Adapted and adaptive linear time-frequency representations: a synthesis point of view

Peter Balazs, Monika Dörfler, Matthieu Kowalski, Bruno Torrésani

Adapted and adaptive linear time-frequency representations: a synthesis point of view

P. Balazs, *Senior Member, IEEE,*, M. Dörfler, M. Kowalski, and B. Torrésani, *Member, IEEE,*

Abstract

To display the time and frequency content of a given signal a large variety of techniques exist. In this paper, we give an overview of linear time-frequency representations, focusing mainly on two fundamental aspects. The first one is the introduction of flexibility, more precisely the construction of time-frequency waveform systems that can be adapted to specific signals, or specific signal processing problems. To do this, we base the constructions on frame theory, which allows a lot of options, while still ensuring perfect reconstruction. The second aspect is the choice of the synthesis framework rather than the usual analysis framework. Instead of the correlation of the signal with the chosen waveforms, i.e. the inner product with them, we look at how the signals can be constructed using those waveforms, i.e. find the coefficient in their linear combination. We show how this point of view allows the easy introduction of prior information into the representation. We give an overview over methods for transform domain modeling, in particular those based on sparsity and structured sparsity. Finally we present an illustrative application for these concepts: a denoising scheme.

I. Introduction

When we speak, we put words together, made up of vowels and consonants with quite distinct time-frequency characteristics. When we play music, we very often synthesize sound according to a time-frequency prescription, used for music notation. In both cases, time-frequency grains are implicitly used to synthesize sound. In general, a signal's existence starts with its synthesis.

P. Balazs is with the Acoustics Research Institute (ARI), Austrian Academy of Sciences, Wohllebengasse 12-14, 1040 Vienna, Austria

M. Dörfler is with the Numerical Harmonic Analysis Group (NuHAG), Faculty of Mathematics, University of Vienna, Alserbachstraße 23, 1090 Vienna, Austria

M. Kowalski is with the Laboratoire des Signaux et Systèmes (L2S) UMR 8506 SUPELEC-CNRS-Univ Paris-Sud, 3, rue Joliot-Curie, 91192 Gif-sur-Yvette cedex, France

B. Torrésani is with Aix-Marseille Université, CNRS, Centrale Marseille, LATP, UMR 7353, 13453 Marseille, France

However, the so-called *dictionaries* of time-frequency waveforms that are commonly used in time-frequency analysis (Gabor, wavelets,...) often does not qualify as physically sensible time-frequency grains. Their construction rules are simple, but rigid; therefore, in many applications, extra flexibility is needed. It is one of the main focuses of this survey paper to show how such flexibility can be introduced into the construction of time-frequency waveform systems. To do that, we will mainly rely on *frame theory*, which provides a convenient mathematical framework allowing a perfect control of signal synthesis.

The approach we shall base most of this paper upon is the *synthesis approach*. Instead of asking the question,

- *What is the best analysis for a signal of interest?*

we ask,

- *How is a signal best synthesized?*

Best, here, means, with the least effort (numerical cost), and the most satisfying results depending on the application (e.g. in denoising, with the best noise-suppression, but least artifacts). When working in the context of orthonormal bases, the synthesis and analysis questions are analogous by duality. The situation becomes both more subtle and rich when using more general time-frequency dictionaries. In particular, the introduction of redundancy, that is, allowing for dictionaries with a number of members higher than the dimension of the signal space, leads to a significant increase in design freedom. Redundant dictionaries always allow for (infinitely many) different ways to synthesize a signal of interest, or its components. By time-frequency representation we will therefore denote the family of coefficients from which the signal is synthesized, rather than the family of analysis coefficients, given as the inner products of a signal with the time-frequency waveforms. A nice feature of the synthesis approach is that it provides a generic way of introducing prior information, or constraints, into the representation. Considering this attempt to exploit prior knowledge, the question therefore becomes

- *How is a signal best synthesized, privileging certain behavior of its synthesis coefficients?*

As a popular example, the desired behavior is often that the coefficient family is sparse. Moreover, in real world signals, it rarely happens that isolated significant coefficients show up in the time-frequency plane. They rather tend to cluster in groups, or regions. A detailed account of approaches that incorporate this kind of behavior into time-frequency representations, together with a description of the corresponding algorithms, is another focus of this paper.

Summarizing, in this paper we elaborate on two basic lines of thought: the *adaptation of the dictionary*

chosen for a particular application, by designing frames according to certain time-frequency characteristics, and the *adaptation of the synthesis coefficients* used in the expansion obtained by means of a convenient dictionary.

The rest of the paper is organized as follows. We briefly review in Section II the basic principles of frame theory and time-frequency analysis that are needed in this paper. We then address in Section III the introduction of flexibility into time-frequency frame constructions, and then turn in Section IV to a discussion of synthesis domain modeling. Numerical examples are discussed in Section V, before briefly summarizing the main conclusions and perspectives.

Pictures, sound files and accompanying codes can be found at http://www.kfs.oeaw.ac.at/IEEESPM_AdaptiveAdapted.html.

II. Time-Frequency and Frames

In this paper we investigate the finite-dimensional setting and consider signals of length K, i.e. $\mathbf{x} \in \mathbb{C}^K$ considered as Hilbert space with the inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=0}^{K-1} x[k] \cdot \overline{y[k]}$. We denote the identity matrix by Id.

Linear time-frequency analysis is generally understood from the *transform* point of view: the most prominent example is a localized version of the Fourier transform [1]. However, due to the existence of inversion formulas, time-frequency analysis can also be understood as expanding signals as linear combinations of time-frequency atoms, i.e. generally time-frequency shifted copies of a reference window function $\mathbf{w}$. The prototype of time-frequency analysis is provided by the short-time Fourier transform (STFT) $\mathbf{x} \mapsto STFT_{\mathbf{w}}(\mathbf{x})$ defined by

$$STFT_{\mathbf{w}}(\mathbf{x})[l,m] = \sum_{k=0}^{K-1} \mathbf{x}[k]\mathbf{w}[k-l]e^{-2\pi i m(k-l)/K} \ , \tag{1}$$

where k is the time variable , $l = 0, \ldots, L-1$ the temporal shift and $m = 0, \ldots, M-1$ the frequency bin. Here in the (full) STFT case we have $K = L = M$. No additional boundary conditions are considered here, which, by the involved Fourier series corresponds to a circular assumption. In particular, also the shifts are considered to be circular. Other boundary assumptions are possible.

The STFT can be inverted in many different ways, including the very simple inversion formula (assuming $\|\mathbf{w}\| = 1$)

$$\mathbf{x}[k] = \sum_{l=0}^{K-1} \sum_{m=0}^{K-1} STFT_{\mathbf{w}}(\mathbf{x})[l,m]\mathbf{w}[k-l]e^{2\pi i m(k-l)/K} \ .$$

The latter can be seen as an expansion of $\mathbf{x}$ on the waveform system $\{\mathbf{w}_{lm}\}_{l,m}$: $\mathbf{w}_{lm}[k] = \mathbf{w}[k-l]e^{2\pi i m(k-l)/K}$. This is the *synthesis point of view*.

When the time and frequency variables n and m are subsampled, i.e. $L, M < K$, this is called the *Gabor transform*. The dictionary, here, is given by

$$\mathbf{w}_{lm}[k] = \mathbf{w}[k - la]e^{2\pi im(k-la)/M},$$

where the *hop size* a is chosen, such that $a \cdot L = K$, and $l = 0, \ldots, L, m = 0, \ldots, M$. In this case, the correspondence between analysis and synthesis is not as simple. Still, the Gabor transform can be inverted in a wide range of situations. The inversion takes the form of an expansion of the signal as linear combination of Gabor atoms again. This fact may actually be seen as a by-product of a more general theory, the theory of frames. This theory allows the extension to more general families of time-frequency atoms, which we will use as backbone in this paper.

We start with a short account of general frame theory. It is worth pointing out that, although the representation clearly depends linearly on the coefficients, we will depart from linearity as the coefficients will be allowed to depend non-linearly on the signal.

In the recent literature, the word "dictionary" is often employed to describe waveform systems with respect to which signals can be expanded. In the finite-dimensional setting considered here, dictionaries represent the same object as frames. We shall use both words indifferently.

A. Elements of frame theory

In a synthesis based approach, a representation of a signal $\mathbf{x} \in \mathbb{C}^K$ is searched for by additive synthesis of "building block" signals $\mathbf{u}_\lambda$ as

$$\mathbf{x} = \sum_{\lambda \in \Lambda} \alpha_\lambda \mathbf{u}_\lambda = \mathbf{U}\boldsymbol{\alpha} \ . \tag{2}$$

Here, signals and coefficients are represented as column vectors, and $\mathbf{U}$ is a matrix that contains the vectors $\mathbf{u}_\lambda$ as columns, and $\boldsymbol{\alpha} = (\alpha_\lambda)_\lambda$.

This representation can be chosen such that $\mathbf{U}$ is a square matrix and $\mathbf{U}^H\mathbf{U} = \mathbf{U}\mathbf{U}^H = \mathsf{Id}$, $\Lambda = [0, \ldots, K-1]$. In this case the vectors $\mathbf{u}_\lambda$ form an orthonormal basis, and the expansion coefficients α_λ of a given signal $\mathbf{x}$ are uniquely defined as $\boldsymbol{\alpha} = \mathbf{U}^H\mathbf{x}$, i.e. $\alpha_\lambda = \langle \mathbf{x}, \mathbf{u}_\lambda \rangle$. However, orthonormality is a very limiting condition, in particular for time-frequency building blocks, (see Section II-C), which often must be relaxed.

If the building blocks $\mathbf{u}_\lambda$ are chosen, such that $\mathbf{U}$ is still an invertible square matrix, the expansion coefficients are still uniquely defined, as $\boldsymbol{\alpha} = \mathbf{U}^{-1}\mathbf{x}$. Here the $\mathbf{u}_\lambda$ form a so-called Riesz basis.

The Riesz basis assumption can still be too restrictive; this problem led to the introduction of frames l [2, Chapter 5]. Frames are generally redundant, i.e. a signal has more coefficients than the dimension

of the space (as in STFT or Gabor); in such cases $\mathbf{U}$ is a rectangular $K \times N$ matrix with full rank and $N > K$, and has infinitely many right-inverses $\mathbf{A}$, such that $\mathbf{UA} = \mathsf{Id}$. In the redundant case the expansion coefficients are not uniquely defined. Among the infinitely many families of expansion coefficients, a specific one is often privileged, obtained by using the pseudo-inverse $\mathbf{U}^\dagger$ of $\mathbf{U}$: $\boldsymbol{\alpha} = \mathbf{U}^\dagger \mathbf{x}$. The rows of $\mathbf{U}^\dagger$ form the so-called canonical dual frame $\tilde{\mathbf{u}}_\lambda$. The rows of any other right inverse are called a dual frame.

The matrix $\mathbf{U}$, as in (2), corresponds to the synthesis operator, $\mathbf{U}^H$ to the analysis operator, given by $\left(\mathbf{U}^H x\right)_\lambda = \langle \mathbf{x}, \mathbf{u}_\lambda \rangle$. The so-called frame operator $\mathbf{S}$ is defined by $\mathbf{S} = \mathbf{U}\mathbf{U}^H$, i.e. $\mathbf{S}x = \sum_{\lambda \in \Lambda} \langle \mathbf{x}, \mathbf{u}_\lambda \rangle \, \mathbf{u}_\lambda$. Here $\Lambda = [0, \ldots, N-1]$

The frame operator is a square, positive definite matrix and invertible, because $\mathbf{U}$ has full rank. In this case, reconstruction is straight-forward, since

$$\mathbf{x} = \mathbf{S}\mathbf{S}^{-1}\mathbf{x} = \sum_\lambda \left\langle \mathbf{S}^{-1}\mathbf{x}, \mathbf{u}_\lambda \right\rangle \mathbf{u}_\lambda \;\; = \sum_\lambda \langle \mathbf{x}, \widetilde{\mathbf{u}}_\lambda \rangle \, \mathbf{u}_\lambda. \tag{3}$$

The inversion required to compute the dual frame can often be efficiently performed by using inversion algorithms such as conjugate gradients or iterative approaches, cf. [3] and references therein.

If $\mathbf{S} = A \cdot \mathsf{Id}$ for some constant $A > 0$, the frame is called *tight*. In this case the frame is self-dual (up to the normalization factor A). Starting from any frame $\mathbf{U}$, a tight frame $\mathbf{U}^t$ is given by $\mathbf{U}^t = \mathbf{S}^{-1/2}\mathbf{U}$.

B. Gabor frames

In signal processing time-frequency related frames have been used ubiquitously. The subsampled *Short-time Fourier transform* (STFT) or Gabor transform [3] corresponds to a system

$$\left(\mathbf{G}_{\mathbf{w},a,M}\right)_{k,l+Lm} = \mathbf{w}[k - la] \, e^{2\pi i m(k-la)/M},$$

where $\mathbf{w}$ is the window, a is the hop size and M is the number of FFT bins, i.e. the number of channels. Here $k = 0, \ldots, K-1$, $l = 0, \ldots, L-1$ and $m = 0, \ldots, M-1$. In this setting the number of frame elements is $N = L \cdot M$.

If the system represented by the $K \times (L \cdot M)$-matrix $\mathbf{G}_{\mathbf{w},a,M}$ forms a frame, the existence of expansions as in (3) is guaranteed using a dual frame for computing the coefficients. The canonical dual frame of a Gabor frame has the same structure again, i.e. a Gabor system with the canonical dual window $\widetilde{\mathbf{w}}$ [3, Chapter 5]. Applying the synthesis operator $\mathbf{G}_{\mathbf{w},a,M}$ is equivalent to the overlap add reconstruction method, performing an inverse Fourier transform with the synthesis window $\mathbf{w}$ as weights and subsequent addition.

In the more classical *analysis point of view*, the Gabor transform or STFT of a signal $\mathbf{x}$ with window $\mathbf{w}$ is obtained by applying to $\mathbf{x}$ the analysis operator $\mathbf{G}_{\mathbf{w},a,M}{}^H$ as in (1). This is equivalent to the windowed Fourier transform or modulated filter bank viewpoint of the STFT (or the phase vocoder) [4].

In the 'full STFT' case, i.e. $K = L = M$, the analysis can be inverted by an inverse Fourier transform, summing all channels up and dividing by $\hat{\mathbf{w}}(0) \cdot L$, assuming $\hat{\mathbf{w}}(0) \neq 0$. This amounts to choosing the constant window $\mathbf{v} = 1$ for the synthesis window. Actually, any window $\mathbf{v}$ that is not orthogonal to $\mathbf{w}$ can serve as dual window in that case.

For any other subsampled setting, the inversion can be done in a similar way only if a certain condition, the constant overlap-add constraint, is fulfilled, i.e. $\sum_{l=0}^{L-1} \mathbf{w}\,[k - la] = \text{const.}$ for all k. For frames, also in an analysis approach, dual windows always allow perfect reconstruction. The connection with the synthesis point of view described above corresponds to the usage of $\tilde{\mathbf{w}}$ as analysis window to obtain the perfect reconstruction coefficients for the synthesis window $\mathbf{w}$.

The particular structure of a Gabor system can be exploited to make the calculations more efficient, see e.g. [3], where also several sufficient conditions for a waveform system to form a frame can be found. One particular setting, often used in practice and termed the *'painless case'* [5] gives an immediate frame-criterion and allows for very fast calculation of the canonical dual window. In this setting the number of FFT-bins M is at least equal to the length of the window $\mathbf{w}$. In this case the frame operator $\mathbf{S} = \mathbf{G}_{\mathbf{w},a,M}\mathbf{G}^H_{\mathbf{w},a,M}$ is a diagonal matrix. If, furthermore, its diagonal entries $\sum_{l=0}^{L-1} \mathbf{w}\,[k - la] > 0$, then the system is a frame and $\mathbf{S}$ is easily invertible. The canonical dual window, in this case, corresponds to a re-weighted original window.

C. The benefits of frames and frame theory

Frames have been used for a long time in signal processing and audio applications, mostly implicitly. The parameters usually chosen for standard time-frequency analysis and modeling correspond to a frame with redundancy two or four. (Often the painless case is used, e.g. a Hann window with 75% overlap, with the FFT size equal to the window size.)

So it seems somehow quite natural to use frames, but *what are the concrete advantages of frames and frame theory for signal processing?*

a) Mathematical background: Frame theory has been a very active field of applied analysis during the last 2 decades. For time-frequency analysis frame theory opens the door to generalizations such as the ones to be discussed in the next section.

While the infinite-dimensional theory (for continuous time signals) dominates the theoretical results, analogies with function spaces have proven to be quite relevant for applications using long signals or signals with variable length. For example, the mixed norm models to be discussed in Section IV are in the spirit of the so-called modulation spaces [3, Chapter 11], originally introduced in the continuous time setting.

b) More freedom: From the definitions in Section II-A it is clear that the design restrictions for frames are less severe than for bases. In some applications certain side constraints arise for the used representation. It is often easier and numerically more feasible to construct frames than bases fulfilling these a-priori conditions. For particular constraints, it can even be impossible to construct corresponding bases, while frames can be found. For example, as stated in the Balian-Low theorem, see e.g. [3, Chapter 7], there is no window function which is simultaneously well-localized in both time and frequency and results in a Gabor Riesz basis. That means that in time-frequency analysis, as reflected in the methods used in applications, redundancy is a necessary requirement in order to obtain good time-frequency representations.

c) Guaranteed Perfect Reconstruction: While classical time-frequency approaches, such as those used in the phase vocoder, have been carefully designed to guarantee the invertibility of the transform, frame theory provides a generic framework for transforms and synthesis, with verifiable criteria for arbitrary windows. Besides, these are not limited to STFT or Gabor frames, but can be extended to many different classes of time-frequency frames, see the next section.

d) Sparsity: Last but not least, frame theory provides an adequate setting for time-frequency sparse approximation techniques. While sparse approximation techniques can also be developed for orthonormal and Riesz bases, they turn out to be more efficient when used in conjunction with frames. The language of synthesis and analysis operators we just outlined above turns out to be particularly relevant and adequate for formulating sparse and sparse/structured signal decomposition problems, as we shall see in Section IV below.

III. Introducing flexibility into time-frequency frames:

As pointed out in the previous section, Gabor frames provide more flexibility than Gabor (Riesz) bases. Intuitively, a dictionary that allows for signal-synthesis with few significant coefficients should contain atoms similar to the signal components to be represented. A regular structure of the sampling set, in conjunction with the usage of just one single analysis window is often too restrictive to meet this kind of design criteria arising in applications. Since natural signals usually will have various components with

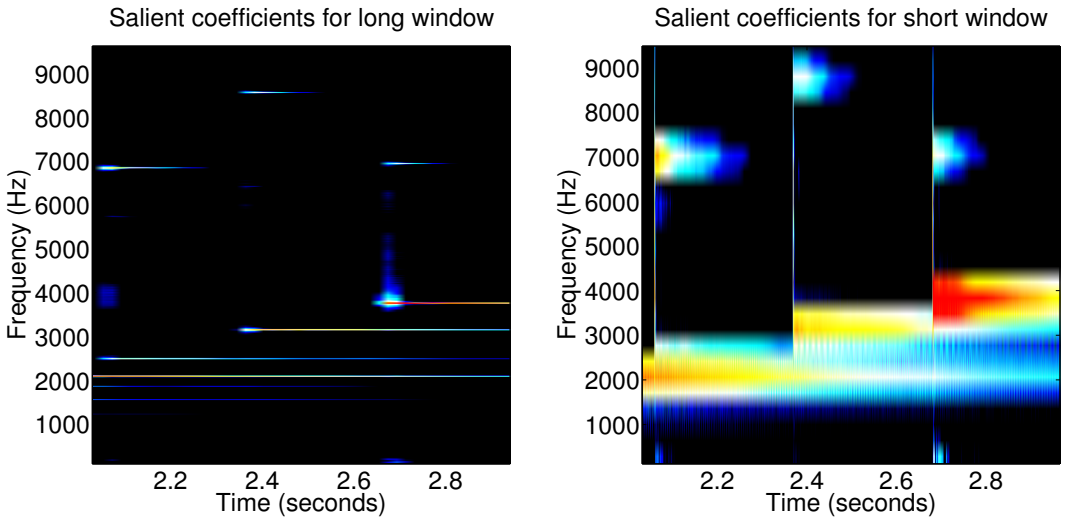

Fig. 1. Glockenspiel-Signal. Gabor representations with long window (92.9 ms, left), resp. short window (2.9 ms, right).

distinct time-frequency localization properties, diverse Gabor frames are differently adapted to certain signal properties. As an illustrating example, Figure 1 shows two versions of Gabor coefficients used for signal synthesis as in (2) with either a narrow (short time duration) or a wide (long time duration) Hann window. In both cases, the redundancy is 4 and the dictionary is a tight frame, hence, appropriate coefficients are obtained as inner products with the same time-frequency shifted window. The plots show the most salient coefficients, that is, those whose value is at least 10% of the biggest coefficient's value. Apparently, signal components, which are localized in time, are much more concisely encoded by the coefficients corresponding to the short window. On the other hand, the more sinusoidal components require many more non-zero coefficients if synthesized with a narrow window, while, as depicted in the left plot, the coefficients corresponding to the long window are comparatively sparse. This example now hints at a fundamental problem one faces when forced to choose one particular size and shape of window for the representation of the entire signal: the trade-off between good time – and good frequency – concentration. Due to this observation, the desire to change, and thus adapt, the time-frequency resolution over time or frequency, or time *and* frequency, arose. It has led to various approaches enabling more flexible resolution of the time-frequency plane.

A. Paving the phase space: Gabor based constructions

Early approaches to a flexible *paving* or *tiling* of the time-frequency plane in a signal-adaptive way mostly suggested the construction of orthonormal bases with well-defined properties, such as wavelet packets, adapted local trigonometric bases (aka modulated lapped transforms) and other local orthogonal bases, see [2] and references therein. The general idea present in all these approaches is to deviate from a regular tiling of the frequency-plane, as present in a STFT or Gabor transform, for the sake of improved time- or frequency-resolution where required.

In the context of redundant time-frequency representations, adaptive time – or frequency – resolution has recently been introduced by investigating a variant of the classical (regular) Gabor frames, denoted *nonstationary Gabor (NSG) frames* [6]. This new approach directly generalizes classical Gabor frames by allowing for a set of general windows $\mathbf{w}_l\,[k - la_l]$ to replace the regular translates $\mathbf{w}[k - la]$, again modulated in order to obtain the entire system, however, with time-varying frequency-sampling parameters:

$$\mathbf{w}_{l,m}[k] = \mathbf{w}_l[k - la_l] \cdot e^{2\pi i m(k-la_l)/M_l},$$

such that the number of channels M_l and hop size a_l depend on the l-th window $\mathbf{w}_l$.

A fast implementation of NSG frames can be based on the condition that the windows have short length. In this situation, it is useful to consider the structure of the frame operator $\mathbf{S}$: for overlapping windows and sufficiently dense frequency samples, corresponding to an FFT-length equal to or higher than the window-length, $\mathbf{S}$ is diagonal and checking the frame condition is very easy, and if possible, the inversion is thus straight-forward. (This is again termed the 'painless' case.) In the currently available implementations, adaptivity is possible in *either* time or frequency. For adaptivity in frequency, fast processing relies on a pre-processing step that entails an FFT applied to either the whole signal or, to allow for real-time applications, to time-slices of the signal, as investigated in [7]. Note that this setting also allows for the implementation of constant-Q transforms, for the original idea see [8] and references therein, wavelet frames [2] and other filter-banks, e.g. inspired by perceptive scales like the ERB scale [9].

Together with Fig. 1, Fig. 2 presents a comparison of the coefficients used in regular Gabor transforms with those obtained from two nonstationary Gabor transforms: one with time adaptivity, where the onsets were first detected, one with a constant-Q scale on the frequency, cf. [7]. Note that, due to the time adaptivity, the onsets are well localized in time, while the good frequency resolution is kept between the onsets. On the other hand, the constant-Q permits a time-frequency representation well adapted for

music, with the best frequency resolution for low frequencies and increasingly fine time resolution at higher frequencies. Several examples of non-stationary Gabor transforms and the synthesized signals with respect to the corresponding dictionaries can be found at http://www.univie.ac.at/nonstatgab/.

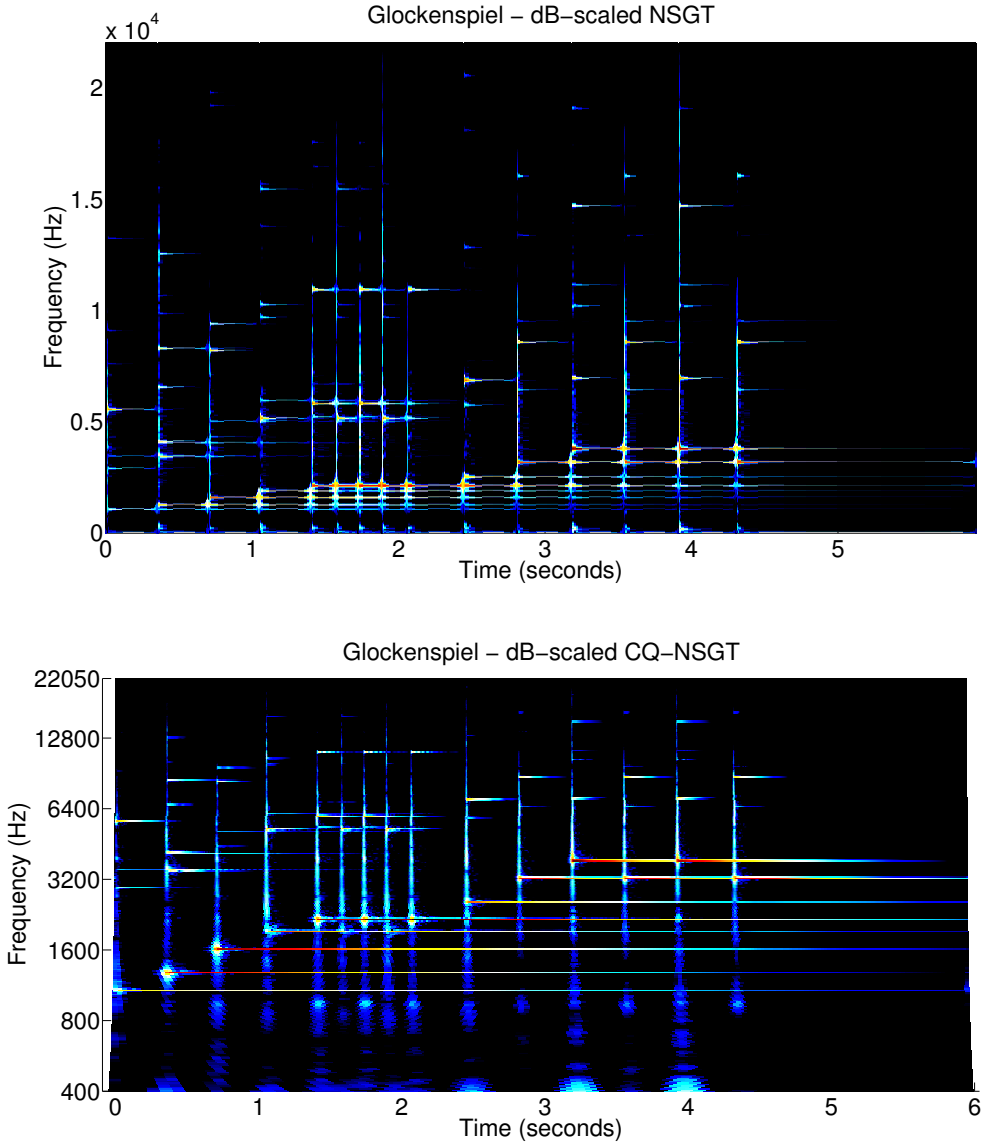

Fig. 2. Glockenspiel-Signal: Nonstationary Gabor representation with time-adaptivity (upper plot) and frequency-adaptivity, corresponding to constant-Q scale (lower plot).

B. Unions of frames, and the components separation problem

A complementary approach to the models described in the previous section is to consider a union of two or more frames or bases, cf. [10]. In both cases, the resulting dictionary is still a frame. The construction usually aims at designing rather different frames – or bases – in order to obtain a good, that is, sparse, representation of each class of expected signal components. While, in the NSGT, adaptation is done locally, the model based on the union of several frames, takes a global point of view to encoding distinct signal components. The underlying assumption can be expressed as the expectation that each class of components, each layer of the signal, has a sparse decomposition within one, and only one, of the frames in the union. The prototypical problem addressed in such a setting is the decomposition of music signals into a tonal, a transient and a residual noise layer. The proposed dictionaries range from the union of wavelet and local cosine bases [11], to the union of Gabor frames with very different window characteristics, to dictionaries which are designed according to even more concrete knowledge about the underlying signal class, as e.g. in [12]. The algorithms described in Section IV below can be adapted to handle such situations. However, the performances of all these separation methods depend crucially on the chosen bases or frames. Quite obviously, the more different the two frames, the better the performance. This statement can be made quantitative by introducing a measure of difference between frames, such as coherence or cumulative coherence, see e.g. [13].

C. Signal-dependent adaptivity

Ideally, the representation system should (to some extend) be adapted to the signal at hand. For example, some MP3-type audio coders perform some prior analysis to switch between short and long windows for the coding. Adaptive transforms were proposed systematically in the seminal work of Coifman and Wickerhauser [14], who proposed to use entropies to measure the fit of the basis with the signal. This approach lead to the so-called *best-basis* algorithm, dedicated to specific, tree-structured, families of time-frequency decompositions. In more general situations, adaptation is often performed via application-driven, ad hoc procedures. This is the case of the approach that was used to perform the analysis leading to Figure 2. In the upper display of Figure 2, the onsets, where short-lived and thus time-concentrated signal components are expected, were automatically extracted from the signal by a standard peak-tracking algorithm. Subsequently, the windows were adapted such that narrow windows are used near the transient signal components. In [6] this algorithm, based on *painless* NSG frames was evaluated and shown to lead to a more sparse representation of the signal of interest.

IV. Transform and coefficient domain modeling:

Redundancy means freedom: given a redundant frame, any signal $\mathbf{x}$ has infinitely many expansions of the form (2). So, to select a satisfactory expansion, additional criteria can be introduced, and transform domain modeling offers a convenient way to exploit prior information about the desired expansion. For example, the choice $\alpha_\lambda = \langle \mathbf{x}, \widetilde{\mathbf{u}}_\lambda \rangle$ results from the minimization of $\|\boldsymbol{\alpha}\|_2$ under the constraint (2).

Extra flexibility is often introduced into transform domain models, to account for possible deviations from the model. In such cases, a standard alternative to (2) is the following model

$$\mathbf{x} = \sum_{\lambda \in \Lambda} \alpha_\lambda \mathbf{u}_\lambda + \boldsymbol{\epsilon} \ , \tag{4}$$

where $\boldsymbol{\epsilon}$ is some residual signal (noise), often modeled in the literature as a Gaussian white noise.

A. Sparse synthesis model

As mentioned above, the minimization of $\|\boldsymbol{\alpha}\|_2$ under the constraint (2) does not give more than standard frame theory. Other signal models are obtained by minimizing other norms of the coefficient sequence, such as ℓ^p norms, under the same constraint.

1) Basis pursuit and penalized regression model: The ℓ^p norm of a sequence is a measure of concentration (for $p > 2$), or spreading (for $p < 2$). The latter case was found particularly interesting in a number of situations, where it was shown to lead to *sparse expansions*, i.e. expansions involving a small number of non-zero (or non-negligible) atoms. Among them, the cases $p \geq 1$ are interesting because the corresponding ℓ^p norm is indeed a norm, and is therefore convex; the cases $p \leq 1$ are also interesting because the corresponding (quasi) norm is non-differentiable, and therefore yields some thresholding operation during the minimization procedure [15]. As a consequence, the case $p = 1$ is often privileged; it leads to the so-called *basis pursuit* problem: given $\mathbf{x} \in \mathcal{H}$, solve

$$\min_{\boldsymbol{\alpha} \in \ell^2(\Lambda)} \|\boldsymbol{\alpha}\|_1 \quad \text{under constraint} \quad \mathbf{x} = \sum_{\lambda \in \Lambda} \alpha_\lambda \mathbf{u}_\lambda \ . \tag{5}$$

When observation noise is introduced into the model, the strict equality constraint is relaxed to (4), which yields the *basis pursuit denoising*, or *Lasso* problem when the ℓ^1 norm is used. The use of ℓ^2 norm is known as *ridge regression* problem, or *Tikhonov regularization*. Replacing the ℓ^1-norm by a functional Φ on the coefficient space, and introducing a tuning parameter μ, the problem to be solved takes the form

$$\min_{\boldsymbol{\alpha} \in \ell^2(\Lambda)} \left[\frac{1}{2} \left\| \mathbf{x} - \sum_{\lambda \in \Lambda} \alpha_\lambda \mathbf{u}_\lambda \right\|_2^2 + \mu \Phi[\boldsymbol{\alpha}] \right] \ . \tag{6}$$

2) Multilayered expansions via coefficient selection: As stressed in Section III-B a finite union of frames is still a frame, to which the techniques described in the previous sections can be applied blindly. However, sticking to the idea that signal components are best encoded when the frame contains waveforms that resemble them, it is natural to construct unions of time-frequency frames specially tailored to account for specific signal features, cp. [12], among many other recent references.

In variational approaches, the signal separation problem addressed in Section III-B may be formulated as follows: given a signal $\mathbf{x}$ and two frames $\mathbf{U}$ and $\mathbf{V}$, splitting $\mathbf{x}$ into two parts $\mathbf{x}_\mathbf{U}$ and $\mathbf{x}_\mathbf{V}$ can be achieved by solving

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\frac{1}{2} \left\| \mathbf{x} - \sum_{\lambda} \alpha_\lambda \mathbf{u}_\lambda - \sum_{\lambda'} \beta_{\lambda'} \mathbf{v}_{\lambda'} \right\|_2^2 + \mu \Phi[\boldsymbol{\alpha}] + \mu' \Phi'[\boldsymbol{\beta}] \right] \tag{7}$$

and the two parts are given by

$$\mathbf{x}_\mathbf{U} = \sum_{\lambda} \alpha_\lambda \mathbf{u}_\lambda \ , \quad \mathbf{x}_\mathbf{V} = \sum_{\lambda'} \beta_{\lambda'} \mathbf{v}_{\lambda'} \ .$$

The properties of the two parts $\mathbf{x}_\mathbf{U}$ and $\mathbf{x}_\mathbf{V}$ obviously depend on the choices of the regularization functionals Φ and Φ', which can be any of the previously described choices – or others –, but also on the hyperparameters μ and μ'. The reader may refer to [16] for instructive examples and further references.

B. Structured sparsity

In some situations, for example when $\lambda = (t, f)$ is a time-frequency index, more purposive models can be introduced via the regularization term Φ. Popular choices for Φ are provided by mixed norms, which we discuss first, before turning to more sophisticated techniques.

1) Mixed norms: Mixed norms offer a flexible framework to generate sensible time-frequency signal models, by promoting different coefficient behaviors as a function of the index. Sticking to the time-frequency case, the following family of norms is considered: given a coefficient sequence $\boldsymbol{\alpha}$, set

$$\|\boldsymbol{\alpha}\|_{p,q} = \left[\sum_t \left(\sum_f |\alpha_{t,f}|^q \right)^{p/q} \right]^{1/p} \tag{8}$$

Clearly, the roles of t and f can be interchanged, leading to a totally different model.

For instance, with the definition given in (8), minimizing $\|\boldsymbol{\alpha}\|_{p,q}$ with $p < 2$ and $q > 2$ will promote diversity within each group of fixed time coefficients, while these groups will be sparsely represented: only a few groups, containing all members, will emerge. In time-frequency dictionaries, this typically happens for time-indexed groups, whenever a transient is present in audio-signals. This situation will be

illustrated in Section V by displaying the transients extracted from a musical signal. The particular case $p = 1$, $q = 2$ leads to the so-called *group-Lasso* problem, introduced in [17].

On the other hand, still with definition (8), minimizing $\|\boldsymbol{\alpha}\|_{p,q}$ with $p > 2$ and $q < 2$ will promote sparsity within each group of fixed time coefficients, with no preference between groups: only few coefficients will be selected within each groups. The particular case $p = 1$, $q = 2$ leads to the so-called *elitist-Lasso* problem, studied in [16]. By switching the role of t and f in (8), the groups will be defined in frequency instead of being defined in time.

For the sake of simplicity, we have so far restricted ourselves to coefficient groups defined by a fixed value of the time (or the frequency) index, i.e. vertical (or horizontal) lines in the time-frequency plane. However, splitting of the time-frequency coefficient domain Λ into fixed-time or fixed-frequency groups can be replaced by an arbitrary splitting of the index set $\Lambda = \cup_{g=1}^{G} \Lambda_g$ into G groups $\Lambda_g = \{\lambda_{gm}, m = 1, \ldots, M_g\}$ (see [18] for a more general discussion on the splitting of the index set).

The regularization functionals described above, and in particular the group-Lasso, are based upon a splitting onto non-overlapping groups. However, as we shall see in the application section below, the need for overlap between groups appears when some short range dependence between groups needs to be incorporated in the synthesis model. We describe below some attempts in that direction.

2) Neighborhood and overlapping groups: Sticking to the convex variational formulation, we shall describe two approaches that have been proposed to introduce overlap between groups. For example, in order to model some time-frequency persistence, one can associate with each time-frequency index λ a neighborhood $\mathcal{N}(\lambda)$. Such neighborhood systems $\{(\lambda, \mathcal{N}(\lambda)),\, \lambda \in \Lambda\}$ thus generate some "groups with overlap". The first possibility is to define a regularization term exploiting this overlap, by solving as in [19] the optimization problem

$$\min_{\boldsymbol{\alpha} \in \ell^2(\Lambda)} \left[\frac{1}{2} \left\| \mathbf{x} - \sum_{\lambda \in \Lambda} \alpha_\lambda \mathbf{u}_\lambda \right\|_2^2 + \mu \sum_{\lambda \in \Lambda} \left(\sum_{\ell \in \mathcal{N}(\lambda)} |\alpha_\ell|^2 \right)^{1/2} \right] . \tag{9}$$

However, as stressed in [20], Lasso type-approaches behave as a *discarding* procedure instead of *selection*: using group-Lasso with overlap, removing any group containing a coefficient removes this coefficient from the final representation. For an intuitive interpretation of this discarding behavior please refer to the thresholding operators in the next subsection. The latent-variable group-Lasso proposed then to solve

$$\min_{\tilde{\alpha}} \left[\frac{1}{2} \left\| \mathbf{x} - \sum_{\substack{\lambda \in \Lambda \\ \ell \in \mathcal{N}(\lambda)}} \tilde{\alpha}_\lambda^\ell \mathbf{u}_\lambda \right\|_2^2 + \mu \sum_{\lambda \in \Lambda} \left(\sum_{\ell \in \mathcal{N}(\lambda)} |\tilde{\alpha}_\lambda^\ell|^2 \right)^{1/2} \right] , \tag{10}$$

where latent variable $\tilde{\boldsymbol{\alpha}}$ can be viewed as a spread version of $\boldsymbol{\alpha}$, in which some coordinates have been duplicated according to the neighborhood structure. Here, a coefficient will be discarded from the final support only if it is discarded from every group it belongs to.

As we will see in the next section, problems (9) and (10) may be difficult to solve efficiently, even when $\mathbf{U}$ is orthogonal. However, each sensible choice of prior term Φ leads to corresponding *shrinkage* strategies which can be exploited within optimization algorithms. Based on this shrinkage operator, new structured shrinkage operators can be defined, to overcome the drawbacks of (9) and (10).

C. Shrinkage, proximal operators and algorithms

Minimization of convex functions like (6) relies on the so-called proximity operators of convex penalties. The proximity operators can be obtained by solving (6) when $\mathbf{U}$ is an orthogonal basis, and typically lead to shrinkage/thresholding operators such as the well-known soft-thresholding. For example, using $\tilde{\mathbf{y}} = \mathbf{U}^H\mathbf{y}$, the proximity operator $\mathcal{S}_\mu : \ell^2(\Lambda) \to \ell^2(\Lambda)$ associated to a mixed norm (8) is defined, in dependence on the parameter μ, by

$$\mathcal{S}_\mu(\tilde{\mathbf{y}}) = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\|\tilde{\mathbf{y}} - \boldsymbol{\alpha}\|_2^2 + \mu \sum_t \left(\sum_f |\alpha_{t,f}|^p \right)^{q/p} . \tag{11}$$

for the Lasso ($p = q = 1$ in (11)) and group-Lasso ($p = 2$ and $q = 1$ in (11)), $\mathcal{S}_\mu$ is given coordinate-wise by

$$\alpha_{t,f} = \frac{\tilde{y}_{t,f}}{|\tilde{y}_{t,f}|}(|\tilde{y}_{t,f}| - \mu)^+ , \tag{12}$$

$$\alpha_{t,f} = \tilde{y}_{t,f} \left(1 - \frac{\mu}{\sqrt{\sum_{f'} |\tilde{y}_{t,f'}|^2}} \right)^+ , \tag{13}$$

respectively, where $(x)^+ = \max(x, 0)$.

In the redundant case, when $\mathbf{U}$ is not orthogonal anymore, solving problem (6) requires an efficient optimization algorithm. When (6) is convex, but not necessarily differentiable, most such algorithms are build upon the proximity operator of the regularizer, and then belong to the iterative shrinkage/thresholding algorithms (ISTA) family. These algorithms perform a gradient descent on the ℓ_2 data term, and then shrink/threshold it by applying the proximity operator of the regularizer. The reader can consult [21] for a review of convex optimization algorithms for Lasso-like problems (6). For the sake of completeness, we provide in Algorithm 1 the pseudocode of the accelerated version FISTA [22]. Nevertheless, closed form expressions for the proximity operators cannot always be obtained. This is particularly true if one

wants to solve the group-Lasso with overlaps (9), where an additional inner loop has to be performed in order to compute the proximity operator.

Based on the proximal operators corresponding to mixed norms, one can construct new thresholding operators in order to *select*, rather than discard, certain coefficients or groups of coefficients. Using the same neighborhood structure as for the group-Lasso with overlaps and latent-group-Lasso, the *windowed-group-Lasso* was introduced in [16] by the following operator:

$$\alpha_\lambda = \tilde{y}_\lambda \left(1 - \frac{\mu}{\sqrt{\sum_{\ell \in \mathcal{N}(\lambda)} |\tilde{y}_\ell|^2}} \right)^+ . \tag{14}$$

In [23], this and related operators were termed "social sparsity"-models and were shown not to be proximity operators. However, the windowed-group-Lasso (14) behaves particularly well in practice, and has the big advantage over the group-Lasso with overlaps and the latent variable group-Lasso, that the dimension of the problem does not explode because of the duplication of variables[1].

Algorithm 1: FISTA

Initialization: $\boldsymbol{\alpha}^{(0)} \in \mathbb{C}^N$, $k = 1$, $\gamma = \|\mathbf{U}\mathbf{U}^H\|$, $\mathbf{z}^{(0)} = \boldsymbol{\alpha}^{(0)}$, $\tau^{(0)} = 1$.

repeat

$\boldsymbol{\alpha}^{(k)} = \mathcal{S}_{\frac{\mu}{\gamma}} \left(\mathbf{z}^{(k-1)} + \frac{1}{\gamma} \mathbf{U}^H (\mathbf{y} - \mathbf{U}\mathbf{z}^{(k-1)}) \right)$;

$\tau^{(k)} = \frac{1}{2} \left(1 + \sqrt{1 + 4{\tau^{(k-1)}}^2} \right)$;

$\mathbf{z}^{(k)} = \boldsymbol{\alpha}^{(k)} + \frac{\tau^{(k-1)} - 1}{\tau^{(k)}} (\boldsymbol{\alpha}^{(k)} - \boldsymbol{\alpha}^{(k-1)})$;

$k = k + 1$

until *convergence*;

One major advantage of ISTA algorithms is that they only involve analysis and synthesis operations. Hence, the frame matrix $\mathbf{U}$ does not have to be explicitly available, fast algorithms can be used for the required matrix-vector multiplications. Moreover, ISTA can be extended in a straightforward way to more general inverse problems.

[1]The website http://homepage.univie.ac.at/monika.doerfler/StrucAudio.html provides a wealth of examples of the structured sparsity approach applied in audio processing.

V. ILLUSTRATIONS

Within the setting of Gabor frames, the combination of the synthesis point of view and the sparsity principle, has proven to be useful in various application, particularly to tackle difficult inverse problems. A natural application area of time-frequency representation is audio processing [24]. Gabor frames are also used successfully in biomedical imaging for problems such as magnetoencephalography (MEG) source localization [25], and also in image processing [26].

Here, we will demonstrate the impact of structured sparsity on the most basic inverse problem: audio-denoising. This simple problem provides a good framework to illustrate the results that can be obtained with such a model. The Matlab toolboxes used for the implementations of the various approaches in the subsequent experiments are:

- The LTFAT toolbox, which provides an implementation of the Gabor analysis and synthesis operations, with a C backend for efficiency; available at http://ltfat.sourceforge.net. See [27] for a review.
- The StrucAudioToolbox, which provides an implementation of various thresholding operator, and in particular the social-sparsity operators; available at http://homepage.univie.ac.at/monika.doerfler/StrucAudio.html
- The NSGT (NonStationary Gabor Transform) toolbox, which provides the implementation, based on the LTFAT toolbox, of the non-stationary Gabor frames, in particular the constant-Q transform. Available at http://www.univie.ac.at/nonstatgab/

The general setup we chose is the following. We work on a 3 second-excerpt of a Jazz signal, containing piano and drums, sampled at 22050 Hz. Then, a white Gaussian noise is added to obtain a final signal to noise ratio (SNR) of 10 dB.

Because of the drums and the impact of the piano hammers on the strings, the signal contains transients, well localized in time. On the other hand, the vibrating piano strings produce sinusoidal components and thus constitute the tonal layer. Consequently, we chose the following direct model for the noisy signal $\mathbf{x}$:

$$\begin{aligned}\mathbf{x} &= transient + tonal + noise \\ &= \mathbf{U}\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\beta} + noise\end{aligned}$$

where $\mathbf{U}$ is a Gabor frame adapted to the transient, and $\mathbf{V}$ a Gabor frame adapted for the tonal. For $\mathbf{U}$, we chose a tight frame with a short Hann window of length about 6 ms. For $\mathbf{V}$, we played with two different choices: one tight frame with a long Hann window of length about 46 ms, and a constant-Q Gabor frame, with 12 bins per octave.

The denoised signal, as well as the two denoised layers, are obtained by running the FISTA Algorithm 1. The model corresponding to the different signal layers is completed, in correspondence to the methods described in Section IV, by choosing the following thresholding operators for the two layers:

- For the transient layer, we stick to the group-Lasso thresholding operator (13) in order to obtain sharp transients, well localized in time. Accordingly, the groups are defined as to contain *all* the frequencies for each time index. (Gabor-GL)
- For the tonal layer, we compared two different thresholding operators:
  - The soft-thresholding operator corresponding to the Lasso. By this choice, no structures are taken into account. This operator was used with both dictionaries corresponding to the tonal layer, resulting in Gabor-L and CQ-Gabor-L.
  - The windowed-group-Lasso social-sparse thresholding operator (14), with a time-neighborhood of 4 coefficients before and after the considered time-frequency coefficients. Such a neighborhood promotes a *time persistent* structure. This operator is applied to the regular tight Gabor frame. resulting in Gabor-WG.

In summary, we have three models to compare, hereafter referred to as Gabor-GL+Gabor-L, Gabor-GL+Gabor-WGL, Gabor-GL+CQ-Gabor-L. The threshold parameters were tuned in order to reach the best SNR, that are respectively: 17.6 dB, 18.5 dB and 18.1 dB.

The time-frequency synthesis coefficients obtained after convergence of the algorithm using each of the corresponding thresholding operators are depicted in Fig. 3. In the left-hand plots, the three obtained transient layers are shown. As expected, the use of the group-lasso penalty provides very sharp transients, well localized in time. When listening to the reconstructed transient signal for Gabor-GL+Gabor-L and Gabor-GL+Gabor-WGL, one can hear the attack of the piano. The transient components are thus well-captured by the model. With the Gabor-GL+CQ-Gabor-L model, the expected transient layer has very little energy. Indeed, with the CQ dictionary, a lot of transient information is also captured in the high frequencies as the window becomes shorter, and then are present in the "tonal" layer.

The estimated tonal components are shown in the right-hand plots. One can see that the three models (Gabor-L, Gabor-WGL, CQ-Gabor-L) provide three rather distinct representations. The use of the windowed-group-Lasso operator provides a time-frequency representation which is more structured, with a certain time persistence of the coefficients. Moreover, this model outperforms the others, in particular the plain Lasso, in terms of SNR. It should be pointed out that according to perceptive criteria, the results obtained from the structured approach are even more convincing than what SNR suggests (though an

objective evaluation of such criteria objectively requires much care, and is extremely time-consuming if one wants to go beyond a rough estimation). Finally, for CQ-Gabor-L the a-priori known signal-structure is directly modeled in the dictionary, rather than by the threshold operator. We obtain time-frequency coefficients which are better localized in frequency for the low frequency, and correspond to the harmonic structures of the musical signal, but we also obtain transient information in high frequencies.

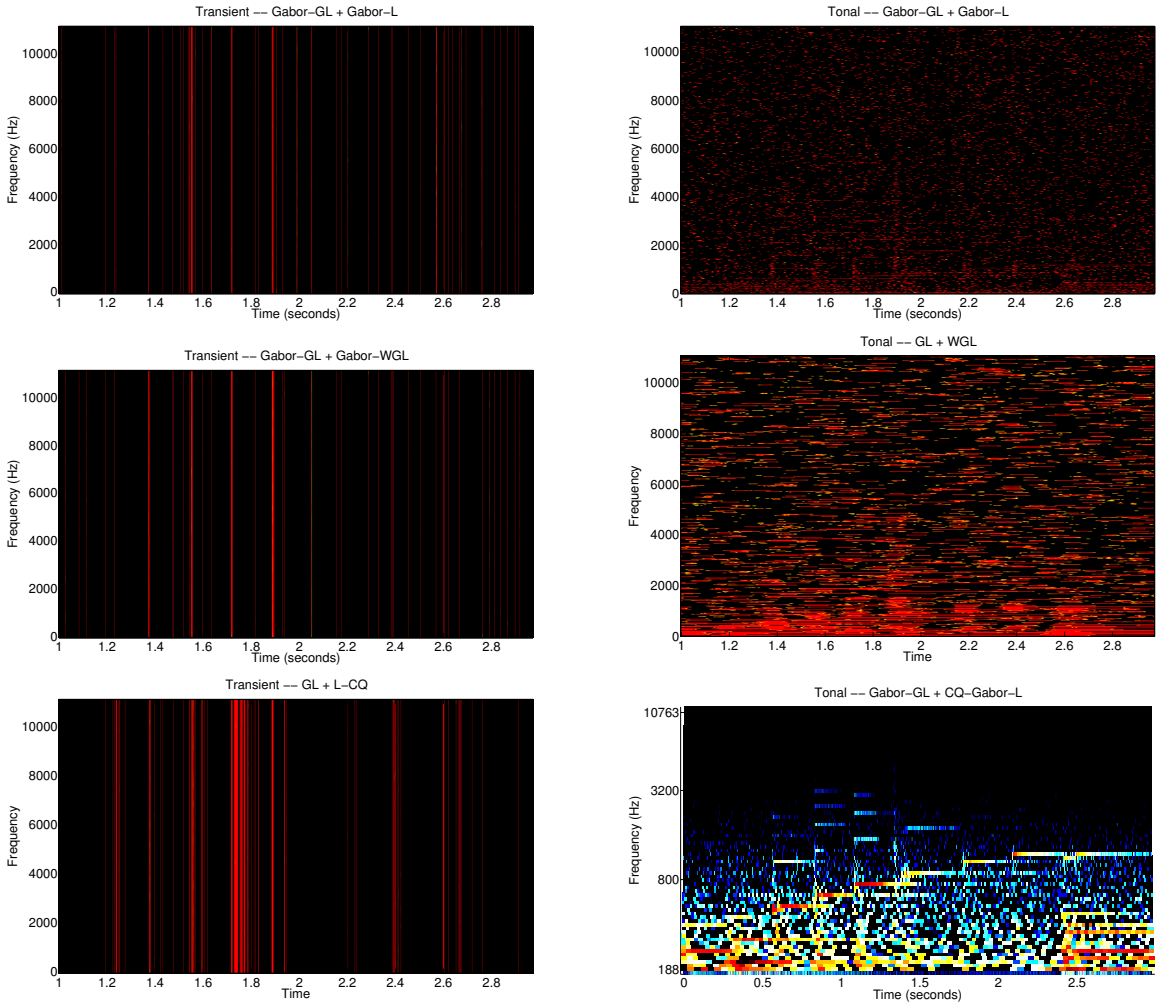

Fig. 3. Time-Frequency representations of each denoised layers (Left: transients – Right: Tonals), for the three methods. From top to bottom: Gabor-GL + Gabor-L, Gabor-GL + Gabor-WGL, Gabor-GL + CQ-Gabor-L

VI. Conclusions, extensions and open problems

Usual linear time-frequency analysis relies on what we called here the analysis coefficients of a Gabor frame. As a result, once the window is fixed, analysis leads to a unique set of coefficients associated to

the signal. In contrast, by adopting the synthesis point of view, more flexibility is introduced into the choice of the time-frequency coefficients, as models can easily be incorporated.

We have also shown that the use of frame theory allows one to introduce extra flexibility into time-frequency analysis, by extending standard Gabor and STFT representations to other waveform systems, which can be more relevant in a number of practical situations, as we exemplified above. Frame theory also has the merit to permit the development of efficient and stable algorithms for frame synthesis, which is essential for practical implementations.

Moreover, the synthesis model is flexible enough in order to model some natural structures on the coefficients. We choose to present here the variational approach, and more specifically convex models which provide efficient algorithms. However, we have seen that iterative thresholding algorithm can be easily used more heuristically with new thresholding operators such as the social sparse WG-Lasso. Of course, other approaches can be used, such as greedy algorithms [28] or Bayesian models [29].

Many extensions of what we have presented here can be imagined, building on the framework we have described. For example, a related concept of structured thresholding, that use a different approach is *perceptual sparsity*. There, those transform domain coefficients whose removal does not produce audible difference are simply erased. To find these perceptually irrelevant coefficients an algorithm based on psychoacoustic models is used [30]. Such a thresholding strategy could be embedded in an ISTA algorithm. Another straightforward extension is the use of social sparsity with NSGT, to provide a very flexible framework in order to have very structured synthesis coefficients. Also related to perception is the concept of ERBlets [9]. Here a NSGT is used on the Fourier side, where the filters are adapted to the ERB scale and bandwidth, which were measured through psychoacoustical measurements. As opposed to other approaches, by relying on frame theory, perfect reconstruction is always guaranteed.

Still, a lot of open problems remains. For applications the problems arise of how to integrate a-priori knowledge, which is often given only heuristically. For example, it would be very natural, to use a structured sparsity approach for the estimations of formants/anti-formants in a speech signal. The hard part is how to integrate the a-priori knowledge of vowels and consonants in such a method. Another related example is how can we learn automatically the onset for the NSGT without any preprocessing step. On a more theoretical side, the generalization of NSGT to adaptivity in two dimensions should be investigated.

As mentioned in the introduction, we have deliberately limited the developments of this paper to the finite-dimensional situation, i.e. the case of discrete, finite-length signals. However, most of the non-algorithmic developments carry over to the infinite-dimensional case (e.g. $L^2(\mathbb{R})$): abstract frame theory,

Gabor theory, and also the flexible Gabor constructions such as continuous time versions of NSGT. For instance, the problem of finding versions of NSGT that yield arbitrary tiling of the time-frequency plane is still open in the infinite-dimensional setting as well. It is also worth stressing that infinite-dimensional theories are of some practical relevance, when it comes to processing very long discrete signals, and the intuition provided by function spaces has often be very relevant when transposed to finite-dimensional situations.

ACKNOWLEDGMENTS

P. Balazs is supported by the Austrian Science Fund (FWF) START-project FLAME ('Frames and Linear Operators for Acoustical Modeling and Parameter Estimation'; Y 551-N13); M. Dörfler is supported by the WWTF project *Audiominer* (MA09-24); B. Torrésani is supported by the European project UNLocX, grant number 255931, and by the ANR project Metason ANR-10-CORD-010 ; M. Kowalski, benefited from the support of the "FMJH Program Gaspard Monge in optimization and operation research", and from the support to this program from EDF.

THE AUTHORS

Peter Balazs (peter.balazs@oeaw.ac.at) received the PhD (2005) and the habilitation (2011) degree at the University of Vienna in mathematics, in cooperation with the Numerical Harmonic Analysis Group (NuHAG). He has been a member of the Acoustics Research Institute (ARI) of the Austrian Academy of Science since 1999. He also worked in Marseille, France both at the Laboratoire d'Analyse Topologie et Probabilités (LATP), and the Laboratoire de Mécanique et d'Acoustique (LMA), from 2003 to 2006; and with the Unité de physique théorique et de physique mathématique (FYMA) from the Université Catholique de LouvainLouvain-La-Neuve in 2005. He is the founder of the working group "Mathematics and Acoustical Signal Processing" at ARI. He is the acting director of ARI since 2012. He is interested in time-frequency analysis, Gabor analysis, numerical analysis, frame theory, signal processing, acoustics and psychoacoustics.

Monika Dörfler (monika.doerfler@univie.ac.at) obtained her PhD in Mathematics from the University of Vienna and is a researcher at the Faculty of Mathematics. She is working in the field of applied mathematics for audio signal processing with a particular focus on the interplay of local and global aspects of time-frequency analysis.

Matthieu Kowalski (matthieu.kowalski@lss.supelec.fr) received the engineering degree in computer science from the Université de Technologie de Compiègne (UTC) in 2005, and the master degree in

Mathematics Vision and Learning from the Ecole Normale Supérieur, Cachan, the same year. He received the PhD degree in applied mathematics from the University of Provence in 2008. His thesis was axed on sparse time-frequency decompositions. He is now an assistant professor at the University of Paris-Sud, as a member of GPI research group in the L2S Lab. His research focuses on Inverse Problems and structured sparse approximations, with applications to audio signal and M/EEG.

Bruno Torrésani (bruno.torresani@univ-amu.fr) received the PhD degree in mathematical physics from Université d'Aix-Marseille I in 1986, and the habilitation degree from Université d'Aix-Marseille II in 1992. He was researcher at CNRS from 1988 to 1998 at Centre the Physique Théorique, Marseille, France, and is now professor in Mathematics at Université d'Aix-Marseille, and the head of LATP, the mathematics laboratory. He held associate professor positions at Université de Louvain la Neuve, Belgium, University of California at Irvine (USA), and Universidad de La Plata (Argentina). His research interests are mainly in mathematical signal processing, including applied harmonic analysis and functional analysis, probabilistic modeling and statistics, and applications to various domains such as audio signal processing, genomics and neurosciences.

REFERENCES

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

[2] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Prentice Hall, 1995.

[3] K. Gröchenig, *Foundations of Time-Frequency Analysis*, ser. Appl. Numer. Harmon. Anal. Birkhäuser Boston, 2001.

[4] M. Dolson, "The phase vocoder: a tutorial," *Comput. Music. J.*, vol. 10, no. 4, pp. 11–27, 1986.

[5] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions," *J. Math. Physics*, vol. 27, no. 5, pp. 1271–1283, 1986.

[6] P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus, and G. A. Velasco, "Theory, implementation and applications of nonstationary Gabor frames," *J. Comput. Appl. Math.*, vol. 236, no. 6, pp. 1481–1496, 2011.

[7] N. Holighaus, M. Dörfler, G. Velasco, and T. Grill, "A framework for invertible, real-time constant-Q transforms," *IEEE Trans. Audio, Speech Lang. Processing*, vol. 21, no. 4, pp. 775–785, 2013.

[8] J. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.

[9] T. Necciari, P. Balazs, N. Holighaus, and P. Soendergaard, "The ERBlet transform: An auditory-based time-frequency representation with perfect reconstruction," in *Proc. Int. Conf. Audio Speech and Signal Processing (ICASSP)*, 2013, p. accepted.

[10] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inform. Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[11] S. Molla and B. Torrésani, "A hybrid scheme for encoding audio signal using hidden Markov models of waveforms." *Appl. Comput. Harmon. Anal.*, vol. 18, no. 2, pp. 137–166, 2005.

[12] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech, and Lang. Processing*, vol. 16, no. 1, pp. 116–128, 2008.

[13] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization," *the Proceedings of the National Academy of Sciences*, vol. 100, pp. 2197–2202, March 2003.

[14] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, pp. 713–718, 1992.

[15] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with a certain family of non-convex penalties and DC programming," *IEEE Trans Signal Processing*, vol. 57, no. 12, pp. 4686–4698, 2009.

[16] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients," *Signal, Image Video Process.*, vol. 3, no. 3, pp. 251–264, 2008.

[17] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J Royal Stat Society-Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[18] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, 2011.

[19] I. Bayram, "Mixed norms with overlapping groups as signal priors," in *Proc. Int. Conf. Audio Speech and Signal Processing (ICASSP)*, May 2011.

[20] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proc. of International Conference on Machine Learning (ICML)*, 2009.

[21] M. Zibulevsky and M. Elad, "L1-L2 optimization in signal and image processing," *IEEE Signal Processing Mag.*, vol. 27, no. 3, pp. 76–88, 2010.

[22] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[23] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE Trans. Signal Processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

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

[25] A. Gramfort, D. Strohmeier, J. Haueisen, M. Hamalainen, and M. Kowalski, "Functional brain imaging with m/eeg using structured sparsity in time-frequency dictionaries," *Neuroimage*, vol. 70, pp. 410–422, 2013.

[26] T. S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, 1996.

[27] P. Soendergaard, B. Torrésani, and P. Balazs, "The linear time frequency analysis toolbox," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 10, no. 4, p. 1250032, 2012. [Online]. Available: http://ltfat.sourceforge.net

[28] R. Gribonval and M. Nielsen, "Beyond sparsity : recovering structured representations by ℓ^1-minimization and greedy algorithms." *Adv. Comp. Math.*, vol. 28, no. 1, pp. 23–41, 2008.

[29] C. Févotte and S. Godsill, "A bayesian approach to blind separation of sparse sources," *IEEE Trans. Audio, Speech Lang. Processing*, vol. 14, no. 6, pp. 2174–2188, 2006.

[30] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, "Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 18, no. 1, pp. 34–49, 2010.