)|^2}\,. \tag{3}$$

Remark 4.2: Formula (1) shows that the atoms needed for the reconstruction of f are the time-frequency shifts of $\widetilde{g}$, according to the lattice Λ. From the identity (3) 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* or *efficient* those algorithms whose computational order is due to the FFT of short-length signals.

A. Nonstationary Gabor frames

A strategy to get an adaptive framework preserving a fast reconstruction method, in the sense of Remark 4.2, is given by nonstationary Gabor frames, (see [11], [12]): we 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 sampling set Λ, as depicted in Figure 2: for each time index l, a window g_l is chosen among the different set considered, which is centered at time a_l;

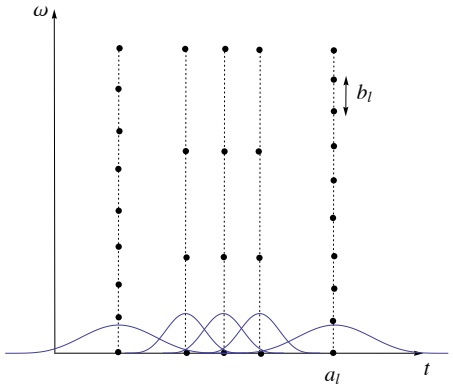

Fig. 2. *Time-frequency centers for a nonstationary Gabor frame in the time case, with variable time locations a_l and frequency steps b_l, depending on the time index l.*

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. A typical strategy for the windows choice, inspired by the wavelet approach, is to scale an original window g as follows,

$$g_l(t) = \frac{1}{\sqrt{i}}\, g\left(\frac{t}{i}\right) , \tag{4}$$

where i depends on the index l and varies in a finite set I of positive scaling factors. Referring to the time case, a nonstationary Gabor frame is thus given by the atoms

$$g_{k,l}(t) = g_l(t)\, \mathrm{e}^{2\pi i k b_l t} \ , (l,k) \in \mathbb{Z}^2 , \tag{5}$$

where b_l is the frequency step associated to the window g_l. For nonstationary Gabor frames there exist a painless case for the calculation of the dual, whose conditions are detailed in the following theorem ([11], Theorem 1).

Theorem 4.3: 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,

$$\mathrm{S}f(t) = \left(\sum_{l\in\mathbb{Z}} \frac{1}{b_l} |g_l(t)|^2\right) f(t) . \tag{6}$$

As a consequence, if $\sum_{l\in\mathbb{Z}} \frac{1}{b_l}|g_l(t)|^2 \simeq 1$, then the set (5) is a frame whose dual frame is given by

$$\widetilde{g}_{k,l}(t) = \frac{g_l(t)\, \mathrm{e}^{2\pi i k b_l t}}{\sum_{l\in\mathbb{Z}} \frac{1}{b_l}|g_l(t)|^2} . \tag{7}$$

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

$$f = \mathrm{D}_{\widetilde{\mathrm{g}}_\mathrm{l}}(\mathrm{C}_{\mathrm{g}_\mathrm{l}} f) , \tag{8}$$

appropriately considering the window and lattice variations at each time location. The expression of the dual frame shows that this formula can be implemented with a fast FFT-based algorithm.

B. Gabor multipliers and weighted frames

Spectral processing techniques are based on analysis manipulations, which determine the desired effect in the resynthesized signal. Gabor multipliers (see [20] 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 4.4: 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

$$\mathbf{G}_{\mathbf{m},\Lambda}^{g^1,g^2}(f) = \mathrm{D}_{\mathrm{g}^2}(\mathbf{m} \cdot \mathrm{C}_{\mathrm{g}^1} f) , \tag{9}$$

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.

The definition of spectral manipulations can be also approached from the point of view of the decomposing atoms; in [21], the concept of weighted frame is introduced.

Definition 4.5: Consider a set of atoms $\{g_{k,l} = M_{bk}T_{al}g\}_{k,l\in\mathbb{Z}}$, for some $a, b > 0$, in $L^2(\mathbb{R})$, and a sequence $\{w_{k,l}\}_{k,l\in\mathbb{Z}}$ of complex numbers. The set $\{w_{k,l}g_{k,l}\}_{k,l\in\mathbb{Z}}$ is a *weighted Gabor frame* for $L^2(\mathbb{R})$ if there exist two positive non zero constants A and B such that for all $f \in L^2(\mathbb{R})$,

$$A\|f\|^2 \leq \sum_{k,l\in\mathbb{Z}} |\langle f, w_{k,l}g_{k,l}\rangle|^2 \leq B\|f\|^2 . \tag{10}$$

We indicate such a frame with $\mathbf{G}^w(g,a,b)$.

Indicating with C_g^w the analysis operator associated to $\mathbf{G}^w(g,a,b)$ and considering $\mathbf{m} = (w_{k,l})_{k,l\in\mathbb{Z}}$, we can write

$$\mathbf{G}_{\mathbf{m},\Lambda}^{g,\widetilde{g}} f = \mathrm{D}_{\widetilde{\mathrm{g}}}(\mathrm{C}_\mathrm{g}^w f) , \tag{11}$$

showing the relation between a Gabor multiplier and a weighted Gabor frame.

V. Algorithm

We outline here the fundamental steps of the proposed algorithm, detailing each of them later in the section. The strategy we adopt to obtain a resolution varying in time and frequency is to first select a fixed number of frequency bands; then, the signal is processed with filters whose impulse responses match the selected bands. The different filtered versions are analyzed with nonstationary Gabor frames in the time case: the resolutions of these frames are automatically time-adapted, optimizing the entropy of several fixed-resolution analyses performed with stationary frames (see Subsection V-A). A unique analysis of the original signal is finally deduced by the different nonstationary ones of its filtered versions: the resolution of the final analysis changes depending on the frequency band, and on the time location within each frequency band.

Disposing of analyses with varying resolution, 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; but still, having analyses with varying resolution requires 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, we discuss the application of a common spectral processing technique, called *time-stretch* (see Section VIII), dealing with analyses with time-adapted resolution. In general, the analysis and re-synthesis algorithms we develop are designed to allow for extensions of existing processing methods, such the ones available in the phase vocoder approach.

For the reconstruction task, two cases can be considered: if the analysis window varies depending on time or frequency individually, or if it depends on both time and frequency. For the first case, nonstationary Gabor frames provide fast algorithms for perfect reconstruction within the painless conditions (see Subsection IV-A).

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. Our approach, denoted as *filter bank*, is focused on a fast algorithm detailed in the following subsections; Subsection V-C, in particular, introduces a variation of this approach, which is implemented in our framework: instead of filters, weight functions are applied to the STFT coefficients of the original signal. This choice is motivated by the growing popularity of spectral techniques in the development of user-oriented sound processing software: thanks to the significant power offered even by standard computers, which lowers the latency introduced by FFT computations, spectral weighting provides an intuitive and efficient tool for the visual-driven design of a large variety of filters. This approach is investigated with several experiments in Section VI and VII.

A. Best window selection and time-adapted analyses

We define here a procedure for the local adaptation of the time-frequency resolution of the spectrogram, according to an entropy-based sparsity criterium. The approach we adopt (see [8] 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.

Given a finite set of index I, we consider different windows g_i of a same type, with varying time-support. We know that each g_i, together with an appropriate lattice Λ_i defined by a time step a_i and a frequency step b_i, forms a frame for our space of signals. Associating the analysis coefficients to the points of the lattice, we can represent the discrete spectrogram $\mathrm{PS}_i f$ by means of the lattice Λ_i. To evaluate the concentration of the different spectrograms $\mathrm{PS}_i f$, we use a normalization of the Rényi entropies; we give the definition in the case of discrete finite spectrograms, which is the one we use when dealing with sounds in a digital format. Given an *entropy order* $\alpha \geq 0,\ \alpha \neq 1$ and m_i, n_i the numbers of rows an columns in the matrix $\mathrm{PS}_i f$, the normalized Rényi entropy is defined as follows:

$$\mathrm{H}_\alpha^R[\mathrm{PS}_i f] = \frac{1}{1-\alpha} \log \sum_{(l,k)} \Big(\frac{\mathrm{PS}_i f(l,k)}{\sum_{(l',k')} \mathrm{PS}_i f(l',k')} \Big)^\alpha + \log \frac{a_i b_i}{m_i n_i} \ , \tag{12}$$

where l and k are the row and column indexes of the matrix $\mathrm{PS}_i f$ which belong to $R \subseteq \Lambda_i$; as detailed in [17], this class of measures is appropriate for the entropy comparison of discrete analyses with different lattices, and we get a better interpretation of the comparison when the analyses are realized with the same time-frequency oversampling: that is, if the product $a_i b_i$ is constant for every index $i \in I$.

The local evaluation (12) takes into account a certain subset R of the analysis coefficients, depending on the envisaged localization: when f is a sound of finite duration, its essential time-frequency support can be inscribed in a rectangle $\overline{R} = \mathrm{supp}(f) \times [-\Omega/2, \Omega/2], \Omega \in \mathbb{R}^+$, whose horizontal side is the support of f, and whose vertical side is the essential support of $\widehat{f}$. The localization we are interested in, is realized by choosing a rectangle $R \subseteq \overline{R}$, and a set of time-frequency shifts of R which cover $\overline{R}$; the area within a particular shift of 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 3), as a solution of the following optimization problem,

$$\min_{i \in I} \mathrm{H}_\alpha^R[\mathrm{PS}_i f] \ . \tag{13}$$

At each step of our algorithm, the rectangle R is shifted in the time-frequency plane with a certain overlap with the

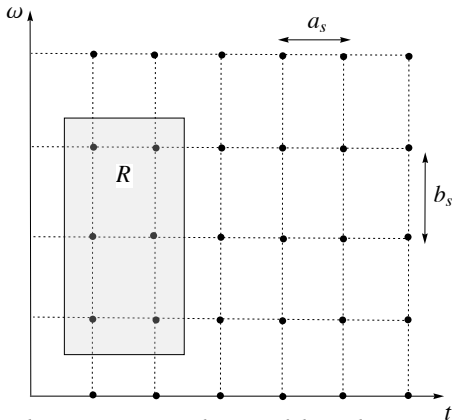

Fig. 3. *Graphic representation of a step of the evaluation procedure: for a given spectrogram* $\mathrm{PS}_i f$ *and a time-frequency shift of the rectangle* R*, the Rényi entropy of the coefficients within* R *is calculated (see Section V-A).*

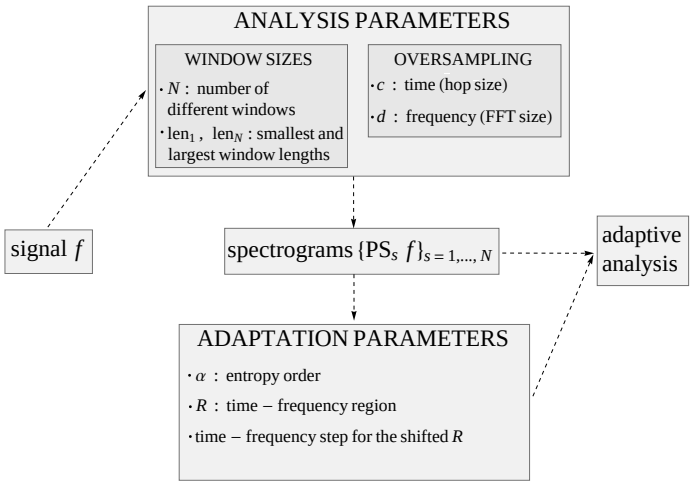

Fig. 4. *Graphic representation of the main steps performed by the algorithm for the automatic local adaptation of the spectrogram window size: the two main blocks are depicted by means of their parameters, while the mid line shows input and outputs (see Section V-A).*

previous position. Within the area of the shifted R, the best coefficients are defined as the ones which belong to the optimal spectrogram $\mathrm{PS}_i f$ in the sense of problem (13); 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. In the case of time-adapted analyses, the entropy evaluation is recursively performed on segments $[t', t'']$ of the signal, taking into account the whole frequency spectrum: the horizontal side of the rectangle R is $[t', t'']$, while the vertical one is the whole frequency dimension. The window g_i associated to the sparsest local analysis is assigned as best window to all the points $(t, \omega) \in \Lambda_i \cap R$. The global time-adapted analysis of the signal is finally realized considering the best windows selected at each time location of the obtained composite lattice; as $\mathbf{G}(g_i, a_i, b_i)$ are Gabor frames for all $i \in I$, by appropriately choosing the time shift factor the rectangle R, the selected windows and the composite lattice form a nonstationary Gabor frame in the time case.

The continuous extension of measure (12), where sums are replaced by integrals and the second term in the right part vanish, is given by

$$\mathrm{H}_\alpha(\mathrm{PS}f) = \frac{2\alpha}{1-\alpha} \log_2 \frac{\|\mathcal{V}_g f\|_{2\alpha}}{\|\mathcal{V}_g f\|_2} \tag{14}$$

The main advantage of using Rényi entropies emerges from the comparison of this form with the Kurtosis-like measure $\|\mathcal{V}_g f\|_4^4 / \|\mathcal{V}_g f\|_2^2$ used in [4]: the norm applied to the spectral coefficients at the numerator introduces a biasing depending on the α parameter. Different values of α determine different concepts of sparsity (see [22]), as the influence of weak spectral coefficients (high partials, as well as noise) on the entropy measure changes accordingly: if $\alpha >> 1$, only the main spectral peaks are taken into account, while when $\alpha = 0$, there is no difference between large and small peaks. With $\alpha = 0.3$, the measure has an analogy with the power law used to map loudness levels in phons to perceived loudness in sones; this value provides an appropriate measure for wide range of vocal and music sounds, and has been used for all the examples and tests proposed in Section VI and VII. The α parameter, as well as the other adaptation parameters indicated in Figure 4, are not meant to be changed by the user: the designers of a specific application have to set them, to tune the required adaptivity; one could even imagine to include an expert user mode, where certain resolutions are privileged among others, but a refined control of these parameters requires a deep knowledge of the entropy measures introduced.

B. Filter bank

We define here a novel approximation method based on analyses with resolution changing in time and frequency, indicating theoretical bounds for the reconstruction error. We extend the results in [23] to the case of filtered signals, obtaining the approach we indicate as *filter bank* in the case of stationary Gabor frames (see Figure 5).

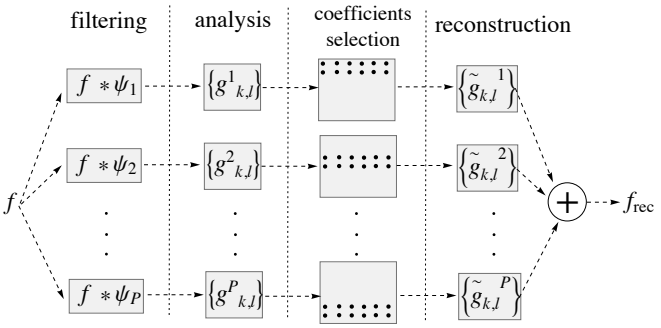

Fig. 5. *Block diagram detailing the steps of the filter bank approach (see Subsection V-B): the signal is first filtered with a bank of P band-pass filters; the filtered signals are analyzed with P nonstationary Gabor frames; the coefficients in the analyses are selected depending on the corresponding 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.*

In Subsection V-C, we define 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 VI).

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 [23].

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 ,

$$\mathcal{V}_p f(t,\omega) = \int_{\mathbb{R}} f(\tau)\overline{g^p(\tau - t)}\, \mathrm{e}^{-2\pi i \omega \tau}\, \mathrm{d}\tau \ , \qquad (15)$$

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

$$c_{k,l}^p = \langle f_p, M_{b_p k} T_{a_p l} g^p \rangle \ , \ (k,l) \in \mathbb{Z}^2 \ ; \qquad (16)$$

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 [23] 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 f_p°, given by

$$f_p^\circ = \sum_{k \in K^p} \sum_{l \in L^p} c_{k,l}^p M_{b_p k} T_{a_p l} \widetilde{g^p} \ , \qquad (17)$$

verifies the following inequality,

$$\left\| f_p - f_p^\circ \right\|_2 \leq C_p(\epsilon_p + \epsilon) \|f_p\|_2 \ , \qquad (18)$$

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 $l a_p$ for which support of f_p overlaps with support of g^p shifted by $l a_p$; the set K^p contains the frequency positions $k b_p$ for which essential support of $\widehat{f_p}$ overlaps with essential support of $\widehat{g^p}$ shifted by $k b_p$. Then the cardinality of L_p equals

$$|L^p| = 2 \left\lceil \frac{\beta + T_p + W_p}{2 a_p} \right\rceil - 1 \ ; \qquad (19)$$

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

$$|K^p| = \left\lceil \frac{\Omega_p^2 - \Omega_p^1 + \alpha_p}{b_p} \right\rceil \ . \qquad (20)$$

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 (18): for every $\epsilon > 0$, with the appropriate sets and constants we have

$$\begin{aligned} \left\| f - \sum_p f_p^\circ \right\|_2 &\leq \left\| f - \sum_p f_p \right\|_2 + \left\| \sum_p f - \sum_p f_p^\circ \right\|_2 \\ &\leq \epsilon_\Omega \|f\|_2 + C_P(\epsilon_P + \epsilon) \sum_p \|f_p\|_2 \, . \end{aligned} \qquad (21)$$

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

$$\left\| f - \sum_p f_p^\circ \right\|_2 \leq (\epsilon_\Omega + C_\psi C_P(\epsilon_P + \epsilon)) \|f\|_2 \ . \qquad (22)$$

Remark 5.1: The choice of the ψ_p functions has an influence on the error we obtain: assuming to work with S_0 windows (see [18]), 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, for chosen ϵ, the number of coefficients used in the expansion (17) depends on the windows g^p; the better concentration of g^p in frequency, the less frequency coefficients are required to achieve ϵ accuracy. 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.

C. Filter bank approach with Gabor multipliers

Spectral processing techniques often avoid manipulations in the signal domain, privileging modifications of the analysis coefficients, followed by the re-synthesis. We look for an estimate like the one in equation (21) when working with Gabor multipliers instead of filters: considering the scheme in FIgure 5, the variation we develop consists in inverting the analysis and the filtering stage; then, filtering is performed now by means of weight functions applied to the spectral coefficients. In particular, we want to replace each filter ψ_p with a Gabor multiplier $\mathbf{G}_{\mathbf{m}_p, \Lambda_p}^{g^p, \widetilde{g^p}}$, 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,

$$W_p f(t,\omega) = \mathcal{V}_p f(t,\omega) \mathbf{m}_p(t,\omega). \qquad (23)$$

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, $d_{k,l}^p = W_p(a_p l, b_p k)$; indeed, if we write $g(\tau - t) = g_t(\tau)$, then

$$\begin{aligned} \mathcal{V}_p f_p(t,\omega) &= ((\widehat{f} \cdot \widehat{\psi_p}) * \widehat{g_t^p})(\omega) \ , \\ W_p f(t,\omega) &= (\widehat{f} * \widehat{g_t^p})(\omega) \cdot \widehat{\psi_p}(\omega) \ ; \end{aligned} \qquad (24)$$

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 [24], [25]): given an underspread operator H, its best approximation by a Gabor multiplier $\mathbf{G}_{\mathbf{m},\Lambda}^{g_1,g_2}$ 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: 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^*,

$$|c_{k,l}^p - d_{k,l}^p| \leq \frac{\epsilon_p^*}{PKL} \ , \tag{25}$$

where KL is the number of coefficients in the expansion (17); here, we assume this inequality to hold. Using the coefficients $d_{k,l}^p$ in the same expansion, we obtain

$$f_p^* = \sum_{k\in K^p}\sum_{l\in L^p} d_{k,l}^p M_{b_p k} T_{a_p l}\widetilde{g^p} \ , \tag{26}$$

and

$$\|f_p^\circ - f_p^*\|_2 \leq \frac{\epsilon_p^*}{P}\cdot\|\widetilde{g^p}\|_2 \ . \tag{27}$$

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

$$\left\|f_p - f_p^*\right\|_2 \leq C^p(\epsilon_p+\epsilon)\|f_p\|_2 + \frac{\epsilon_p^*}{P}\|\widetilde{g^p}\|_2 \ . \tag{28}$$

Writing $\epsilon_P^* = \max_p \epsilon_p^*$ and $\|\widetilde{g^P}\|_2 = \max_p \|\widetilde{g^p}\|_2$, we can rewrite the estimate (22) as follows,

$$\left\|f - \sum_p f_p^*\right\|_2 \leq C_\psi C_P(\epsilon_P+\epsilon)\|f\|_2 + \epsilon_P^*\|\widetilde{g^P}\|_2 \ . \tag{29}$$

As we are working with Gabor frames in the painless case, we can further precise 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 VI, we provide examples of the reconstruction error obtained for given choice of the above functions.

As we have shown in [17], the estimate obtained in the expansion (17) can be extended to the nonstationary case; a similar extension for the estimate in (29) will be treated in a separated contribution.

VI. Re-synthesis from adaptive analyses

We focus here on the approximation method introduced, in the Gabor multipliers case, considering two frequency bands, so $P=2$: given a signal f and a reconstruction f_{rec} obtained with that method, we measure its accuracy by means of the maximum of the absolute value of the error $er\_peak = \|f - f_{rec}\|_\infty$, and the *RMS* (Root Mean Square) of the error, that is

$$er\_rms = \sqrt{\frac{\sum_{n=1}^L (f[n]-f_{rec}[n])^2}{\sum_{n=1}^L f[n]^2}} \ , \tag{30}$$

where L is the signal length.

We first detail our approach in terms of stationary Gabor frames, which is also the case which the estimates in Subsection V-C refer to. Then, we will extend the methods to the nonstationary case, which is used in our framework.

A. A stationary Gabor frame for each frequency band

Using the notation introduced in Section IV, and Subsection IV-B 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, possibly considering an overlap. The reconstruction formula is thus given by

$$f_{rec} = \mathrm{D}_{\widetilde{\mathrm{g}}}(\mathrm{C}_{\mathrm{g}}^{w_1} f) + \mathrm{D}_{\widetilde{\mathrm{h}}}(\mathrm{C}_{\mathrm{h}}^{w_2} f) \ . \tag{31}$$

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

To investigate the effects of the weighting technique, we consider a basic signal whose energy is concentrated at the frequency point where the two weights vanish: f is a sinusoid with sinusoidal frequency modulation, sampled at 44.1kHz. Then, we measure the reconstruction error obtained with binary weights, as well as the reduction obtained allowing the weights for an overlap. The sinusoid frequency starts at 350Hz, and the modulation varies between 130Hz and 570Hz with a period of half a second.

For the weighting functions, the following frequencies are given: $\Omega_{cut} = 350$Hz, $\Omega_1 = 200$Hz and $\Omega_2 = 500$Hz. We first consider two binary masks, whose coefficients vanish, respectively, at the frequencies above and below Ω_{cut}; then, two masks w_1 and w_2 with a linear crossfade are taken: having $Ny = 22.05$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 $w_2 = 1-w_1$. The windows g and h are Hanning windows of size 512 and 4096 samples. The aim is to show that the reconstruction error can be reduced, appropriately choosing the overlap of the two masks, according to the signal spectral energy.

Table I shows the errors obtained: with the overlap 200-500Hz, which does not include the whole modulation range, the error reduction obtained with the linear crossfade is limited. 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 (26); as too few coefficients are considered for the reconstruction of the two individual bands,

then the error on the global reconstruction is still considerable. The last line of Table I shows, as expected, that increasing the

TABLE I
Reconstruction error when f is a sinusoid with sinusoidal modulation: the masks are indicated on the left, together with their frequency significant values (see Subsection VI-A).

Weight method	Parameters	er_peak	er_rms
Binary mask	$freq_{cut} = 350$Hz	0.5102	0.0967
Linear cross	$freq_1 = 200$Hz $freq_2 = 500$Hz	0.1856	0.0725
Linear cross	$freq_1 = 50$Hz $freq_2 = 650$Hz	0.0576	0.0262

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 we have approximations with er_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-frequency point.

Figure 6 shows the composed spectrogram obtained with the binary masks, and the consequent reconstruction error. The same, in Figure 7, for weights with linear crossfade 50-650Hz. We thus see that the spectral energy of the error with overlapping weights is lower, and more uniformly distributed.

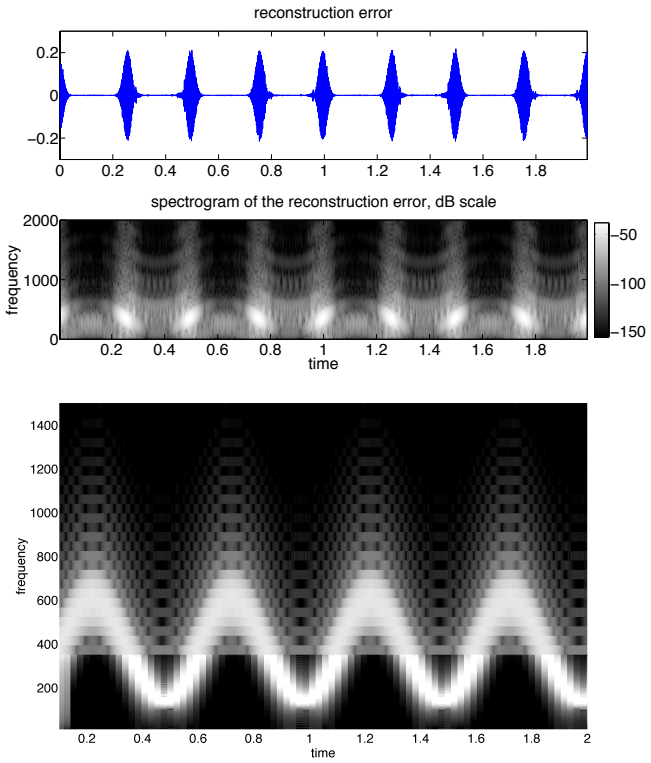

Fig. 6. *Composed spectrogram of a sinusoid with sinusoidal frequency modulation: the signal is analyzed with two different windows, the spectrogram coefficients are weighted with two binary masks and then summed together (see Subsection VI-A). On top, the reconstruction error obtained, and its spectrogram. Times are indicated in seconds, frequencies in Hertz.*

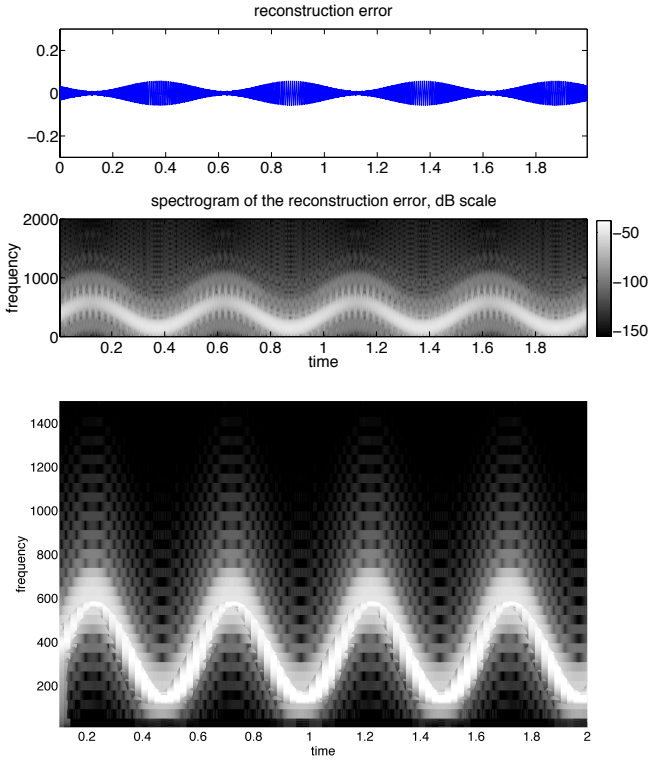

Fig. 7. *Composed spectrogram of a sinusoid with sinusoidal frequency modulation: the signal is analyzed with two different windows, the spectrogram coefficients are weighted with the w_p masks, with overlap 50-650Hz, and then summed together (see Subsection VI-A). On top, the reconstruction error obtained, and its spectrogram. Times are indicated in seconds, frequencies in Hertz.*

B. A nonstationary Gabor frame for each frequency band

The test we have shown has been obtained with two stationary Gabor frames, each one associated to a frequency band. In our framework, we extend this methods to nonstationary Gabor frames. With the different scalings g_i of a same window function, and appropriate lattices Λ_i, we realize the analyses $\mathcal{V}_{g_i} f$ and their weighted versions $\mathcal{V}_{g_i} f(t,\omega) w_p(\omega)$. These weighted analyses are used for the reconstruction, after the automatic time-adaptation procedure detailed in Subsection V-A: at the end of the automatic selection of the window, 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,

$$f_{rec} = \mathrm{D}_1(\mathrm{C}_1^{w_1} f) + \mathrm{D}_2(\mathrm{C}_2^{w_2} f) \ . \tag{32}$$

VII. Time-frequency adaptation and music

When a music signal presents heterogeneous spectral components, the choice of a fixed resolution, or even of a time-dependent resolution, may not be sufficient to resolve them. As an example, we consider a sound sample 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: this 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 in V-A, would not be appropriate within all the frequency regions.

The right part of Figure 8 shows the adapted spectrogram we obtain analyzing this sound with the following setup:

- $N = 8$ Hanning windows, whose length varies between $len_1 = 1024$ and $len_N = 4096$ points, which correspond to about 23 milliseconds and 93 milliseconds, as the sampling rate SR is 44.1kHz;
- $c = 0.15$ 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 139 milliseconds;
- 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 116 milliseconds.

Then, all the spectrograms are weighted with a binary mask setting to 0 the coefficients above 1kHz before the entropy evaluation, and the consequent window selection. The chosen mask rises a window choice adapted to the frequency area where the first harmonics of the two instruments are predominant. Nevertheless, 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 range. The complementary analysis, where the window selection is adapted to the coefficients below 1kHz, is shown in the left part of Figure 8.

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 as shown in the left spectrogram of Figure 8, 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 the right spectrogram. 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 9. Table II shows the reconstruction error obtained on this music signal, with the analysis-weight method detailed in Section VI. Here, 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. A further ongoing research project is focused on a prior target to achieve, that is a strategy to reduce the error by varying the weighting functions, once chosen their overlap: this would lead to an effective control of the error by an intuitive setting, closely related to filter design.

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

Weight functions	Parameters	er_peak	er_rms
Binary mask	$freq_{cut} =$ 1kHz	$4.7{\cdot}10^{-3}$	$4.4{\cdot}10^{-3}$
Linear cross	$freq_1 =$ 750Hz $freq_2 =$ 1.25kHz	$3.4{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$
Linear cross	$freq_1 =$ 500Hz $freq_2 =$ 1.5kHz	$3.7{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$

The reconstruction error and its spectrogram are shown in Figure 10: comparing this figure with the ones of the adapted analyses for the two different bands (Figures 8 left and right), we see that the error energy is concentrated at the time locations where the window choice differs within the two bands, and within a frequency range determined by the overlap of the two masks. As the aim of these representations is to ameliorate sound processing algorithms, the perceived quality of the reconstruction is determinant, beside the objective error measure: some preliminary tests with expert listeners show that, even when binary masks are used, the re-synthesized signal is perceived as identical to the original one. This allow to argue that, even in the case of multiple bands, the error should be perceptually masked by the relevant components. Further investigations have to characterize the error, in particular when applying processing techniques that may enhance the error.

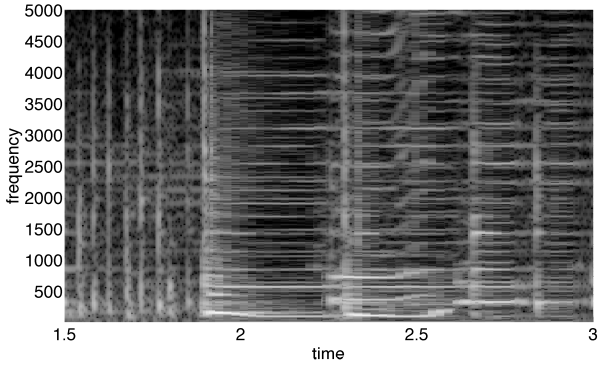

Fig. 9. *Adaptive analysis of a sound sample with tabla and sitar (see Section VII); the frequency range is limited to enhance readability: the resolution is adapted in time and in two frequency bands, above and below 1kHz. Times are indicated in seconds, frequencies in Hertz.*

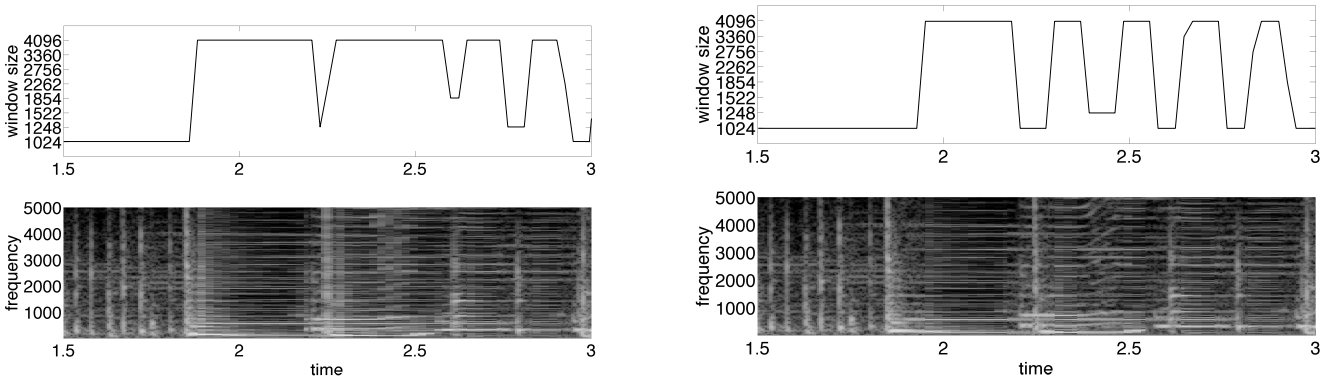

Fig. 8. *Adaptive weighted analysis of a sound sample with tabla and sitar (see Section VII): the binary masks for the adaptation select alternatively the coefficients below 1kHz (on the left) or above 1kHz (on the right); the frequency range is limited to enhance readability: on top, the best window 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). Times are indicated in seconds, frequencies in Hertz, window sizes in number of points.*

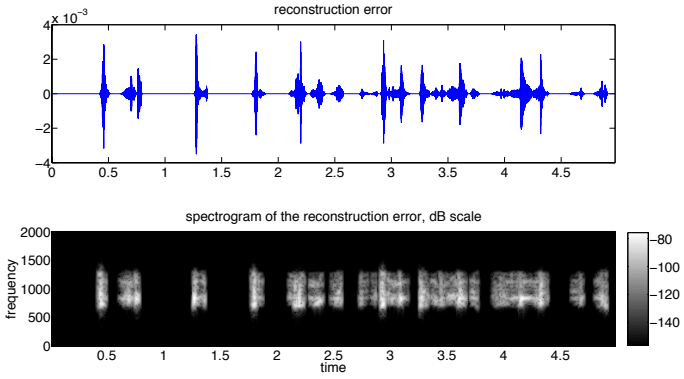

Fig. 10. *Spectrogram of the reconstruction error given by the analysis-weight approach, on a sound sample with tabla and sitar (see Section VII); the frequency range is limited around the overlap of the weighting masks, from 750Hz to 1.25kHz. Times are indicated in seconds, frequencies in Hertz.*

VIII. Perceptive Evaluation of Adaptive Time-Stretching

We have realized a perceptive test[1] to evaluate the quality of spectral processing techniques based on analyses whose resolution is adapted only in time (see Section V-A): the listener is asked to compare three different time-stretches (the duration of a sound is increased without changing its pitch) of given sound samples. The three time-stretches are realized with the extended phase vocoder SuperVP[2]: two of them use constant windows, of lengths $len_1 = 1024$ and $len_2 = 4096$ samples, respectively (the sampling rate is 44.1kHz, all sounds are in standard cd format); the third adopts a variable window, computed with the algorithm detailed in Subsection V-A, with 8 windows whose length vary between len_1 and len_2. All the methods use the advanced option for transients preservation available in SuperVP.

The test has been performed on 43 listeners wearing headphones and declaring to be familiar with comparing sound transformation examples: they are asked to evaluate how natural the dilatation is, compared to the original sound, on a scale going from 1 (*inacceptable*) to 5 (*perfectly natural*).

[1] see http://recherche.ircam.fr/equipes/analyse-synthese/liuni/form_mp3.php
[2] see http://anasynth.ircam.fr/home/english/software/supervp

Figure 11 shows the results of the test: the left graph compares the mean scores of the three methods, where the adapted window surmounts the two constants ones, as the chosen sound samples are complex and no constant window is appropriate for all of them. In the right graph, the best score of the constant window is taken for each sound, and the mean of these best scores is calculated: the adapted window slightly surpasses the best constant one, which proves that state-of-the-art time-stretching methods with automatic adaptation of the window have comparable performances with the ones obtained through an accurate choice between two fixed windows.

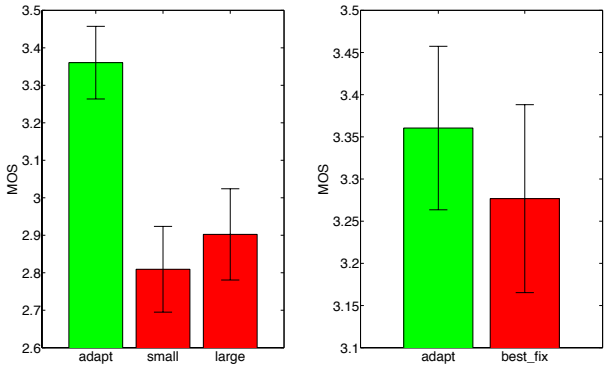

Fig. 11. *Results of the perceptive test detailed in Section VIII: on the left, mean opinion scores (MOS) of the three different methods; on the right, mean score of the adaptive method versus the mean of the best fixed resolution ones.*

IX. Conclusion

We introduced an algorithm for sound analysis and re-synthesis with local automatic adaptation of time-frequency resolution: our method is intended to provide a signal representation with optimal local time-frequency information, with a high quality of the analysis/re-synthesis process. The computational order of the algorithm is determined by the short-length FFT performed: this high efficiency is obtained at the price of an error in the reconstruction, whose theoretical bound is provided.

At present, there are no common sound processing techniques dealing with time-frequency adapted analyses like the

ones we introduce: therefore, when varying the resolution both depending on time and frequency, 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 be done iteratively on the different frequency bands, according to the weighted analyses, the fundamental task being to conceive appropriate strategies to treat the overlapping zones, depending on the specific sound treatment. These aspects, together with the objective and perceptual investigation of the reconstruction error, are the core of the ongoing research of the authors on this topic.

Together with sound transformation, the optimal local time-frequency resolution guarantees a solid ground to develop adaptive high-quality applications: sound object localization and separation with adapted time-frequency precision, as well as information retrieval with optimal local resolution, among several others.

ACKNOWLEDGMENT

The authors would like to thank Monika Dörfler and Peter Balazs for the precious collaboration.

REFERENCES

[1] A. Rényi, "On Measures of Entropy and Information," in *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, Berkeley, California, June 20-30, 1961, pp. 547–561.

[2] C. Beck and F. Schlögl, Eds., *Thermodynamics of chaotic systems*. Cambridge, Massachusetts, USA: Cambridge University Press, 1993.

[3] K. Zyczkowski, "Rényi Extrapolation of Shannon Entropy," *Open Systems & Information Dynamics*, vol. 10, no. 3, pp. 297–310, Sep. 2003.

[4] D. Jones and R. Baraniuk, "A simple scheme for adapting time-frequency representations," *IEEE Trans. Signal Processing*, vol. 42, no. 12, pp. 3530–3535, Dec. 1994.

[5] ——, "An adaptive optimal-kernel time-frequency representation," *Signal Processing, IEEE Transactions on*, vol. 43, no. 10, pp. 2361 –2371, oct 1995.

[6] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. Davies, "Sparse representations in audio and music: From coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995 –1005, june 2010.

[7] I. Tošić and and P. Frossard, "Dictionary learning," *Signal Processing Magazine, IEEE*, vol. 28, no. 2, pp. 27 –38, march 2011.

[8] R. Baraniuk, P. Flandrin, A. Janssen, and O. Michel, "Measuring Time-Frequency Information Content Using the Rényi Entropies," *IEEE Trans. Info. Theory*, vol. 47, no. 4, pp. 1391–1409, May 2001.

[9] F. Jaillet and B. Torrésani, "Time-frequency jigsaw puzzle: adaptive and multilayered Gabor expansions," *International Journal for Wavelets and Multiresolution Information Processing*, vol. 1, no. 5, pp. 1–23, 2007.

[10] M. Dörfler, "Gabor Analysis for a Class of Signals called Music," Ph.D. dissertation, Institut für Mathematik der Universität Wien, 2002.

[11] P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, and G. Velasco, "Theory, implementation and applications of nonstationary gabor frames," *Journal of Computational and Applied Mathematics*, vol. 236, no. 6, pp. 1481 – 1496, 2011.

[12] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The Linear Time Frequency Analysis Toolbox," http://www.univie.ac.at/nuhag-php/ltfat/toolboxref.pdf.

[13] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Phys.*, vol. 27, pp. 1271–1283, May 1986.

[14] D. Rudoy, P. Basu, and P. Wolfe, "Superposition frames for adaptive time-frequency analysis and fast reconstruction," in *IEEE Trans. Sig. Proc.*, Cambridge, Massachussets, May 2010, pp. 2581–2596.

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

[16] M. Dörfler, "Quilted Gabor frames - A new concept for adaptive time-frequency representation," *Advances in Applied Mathematics*, vol. 47, no. 4, pp. 668 – 687, 2011.

[17] M. Liuni, "Automatic Adaptation of Sound Analysis and Synthesis," Ph.D. dissertation, Universitá di Firenze - IRCAM, Université Paris VI, 2012. [Online]. Available: http://articles.ircam.fr/textes/Liuni12a/index.pdf

[18] K. Gröchenig, Ed., *Foundations of Time-Frequency Analysis*. Boston, Massachusetts, USA: Birkhäuser, 2001.

[19] H. Feichtinger and T. Strohmer, Eds., *Gabor analysis and algorithms*, ser. Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Boston Inc., 1998.

[20] H. G. Feichtinger and K. Nowak, "A first survey of gabor multipliers," in *Advances in Gabor Analysis*. Birkhäuser, 2002, pp. 99–128.

[21] P. Balazs, J.-P. Antoine, and A. Griboś, "Weighted and controlled frames: mutual relationships and first numerical properties," *Int. J. Wav, Mult. Info. Proc.*, vol. 8, no. 1, pp. 109–132, 2010.

[22] M. Liuni, P. Balazs, and A. Röbel, "Sound analysis and synthesis adaptive in time and two frequency bands," in *Proc. of DAFx11*, Paris, France, September 19-23, 2011.

[23] E. Matusiak and Y. C. Eldar, "Sub-Nyquist sampling of short pulses," *IEEE Trans. Sig. Proc.*, vol. 60, no. 3, pp. 1134–1148, Mar. 2012.

[24] M. Dörfler and B. Torrésani, "Spreading function representation of operators and Gabor multiplier approximation," in *Proceedings of SAMPTA07*. NuHAG;MOHAWI, June 2007.

[25] M. Dörfler and B. Torresani, "Representation of operators in the time-frequency domain and generalized Gabor multipliers," *J. Fourier Anal. Appl.*, vol. 16, no. 2, pp. 261–293, 2010.

Marco Liuni received the Ph.D. degree in mathematics from Florence University and the Ph.D. degree in signal processing from UPMC Paris 6 University in 2012. He is currently postdoctoral fellow in the Analysis/Synthesis Team at IRCAM. His interests lie in sound processing, source separation and computer music.

Axel Röbel received the Diploma in electrical engineering from Hanover University in 1990 and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin in 1993. In 1994 he joined the German National Research Center for Information Technology (GMD-First) in Berlin where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996 he became assistant professor for digital signal processing in the communication science department of the Technical University of Berlin. In 2000 he was visiting researcher at CCRMA Standford University, where he worked on adaptive sinusoidal modeling. In the same year he joined the IRCAM to work on sound analysis, synthesis and transformation algorithms. In summer 2006 he was Edgar-Varese guest professor for computer music at the Electronic studio of the Technical University of Berlin and currently he is head of the Analysis/Synthesis Team at IRCAM. His current research interests are related to music and speech signal analysis and transformation.

Ewa Matusiak received the Master's degree in mathematics and electrical engineering from Oklahoma University, Norman, in 2003 and the Ph.D. degree in mathematics from the University of Vienna in 2007. She was a Postdoctoral Fellow in the Department of Electrical Engineering at the Technion - Israel Institute of Technology, Haifa, Israel from November 2008 to August 2010. She is currently with University of Vienna. Her interests lie in Gabor analysis, sampling theory, uncertainty principle and sparsity.

Marco Romito received the Ph.D. degree in mathematics from Pisa University in 2001. He is currently associate professor in probability at Pisa University and formerly lecturer in mathematical analysis at the University of Firenze. His research interests are in stochastic analysis, stochastic PDEs and the mathematical theory of turbulence.

Xavier Rodet 's research interests are in the areas of signal and pattern analysis, recognition and synthesis. He has been working particularly on digital signal processing for speech, speech and singing voice synthesis and automatic speech recognition. Computer music is his other main domain of interest. He has been working on understanding spectro-temporal patterns of musical sounds and on synthesis-by-rules. He has been developing new methods, programs and patents for musical sound signal analysis, synthesis and control. He is also working on physical models of musical instruments and nonlinear dynamical systems applied to sound signal synthesis. Xavier Rodet is currently emeritus researcher at IRCAM in the Analysis/Synthesis Team.