Sparse and Structured Decompositions of Signals With the Molecular Matching Pursuit

Laurent Daudet, *Member, IEEE*

***Abstract*—This paper describes the Molecular Matching Pursuit (MMP), an extension of the popular Matching Pursuit (MP) algorithm for the decomposition of signals. The MMP is a practical solution which introduces the notion of structures within the framework of sparse overcomplete representations; these structures are based on the local dependency of significant time-frequency or time-scale atoms. We show that this algorithm is well adapted to the representation of real signals such as percussive audio signals. This is at the cost of a slight sub-optimality in terms of the rate of convergence for the approximation error, but the benefits are numerous, most notably a significant reduction in the computational cost, which facilitates the processing of long signals. Results show that this algorithm is very promising for high-quality adaptive coding of audio signals.**

***Index Terms*—Matching pursuit, overcomplete representations, parametric audio coding, time-frequency transforms.**

I. Introduction

FINDING sparse representations of signals has become a major area of research in the last few years (see [1] for a review on methods and recent results). Sparse representations are obviously useful for signal compression [2], [3] but are also relevant in the context of applications such as denoising [4], source separation [5], etc.

Generally, we look for decompositions of a signal $\mathbf{x}$ on a dictionary $\mathcal{D} = \{\mathbf{u_\lambda}\}_{\lambda \in \Lambda}$ of indexed elementary waveforms $\mathbf{u_\lambda}$, in the form

$$\mathbf{x} = \sum_{\lambda \in \Lambda} \alpha_\lambda \mathbf{u_\lambda}. \tag{1}$$

It is sometimes useful to use a dictionary $\mathcal{D}$ that is overcomplete, which means, in finite dimension N, that $\mathcal{D}$ spans the whole space and has more than N elements. In that case, the above decomposition (1) is not unique and one has to select a decomposition according to a sparsity criterion or some other appropriate metric. Indeed, in many applications, useful representations are the ones where most of the energy of the signal is concentrated into a small number N of coefficients, so that the signal can be approximated using only N terms (Nth order nonlinear approximation)

$$\mathbf{x} = \sum_{i=0}^{N-1} \alpha_{\lambda_i} \mathbf{u_{\lambda_i}} + \mathbf{R_N}. \tag{2}$$

Manuscript received July 30, 2004; revised June 14, 2005. This work was supported by the French Ministry of Research and Technology, under contract ACI "Jeunes Chercheuses et Jeunes Chercheurs" number JC 9034. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

The author is with the Laboratoire d'Acoustique Musicale, Université Pierre et Marie Curie (Paris 6), 75015 Paris, France (e-mail: daudet@lam.jussieu.fr).

Digital Object Identifier 10.1109/TSA.2005.858540

Accurate and compact nonlinear decompositions of the signal $\mathbf{x}$ can only be obtained if the elements $\mathbf{u_\lambda}$ of the dictionary have strong similarities with the signal, in which case the corresponding waveforms can be seen as elementary components of the signals, or "atoms". In general, overcompleteness (i.e., redundancy) in the dictionary is required in order to get truly sparse representations, since real orthonormal bases, such as discrete wavelets or lapped local cosines, cannot be shift-invariant [6], or because the signals have complex structures that cannot be represented using only one class of waveforms. Such dictionaries can be made e.g., by the use of redundant transforms (for instance wavelet packets or Gabor frames), by the union of orthogonal bases, or by parametrized waveforms such as damped sinusoids [7].

Many algorithms have appeared for such atomic decompositions of complex signals. Amongst these, an important class of algorithms is the so-called iterative "greedy" algorithms, such as the **Matching Pursuit (MP)** [8] (or variants such as the Orthogonal Matching Pursuit [9]–[11]). The basic principle of MP is as follows.

(1) Initialization: compute all the inner products $\alpha_\lambda = \langle \mathbf{x}, \mathbf{u_\lambda} \rangle$. Let $\mathbf{R_0} = \mathbf{x}$ and $i = 0$.

(2) Find maximum modulus amongst all inner products: $\lambda_i = \arg\max_\lambda |\alpha_\lambda|$

(3) Update the residual by subtracting the corresponding atom

$$\mathbf{R_{i+1}} = \mathbf{R_i} - \alpha_{\lambda_i} \mathbf{u_{\lambda_i}}.$$

(4) Update the inner products

$$\alpha_\lambda \leftarrow \langle \mathbf{R_{i+1}}, \mathbf{u_\lambda} \rangle.$$

(5) if $|\alpha_{\lambda_i}| < \varepsilon_{\text{stop}}$ then stop, otherwise $i \leftarrow i + 1$ and iterate to step (2).

When the algorithm has stopped (after N iterations), the signal x is estimated using (2). At every iteration, the algorithm is "optimal" in the sense that it selects the atom that minimizes the residual energy (hence the "greedy" nature of the search). Note that other criteria for best atom selection can be employed [12]. Also, it should be emphasized that this procedure is only optimal at every iteration and not globally : it is not true in general that MP after N iterations will provide the best N-term approximation of the signal [1], [10]. Finally, other stopping criterion can be used: for instance, in compression applications, the algorithm can be stopped when the available bit budget is reached, or when all remaining tones are masked by the previously detected components.

The main limitation of the above method is its intrinsic complexity. At every iteration there are two stages that may be computationally expensive: step (2) that looks for the maximum of the inner product, which can be lengthy if the dictionary is large; step (4) when one has to update the inner products of the residual with every element of the dictionary.

In order to reduce the complexity of these potential bottlenecks, in some cases it is necessary to consider "fast" schemes, that are generally somewhat less effective (in a rate-distortion sense) than a full MP. For instance, in order to accelerate the search step (2), the parameters can be organized hierarchically, and the search conducted only on first ones (in [13], a frequency chirp parameter is fixed to zero in the search step and estimated *a posteriori*). Also, a set of all local maxima can be stored and updated adaptively [14]. Similarly, "Weak Matching Pursuits" stop the search when it has found an atom that is nearly optimal, i.e., a λ_i such that $|\alpha_{\lambda_i}| \geq a \max_\lambda |\alpha_\lambda|$, where $a \in (0,1]$ is a fixed *weakness parameter* [1], and $\max_\lambda |\alpha_\lambda|$ is assumed nearly constant for a few successive iterations before being recomputed explicitly. For a fast update of the inner products [step (4)], one can in some cases pre-compute all cross-products of the basis functions, but for a size-N dictionary this requires memory calls to potentially very large $N \times N$ look-up tables. Such implementation schemes have made it possible to use MPs for the analysis and representation of "real" data (i.e., large), and significant advances have been made in the context of still image coding (see for instance contributions by Vandergheynst and collaborators—e.g., [15]). However, even in such cases typical computational requirements are still high, and so far none of the above methods could realistically be applied globally to very large datasets, such as audio files (at CD-quality 44.1-kHz sampling rate, a mono 5-min song has 1.3×10^7 samples). For such long signals, existing practical solutions use only local searches, on a frame-by-frame basis [16]–[18]. The work presented here shows that, under certain assumptions, it is possible to apply MP globally on the whole data (at least on a duration that is longer than the typical note duration, i.e., a few seconds), and to extract components whose time duration range from a few milliseconds to a few seconds.

In this paper, we consider another class of "fast" suboptimal MPs, where the above suboptimality is compensated by a better description of the structure of the signal. In MP decompositions, the localization of the selected atoms in the time-frequency/time-scale planes is not uniform but reveals some of the intrinsic structure of the analyzed signal. The goal of this paper is to make use of this structural information, by grouping together atoms of the same class (i.e., belonging to the same orthonormal basis) with neighboring time-frequency/time-scale parameters. These groups of atoms will be referred to as "molecules". Now, we want to design a "Molecular Matching Pursuit" (MMP), where at any given iteration i one full molecule M_i of m_i significant atoms is estimated and subtracted from the residual. From a signal compression perspective, grouping significant coefficients into simple structures can offer a significant advantage in terms of coding cost, as encoding the shape of a structure usually requires less information than individually encoding the indices of the coefficients contained in it [19], [20] (in the same way as run-length encoding binary images usually results in much less data than entropy coding independent sample values). From a signal analysis point of view, these molecules provide relevant information about the structure of the signal. Finally, from a computational complexity perspective, there is also a significant advantage in considering such molecules: at each iteration, m_i atoms are picked up at once, and therefore the inner product update [step (4)] is performed jointly for m_i atoms. The difficulty lies in that, if not designed carefully, such improvements could be at a cost of a very large increase in the complexity of the search step (2).

For the rest of this paper, we will focus on audio (musical) signals, as these are a good example of well-structured data. Audio signals can be well modeled as a sum of three components [21]: the tonal part (sum of sinusoids with slowly varying amplitude and frequency), transients (well-localized at the attack of notes) and a residual (modeled as locally stationary filtered white noise). This relatively simple structure (although covering a wide range of signals) allows us to work with a small degree of redundancy in our dictionary. Here, we take a 2-times redundant dictionary $\mathcal{D} = \mathcal{C} \cup \mathcal{W}$, where $\mathcal{C} = \{\mathbf{c_i}\}_{i=1...N}$ is an orthogonal basis of lapped cosines (also called a modified discrete cosine transform (MDCT) basis), and $\mathcal{W} = \{\mathbf{w_j}\}_{j=1...N}$ is an orthogonal basis of discrete wavelets [or discrete wavelet transform (DWT)]. Note that this hybrid model is additive, i.e., the tonal component (represented by MDCT atoms) and the transient component (represented by DWT atoms) can coexist at the same time, as opposed to some audio coding algorithms [22] that implement a switch between a MDCT basis and a DWT basis around attacks.

Although the representations described here are specifically tailored to sounds, similar methods can also be applied to images [23], with two-dimensional (2-D) dictionaries made of different orthogonal bases to represent edges, slowly varying areas and textures. More generally, the MMP is a relevant technique when the coefficients display strong structural information simultaneously in a small number of bases that are mutually weakly coherent, and where these structures correspond to different features of the signal.

For audio signals, the idea of selecting at each iteration a group of atoms in a MP framework has already been developed in the so-called harmonic matching pursuit [14], which looks for harmonic structures made of (quasi) harmonically related Gabor atoms. This approach gives good results for the analysis of harmonic data, but its range of applicability is limited by the following three factors: first, it requires the estimation of a large set of (potentially continuous) parameters, making it unsuitable for compression purposes (the cost of coding the parameters, i.e., the significance map, becomes prohibitive). Second, the best analysis results require a significant amount of information on the data, such as the number of partials or the time envelope near the attack, that must be known or assumed *a priori* or estimated by some other means. Third, a large fraction of musical sounds that are difficult to represent through standard (local Fourier-based) analysis are not harmonic (e.g., percussive sounds). For real data, a given harmonic partial cannot simply be described by a single Gabor atom: it has frequency and amplitude modulations, it may have a sharp or a very slow attack transient; and therefore a given note requires a potentially large

number of (independent) harmonic atoms. One of the claims of our paper is that the *local* time-frequency/time-scale grouping is a stronger and more robust assumption about the structure of real audio signals than the harmonicity. We shall also see that this type of structure has additional benefits, such as the ability to control pre-echoes and to interpolate sinusoids in the MDCT domain.

Finally, one may see the main goal of this article as to give practical solutions for unifying two recent paradigms that have emerged separately in the signal processing community within the last few years: sparse representations in overcomplete spaces (see, for instance, [1], [7], [8], [15]), and structured representations (such as EZW [19] or SPIHT [20] in image coding). So far, algorithms for sparse signal representations in overcomplete spaces have not considered dependencies between significant atoms; and algorithms that have considered structures in the significance map only work within orthonormal bases. Unifying these two concepts is possible here because we consider dictionaries made by concatenation of a small number of orthonormal bases that are sufficiently incoherent, i.e., with sufficiently different time-frequency localization properties. This is a major difference with the harmonic matching pursuit, where the dictionary is made of a large number of very coherent atoms that make it more difficult to consider *local* time-frequency/time-scale structures. In that case, it is not possible to compute simply the energy of a group of atoms as the sum of the energy of the individual atoms within that group, since neighboring atoms are not (even approximately) orthogonal.

The paper is organized as follows: after a presentation of the two types of "molecules" (Section II), we describe the plain MMP decomposition algorithm (Section III). Results are presented in Section IV, and finally, refinements of the decomposition are found in Section V. The conclusion (Section VI) discusses the generality of the method and future directions for research.

II. Molecules as Coherent Sets of Atoms

In this section, we describe what we call "molecules" of time-frequency or time-scale atoms. We restrict ourselves to redundant spaces $\mathcal{D}$ that are the concatenation of orthogonal bases, for instance $\mathcal{D} = \mathcal{C} \cup \mathcal{W}$, where $\mathcal{C}$ denotes the basis of MDCT atoms and $\mathcal{W}$ the basis of the DWT atoms. In order to eliminate the redundancy *within one molecule*, we will consider only molecules that are formed of one class of atoms: "tonal molecules" that are clusters of MDCT atoms, or "transient molecules" that are clusters of DWT atoms.

A. Tonal Molecules: Clusters of MDCT Atoms

Molecules of MDCT atoms are used to represent the tonal part of the signals. The MDCT [24] is an orthogonal local cosine transform with smooth windows that satisfy a perfect reconstruction condition. Typically, one uses sinusoidal windows given by $g_p[n] = \sin[(\pi/2L)(n - pL + (L/2))]$, where L is the window half-length and the stride, i.e., the hop size between two analysis frames. The MDCT expansion of the signal $\mathbf{x}$ is given by

$$\beta_{p,k} = \langle \mathbf{g}_{p,k}, \mathbf{x} \rangle \tag{3}$$

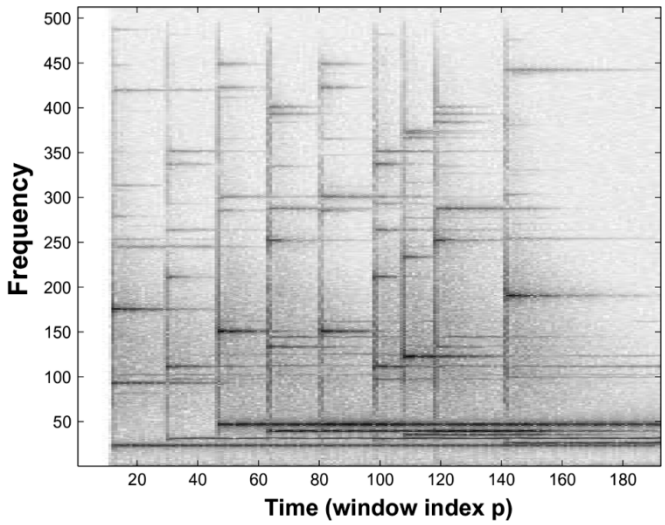

Fig. 1. Time-frequency representation of the MDCT of a 2.2-s recording of a glockenspiel. Note the narrow horizontal structures corresponding to the partials.

where

$$\mathbf{g}_{p,k}[n] = g_p[n]\sqrt{\frac{2}{L}}\cos\left[\frac{\pi}{L}\left(k+\frac{1}{2}\right)(n+n_p)\right] \tag{4}$$

with $k = 0 \ldots L-1$ and $n_p = (L+1)/2 - pL$. Note that, due to the overlap between frames, for a given time frame p the set $\{\mathbf{g}_{p,k}\}_{k=0\ldots L-1}$ does not constitute a basis, and this transform can only be seen as an orthonormal transform globally on the whole signal, with proper boundary conditions.

In musical signals, the tonal part is made of so-called "partials", which can be described as sinusoids with slowly varying parameters (amplitude and frequency). For the sake of simplicity, we will restrict our model to signals where the instantaneous frequencies of the partials remain constant or approximately constant (fluctuations within one frequency bin), although our system can readily be expanded to slowly varying frequencies.

The MDCT coefficients of a stationary sinusoid, with frequency ω_0, are distributed around the center frequency bin $k_0 = \lfloor \omega_0 L/\pi \rfloor$ with a relatively slow decay (in $1/k^2$). Here, we define tonal molecules as horizontal structures in the MDCT time-frequency plane (well identified on Fig. 1). Since the minimum number of MDCT coefficients required to effectively represent a stationary sinusoid [25] is three per window (these can be seen as the equivalent of frequency, phase, and amplitude), we will define the tonal molecules as "tubes" with a width in frequency equal to three bins $(k_0 - 1, k_0, k_0 + 1)$, and with an arbitrary time duration (see Fig. 3).

B. Transient Molecules: Dyadic Trees of Discrete Wavelet Coefficients

Similarly, one can construct "transient" molecules by grouping together wavelet coefficients. A basis $\mathcal{W} = \{\mathbf{w}_{b,a}\}$ of DWT atoms is organized according to two coefficients: a scale coefficient b ($b = 1 \ldots J$, where J denotes the largest scale) and a time coefficient a. With a proper shift at the origin, a given wavelet $\mathbf{w}_{b,a}$ is localized around the time $t = 2^b a$. Therefore, discrete wavelets are organized as dyadic trees [19],

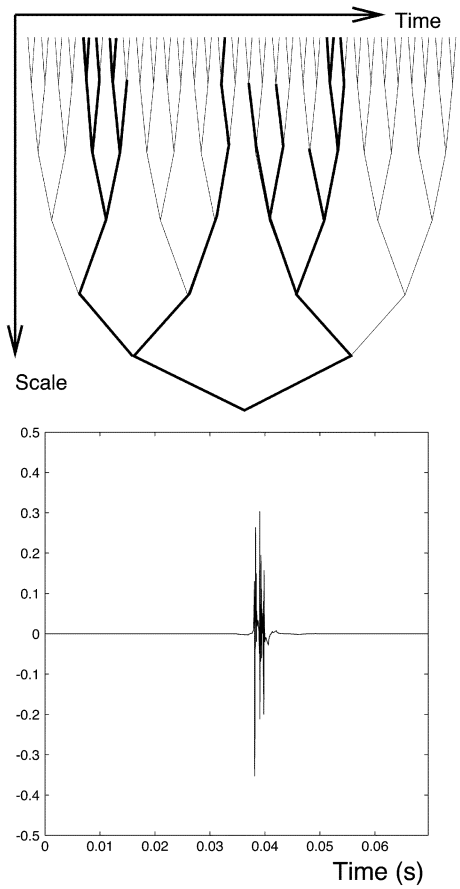

Fig. 2. Top: discrete wavelet coefficients are organized as dyadic trees. The subtree in thick lines represents a transient molecule in this grid. Bottom: corresponding time-domain waveform.

[20], where every atom $\mathbf{w}_{b,a}$ $(b = 2 \dots J)$ has two "children" $\mathbf{w}_{b-1,2a}$ and $\mathbf{w}_{b-1,2a+1}$. Let $\alpha = \{\alpha_{b,a}\}$ be the corresponding wavelet expansion of our signal

$$\alpha_{b,a} = \langle \mathbf{w}_{b,a}, \mathbf{x} \rangle. \tag{5}$$

The molecules of wavelets will correspond to clusters of significant coefficients $\alpha_{b,a}$ sharing neighboring time localization, while still forming a connected (incomplete) tree [26]. By their very precise time localization (we use generating wavelets with a short support, such as Haar or Daubechies-4 wavelets), wavelet trees are well adapted to the description of transients [27]. Fig. 2 shows the dyadic grid (thin lines) with an example of a transient molecule (thick lines), and the corresponding waveform.

III. Molecular Matching Pursuit

With the above definitions, we would like to design a molecular MP decomposition algorithm that, at every iteration, identifies, and subtracts the most significant molecule. Unfortunately, performing an exhaustive search amongst all possible molecules is not realistic [10]. For instance, in the general case where we allow the frequencies of tonal molecules to vary in time, their number is exponentially growing with m_i, the number of atoms in a molecule; note that we want to be able to define molecules that may contain a rather large number of atoms—typically between 5 and 100. Similarly, the number of subtrees of a dyadic tree is rapidly growing with the number J of wavelet scales (we typically use $J = 8$). Therefore, the search for the maximum is performed over values of two indices of correlation, one called the *local tonality index* in the MDCT domain for the tonal part, one called the *regularity modulus* in the DWT domain for the transient part. The corresponding molecules are then constructed around the location of the maximum index of correlation.

A. Local Tonality Index in the MDCT Domain

Here, one has to design a scheme for correlations across time for MDCT spectra. However, the MDCT is not invariant through time shifts, and therefore a direct correlation may give rise to an estimation of the peak frequency bin that is not robust to time shifts. Instead, one can design a pseudo shift-invariant representation out of the set of MDCT coefficients $\beta_{p,k}$, called the MDCT *pseudo-spectrum* $\mathcal{S}$, defined as follows [6]:

$$\mathcal{S}_{p,k} = \left(\beta_{p,k}^2 + (\beta_{p,k+1} - \beta_{p,k-1})^2\right)^{\frac{1}{2}}, \quad \text{for } k = 0 \dots (L-1) \tag{6}$$

where we define $\beta_{p,-1} = \beta_{p,L} = 0$. It has been shown that the pseudo-spectrum of a pure sinusoid will always be maximum in the same frequency bin k_0, for any value of its phase.

At each point (p, k) of the time-frequency plane, we associate a *local tonality index* $\mathcal{T}$ by looking at local averages (in time) of the pseudo-spectrum $\mathcal{S}$

$$\mathcal{T}_{p,k} = \frac{1}{W} \sum_{i=0}^{W-1} \mathcal{S}_{p+i,k} \tag{7}$$

where W represents a time persistence constant.

B. Regularity Modulus in the DWT Domain

Similarly, it is possible to measure the correlations across scales of the wavelet coefficients. From the set $\{\alpha_{b,a}\}$ of DWT coefficients, the following *modulus of regularity* κ has been introduced in [26] as

$$\kappa[2t] = \frac{1}{J} \sum_{(b,a) \in \mathcal{B}_t} |\alpha_{b,a}| \tag{8}$$

where J is the number of wavelet scales and $\mathcal{B}_t$ is the set of ancestors (full branch) of the smallest-scale atom $\mathbf{w_{1,t}}$. Note that κ is defined for every second time sample, and this makes it a very *local* measure of the strength of singularities in the signal.

C. MMP Decomposition Algorithm

With the above definitions, molecular decomposition of audio signals can be implemented using a modified version of the MP algorithm. At every iteration, we now look for the highest value of the indices of correlation (local tonality index $\mathcal{T}$ and regularity modulus κ), identify the corresponding molecule and subtract it from the residual. More precisely, the basic MMP

is as follows (compare with the standard MP described in the introduction):

(1) Initialization: compute all the inner products, i.e., the DWT coefficients $\alpha_{b,a}$ and the MDCT coefficients $\beta_{p,k}$ of the signal $\mathbf{x}$. Let $\mathbf{R_0} = \mathbf{x}$ and $i = 0$.

(2) Compute modulus of regularity κ and local tonality index $\mathcal{T}$, find $K = \max \kappa$ and $T = \max \mathcal{T}$.

(3) Identify the most significant structure. If $T \geq K$ then the most significant structure is of type "tonal molecule"; otherwise ($K > T$) it is of type "transient molecule".

(4) Identify atoms the define the most significant molecule, update the residual, and update the pursuit correlations.

- For tonal molecules, identify the set of weighted MDCT atoms that define the corresponding molecule $M_i = \{\beta_\mu \mathbf{g_\mu}\}_{\mu=1\ldots m_i}$, as described in Section III-D.

Update the residual by subtracting the corresponding atoms:

$\mathbf{R_{i+1}} = \mathbf{R_i} - \sum_{\mu=1\ldots m_i} \beta_\mu \mathbf{g_\mu}$.

Update MDCT coefficients by setting to zero all $\{\beta_\mu\}_{\mu=1\ldots m_i}$, update DWT coefficients by direct recalculation: $\alpha_{b,a} = \langle \mathbf{w}_{b,a}, \mathbf{R_{i+1}} \rangle$.

- For transient molecules, identify the set of weighted DWT atoms that define the corresponding molecule $M_i = \{\alpha_\nu \mathbf{w_\nu}\}_{\nu=1\ldots m_i}$, as described in Section III-E.

Update the residual by subtracting the corresponding atoms:

$\mathbf{R_{i+1}} = \mathbf{R_i} - \sum_{\nu=1\ldots m_i} \alpha_\nu \mathbf{w_\nu}$.

Update DWT coefficients by setting to zero all $\{\alpha_\nu\}_{\nu=1\ldots m_i}$, update MDCT coefficients by direct recalculation: $\beta_{p,k} = \langle \mathbf{g}_{p,k}, \mathbf{R_{i+1}} \rangle$.

(5) if $\max(K,T) < \varepsilon_{\text{stop}}$ then stop, otherwise $i \leftarrow i+1$ and iterate to step (2).

With regards to computation, note that in the implementation of MMP the search step (2) is fast ($\mathcal{O}(N)$), since it is only performed over N time-frequency (MDCT) parameters, and $N/2$ time-scale (DWT) parameters (only at the smallest scale). Similarly, the update (4) of the inner products is fast ($\mathcal{O}(N)$), since only half of them have to be fully recomputed, the other half being updated by a simple difference. The update is performed by computing the new residual $\mathbf{R_{i+1}}$ and then by using fast transforms to compute the new correlations. If a tonal molecule has been selected, the DWT update is $\mathcal{O}(N)$; if a transient molecule has been selected, the MDCT update is $\mathcal{O}(L \log L)$ on each of the N/L windows, therefore $\mathcal{O}(N \log L)$. Note that, for small molecules it may be faster to use a direct update of the coefficients via $\alpha_\nu \leftarrow \alpha_\nu - \sum_{\mu=1}^{m_i} \beta_\mu \langle \mathbf{w_\nu}, \mathbf{g_\mu} \rangle$, which is typically the method used by standard MP [8], and whose complexity scales as $\mathcal{O}(m_i^2)$. For the sake of simplicity, and since we are only interested in the first iterations that are likely to provide large molecules (see Section IV-A), we will not consider this possible improvement.

It should be noted as well that the MMP can incorporate, in a straightforward way, standard modifications of the MP, such as the orthogonalization of the selected atoms [11], or (faster) weak searches [1]. The major difficulty, and arguably the most *ad-hoc* part of the algorithm, lies in step (3), i.e., the identification of the significant molecule, once the maximum index of correlation has been computed. This is the topic of the next two paragraphs.

D. Identification of the Tonal Molecules

This paragraph describes how the tonal molecule is estimated, when at a given iteration i the most significant structure has been estimated as being of type "tonal".

Let p_0 and k_0 be the time (window index) and frequency indices, respectively, corresponding to the maximum local tonality index $\mathcal{T}_{p_0,k_0} = \max_{p,k} \mathcal{T}_{p,k}$. As stated in Section II.A, tonal molecules are by definition tubes with a width equal to 3 frequency bins. The beginning and end windows are determined when looking at the profile of $\mathcal{S}_{p,k_0}$ (for the frequency bin k_0), around $p = p_0$ (see Fig. 3). When looking iteratively forward in time, starting from the time index $p = p_0$, the end window p_{end} is defined as the last window before a sudden drop in the value of $\mathcal{S}_{p,k_0}$ (a ratio threshold of $\mathcal{S}_{p,k_0}/\mathcal{S}_{p+1,k_0} < 3$ is a typical choice), or the last window before $\mathcal{S}_{p,k_0}$ gets below the threshold $\varepsilon_{\text{stop}}$, whichever comes first. Similarly, the beginning window p_{start} is defined, when looking iteratively *backward* in time, starting from the time index $p = p_0$, as the last window before a sudden drop in the value of $\mathcal{S}_{p,k_0}$ (same ratio threshold), or the last window before $\mathcal{S}_{p,k_0}$ gets below the threshold $\varepsilon_{\text{stop}}$, whichever comes first. Note that this procedure has some similarities with the one employed in [14] for the harmonic matching pursuit, with the major difference here being within an orthogonal subset of the dictionary $\mathcal{D}$.

The last step is a final post-thresholding: the tonal molecule M_i is defined as the set of tonal MDCT atoms, within the width-3 tube starting at p_{start} and ending at p_{end}, that are above a given threshold $\varepsilon_{\text{coef}}$

$$M_i = \{\beta_{p,k}; |\beta_{p,k}| > \varepsilon_{\text{coef}}\}_{p=p_{\text{start}}\ldots p_{\text{end}}, k=k_0-1\ldots k_0+1}. \quad (9)$$

This threshold $\varepsilon_{\text{coef}}$ should be chosen adequately, since too large of a value may lead us to neglect all atoms in a given time window, and therefore to construct an apparently broken molecule. A good choice for $\varepsilon_{\text{coef}}$ is therefore $\varepsilon_{\text{coef}} = \varepsilon_{\text{stop}}/2$.

E. Identification of the Transient Molecules

In the case when, at a given iteration i, the most significant structure has been estimated as being of type "transient", the corresponding molecule M_i is constructed in two steps: first, from the wavelet expansion one constructs a *full tree*, i.e., a wavelet tree with branches going from the largest ($b = J$) to the smallest scale ($b = 1$). Let t_0 be the time index of the maximal regularity modulus, and $K = \kappa[2t_0]$. We define the full tree as the set of full branches, sharing the same root as $\mathbf{w_{1,t_0}}$, whose value of κ is above a threshold $\tilde{\kappa} = K/\zeta$ (a typical choice is $\tilde{\kappa} = K/3$). Second, incomplete subbranches are pruned out. This is done in a top-down approach: starting at the smallest scale (leaves of the branches), one prunes out coefficients that are smaller in absolute value than $\varepsilon_{\text{coef}}$. On a given

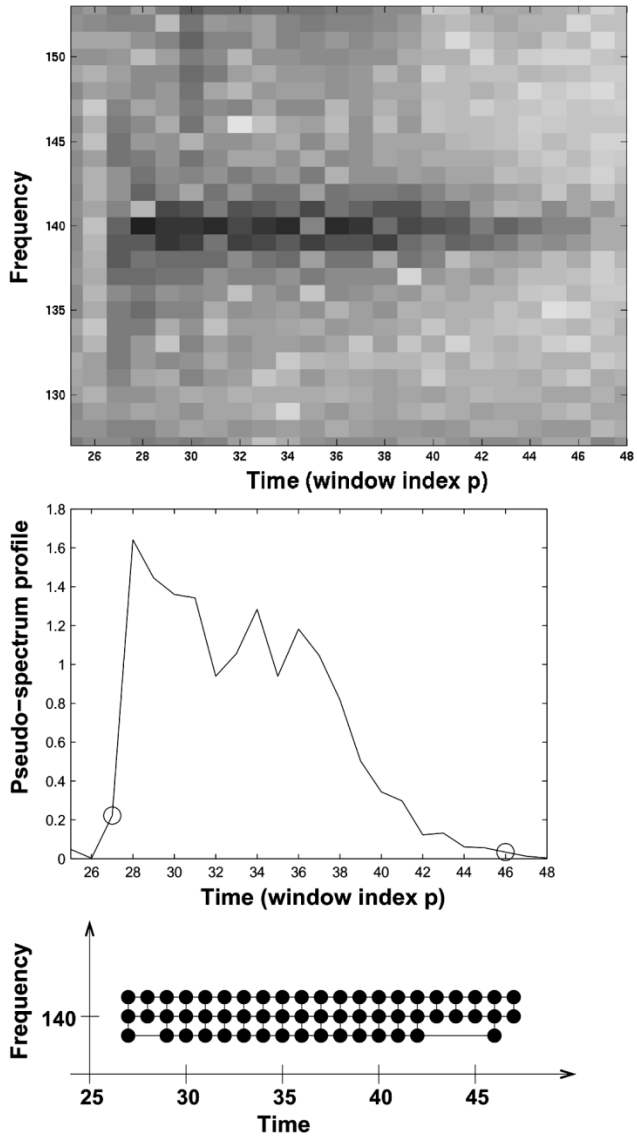

Fig. 3. Construction of a tonal molecule. Top: neighborhood of a tonal molecule in the MDCT domain. Middle: profile of the pseudo-spectrum $\mathcal{S}_{p,k_0}$ for a given molecule. Circles indicate the beginning and end windows. Bottom: corresponding tonal molecule. Empty slots indicate MDCT atoms that are below an established coefficient threshold.

branch, this pruning is stopped whenever a significant coefficient is found (larger than $\varepsilon_{\text{coef}}$), in order to ensure that the tree remains connected.

IV. Results

A. Complexity Reduction

In terms of computational complexity, this algorithm is significantly faster than the standard MP, as one only has to update the inner products with the dictionary once at every iteration, where a whole structure of m_i atoms is subtracted from the signal. Fig. 4 shows the number m_i of significant atoms for the first 200 iterations of the MMP. For this example, the average number is $\langle m_i \rangle = 14.6$, which gives an indication of the expected speedup ratio for the update step (4). Notice that large peaks (m_i up to 134) are observed; these are related to very long partials. As the number of iterations progresses, these

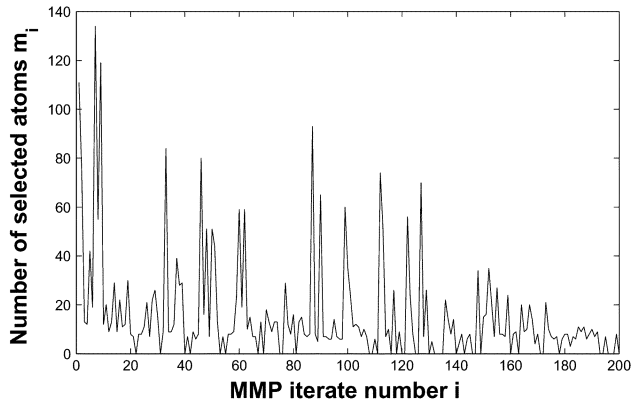

Fig. 4. Number m_i of atoms in the molecule M_i extracted by the MMP algorithm at each iteration.

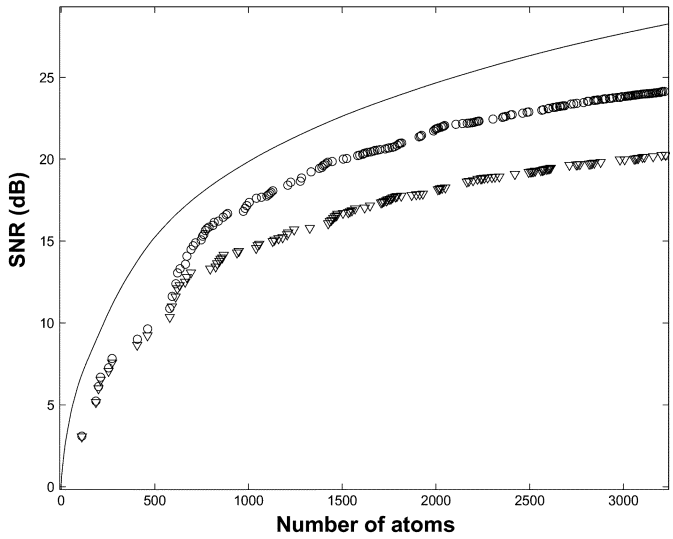

Fig. 5. Signal-to-noise ratio of three MP algorithms, with the same dictionary and on the same glockenspiel signal. Plain Line: standard MP; Triangular marks: plain MMP; circular marks : MMP with pre-echo control and frequency interpolation.

large peaks tend to disappear, which means that after a number of iterations all the "structured" information has already been extracted, and after that point MMP should offer no advantage over the standard MP. If one stops the algorithm at that stage (about 200 iterations in our example), a nonoptimized Matlab implementation of our algorithm runs in about $30\times$ real-time on a standard PC. We expect that an optimized C implementation would be substantially faster and therefore suitable for processing large sound files.

B. Convergence Rate

Let us now compare results in terms of convergence rate of the N-terms approximation for the standard MP and its structured version, the MMP. The plain line in Fig. 5 plots the signal-to-noise ratio (SNR) of the standard MP algorithm, for the glockenspiel signal. The SNR is calculated as $\text{SNR} = 20\log_{10}(\|x\|_2/\|\tilde{x}_N - x\|_2)$, where $\tilde{x}_N$ is the nonlinear approximation of x with the first N atoms selected by the MP. Similarly, SNR values are plotted for the approximations given by the MMP (triangular marks). It is only defined for values of N that correspond to the cumulative number of atoms in

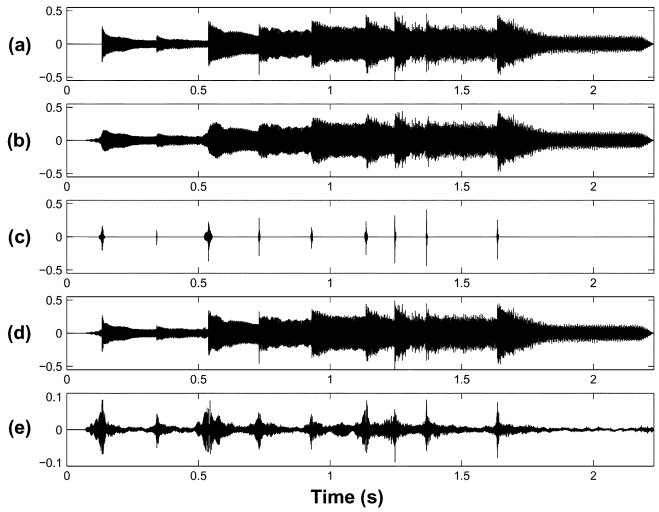

Fig. 6. Waveform separation results of the plain MMP after 292 iterations. (a) Original glockenspiel signal. (b) Tonal part. (c) Transient part. (d) Reconstructed signal ($\mathrm{tonal} + \mathrm{transients}$). (e) Residual (note the different scale).

the signal model after each iteration, namely $\sum_{i=1\ldots i_0} m_i$. In the first few iterations, only tonal molecules are selected, and the elbow around $N = 600$ corresponds to the appearance of transient molecules.

C. Separation Results

Finally, Fig. 6 shows the separation of the signal into the three signal model layers: the tonal part, the transient part, and a small amplitude wide-band residual; corresponding sound files are available on-line [28]. It can be noted that, as expected, the transients are only found at the attack of notes. The reconstructed signal (d) sounds quite similar to the original, although slightly poorer in high-frequency components, and with a noticeable pre-echo artifact. Reduction of this artifact will be discussed in Section V-A.

V. When One Trusts the Signal Model: Additional Benefits

Using the MMP algorithm has additional benefits over the standard MP, as it allows a better enforcement of the signal model. In our three-layer audio signal model, this can be done in two regards: a reduction of the pre-echo that occurs at the beginning of notes, and a frequency-domain interpolation around the spectral lines where the tone is apparently constant.

A. Pre-Echo Reduction

Once a tonal molecule has been selected, it is assumed that this corresponds to one partial of a sound. However, the need for good frequency resolution imposes the choice of relatively long windows (a typical choice for the window half-length is $L = 2048$ samples). With such long windows the well-known "pre-echo" phenomenon occurs at the beginning of sounds that have sharp attacks: in the reconstructed signal, the energy will be smeared out before the actual onset of the note [see the first note onset on the error plot of Fig. 6(e)], and this can be perceptually a very noticeable artifact.

In the MMP algorithm, the pre-echo artifact can be suppressed by adding a negligible amount of side information,

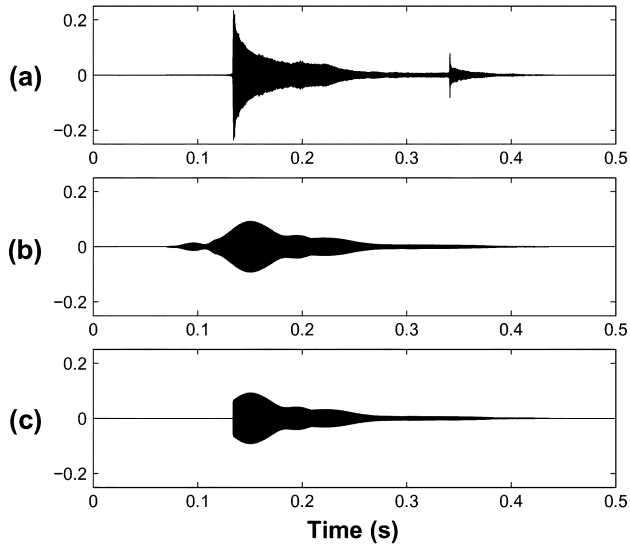

Fig. 7. Pre-echo control mechanism. (a) Original molecule (extracted from a large neighborhood $\mathcal{V}_i$ in the MDCT domain). (b) Reconstructed molecule with atoms from the molecule $\mathcal{T}_i$. (c) Same as (b) but with pre-echo control.

namely the *partial onset time* for tonal molecules. This onset time τ_i for a given tonal molecule $\mathcal{T}_i$ is estimated by minimizing the quadratic error between the actual waveform for this partial only (reconstructed using a large neighborhood $\mathcal{V}_i$ in the MDCT domain around the molecule) and the waveform reconstructed with only the molecule $\mathcal{T}_i$, with the beginning chopped off (see Fig. 7)

$$\tau_i = \arg\min_t \left\| \sum_{\mu \in \mathcal{V}_i} \beta_\mu \mathbf{g}_{\boldsymbol{\mu}} - H_t \sum_{\mu \in \mathcal{T}_m} \beta_\mu \mathbf{g}_{\boldsymbol{\mu}} \right\|_2 \tag{10}$$

where H_t is the t-centered Heavyside function $H_t[n] = 0$ if $n < t, H_t[n] = 1$ otherwise. The domain $\mathcal{V}_i$ uses typically $k_0 \pm 20$ frequency bins, that is a reasonable tradeoff for a good reconstruction of the attack transient without too much influence of other partials.

B. Frequency-Domain Interpolation

As stated in Section II-A, three MDCT coefficients are needed in each window to represent a stationary tone. Reciprocally, given three MDCT coefficients $\beta_{p,k_0-1}, \beta_{p,k_0}$, and β_{p,k_0+1} one can estimate [25] the frequency, amplitude and phase of the corresponding tone. Then, with the assumption that the signal is stationary, it is possible to estimate the contribution of this tone to the MDCT coefficients locally *around* the molecule. Given these parameters, and with the use of a sinusoidal window for the MDCT, one can easily compute analytically [6] the coefficients corresponding to this tone in the whole time-frequency plane. This contribution is not negligible for strong partials since the $\mathcal{O}(1/k^2)$ decay in amplitude around k_0 is relatively slow. Therefore, at no extra cost in the amount of data, it is possible to estimate and subtract the whole contribution of a stationary tone, and not only the center of the main lobe, i.e., the three coefficients per window that are retained.

Fig. 8 shows the separation results for the MMP, when the above two modifications are added. It is clear in Fig. 8(b) and

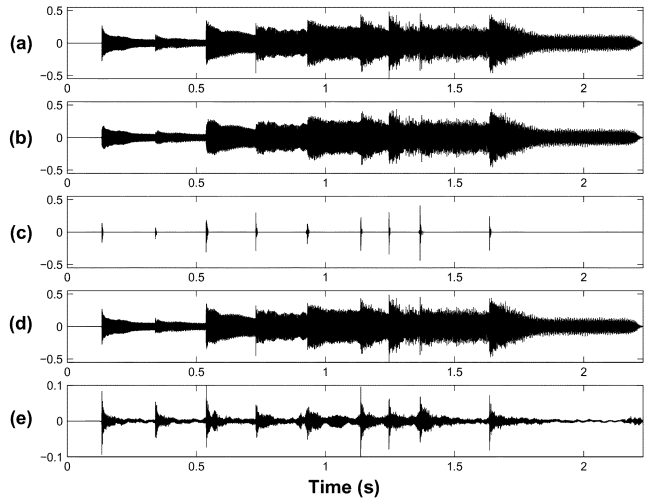

Fig. 8. Waveform separation results of the MMP after 221 iterations, with additional pre-echo control and frequency interpolation. (a) Original glockenspiel signal. (b) Tonal part. (c) Transient part. (d) Reconstructed signal $(\mathrm{tonal} + \mathrm{transients})$. (e) Residual (note the different scale).

(d) that the pre-echo has entirely disappeared, and this results in a significant improvement in the sound quality. Again, corresponding sound files are available on-line [28]. Finally, the circular marks in Fig. 5 indicate that these improvements make it possible to reach a performance that is roughly half-way closer to the "optimal" (within the MP framework).

It should be noted that for signals that do not match the model; for instance' for quickly varying tones or soft attacks, the above modifications do not offer any advantage over the standard MMP, nor do they reduce the performance. Finally, these improvements may require a significant increase in the complexity, since they add loops in the molecule identification process (step 3); and one has to find a tradeoff with the benefits in terms of convergence rate.

VI. Conclusion

In this paper, we have presented the MMP, a modification of the well-known MP algorithm, where we take into account the relationship of the significant atoms to the local structure of the signal. With a dictionary made by concatenation of a small number of orthogonal bases, it is possible to design a practical decomposition algorithm that at every iteration identifies and removes a whole cluster of (orthogonal) atoms. At the cost of a slight suboptimality in the approximation error rate, this offers a number of advantages, most notably it is significantly faster since the inner products update step is made for a large number of atoms at every iteration.

The most promising application of this algorithm is in high-quality coding of audio signals. Indeed, using structure information allows a significant reduction in the coding cost of the significance map (i.e., the set of parameters for the significant coefficients). Coding a tonal molecule only requires coding the position of the first atom and the time duration of the molecule; coding an incomplete wavelet tree, i.e., a transient molecule, requires coding the position of its root plus a maximum of 2 bits per retained coefficient [27], [20]. Also, for multichannel sounds, it is likely that a large portion of the structure information can be shared between channels, resulting in an increased performance. Furthermore, this decomposition is intrinsically progressive, and therefore could be used for scalable coding. Actually, it is conjectured that, for a large class of sounds, an N-atom partial reconstruction of the signal would sound better in the MMP case than in the MP case, in a similar way as, for a partial reconstruction of images, it may be in some cases that one prefers fewer fully-defined objects than all the objects at low resolution. Indeed, the MP/MMP may be a relevant framework to conduct such perceptual studies. Additionally, for coding purposes, the separation between the three classes has some extra benefits: first, it provides information about the audio signal that is highly relevant for the analysis, and this can be used, e.g., for indexing purposes; as opposed to individual atoms, molecules of *coherent* atoms are perceptually relevant, since they represent identifiable features of the signal. Second, it allows a different psychoacoustic model (for the quantization) in each layer, for instance the standard frequency-domain masking for the tonal layer, and a model for temporal masking for the transient domain. Future research will focus on the design of a specific quantizer and coding scheme, as well as the extension to tonal molecules with a nonstationary instantaneous frequency.

Finally, some other signal features can also be taken into account. It should be stressed that the MMP is by no means incompatible with harmonic models, and indeed a possible extension of the MMP is to group together the extracted molecules that belong to a single note, namely its attack and its (potentially harmonic) partials. Ultimately, it is expected that at least for restricted classes of musical sounds, very efficient coding can be performed through the isolation of sound "objects", i.e., a simultaneous transcription and coding. We believe that the MMP may be one efficient tool toward this goal.

Acknowledgment

The author would like to thank the anonymous referees for their valuable comments.

References

[1] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[2] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebart, "Advances in parametric coding for high-quality audio," in *Proc. AES 114th Conv.*, Amsterdam, The Netherlands, 2003.
[3] H. Purnhagen and N. Meine, "HILN—The MPEG-4 parametric audio coding tools," in *Proc. IEEE ISCAS*, Geneva, Switzerland, 2000.
[4] E. Oja, A. Hyvärinen, and P. Hoyer, "Image feature extraction and denoising by sparse coding," *Pattern Anal. and Applic.*, vol. 2, no. 2, 1999.
[5] R. Gribonval, F. Bimbot, and L. Benaroya, "Audio source separation with a single sensor," *IEEE Trans. Speech Audio Process.*, to be published.
[6] L. Daudet and M. Sandier, "MDCT analysis of sinusoids: Explicit results and applications to coding artifacts reduction," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 302–312, 2004.
[7] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. ICASSP*, Munich, Germany, 1997.
[8] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[9] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.
[10] G. Davis, "Adaptive Nonlinear Approximations," Ph.D. dissertation, New York Univ., 1994.

[11] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Signals, Systems, and Computers*, 1993.
[12] S. Jaggi, W. Carl, S. Mallat, and A. Willsky, "High Resolution Pursuit for Feature Extraction," MIT, Cambridge, MA, Tech. Rep., 1995.
[13] R. Gribonval, "Fast matching pursuit with a multiscale dictionary of Gaussian chirps," *IEEE Trans. Signal Process.*, vol. 49, no. 5, pp. 994–1001, May 2001.
[14] R. Gribonval and E. Bacry, "Harmonic decompositions of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, Jan. 2003.
[15] R. Figueras i Ventura, P. Vandergheynst, and P. Frossard, "Low rate and scalable image coding with redundant representations," *IEEE Trans. Image Process.*, to be published.
[16] T. Verma and T. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. ICASSP*, Phoenix, AZ, 1999.
[17] K. Vos, R. Vafin, R. Heusdens, and W. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. AES 17th Int. Conf.*, Sep. 1999.
[18] M. Goodwin, "Multiscale overlap-add sinusoidal modeling using matching pursuit and refinements," in *Proc. IEEE WASPAA*, New Paltz, NY, 1999.
[19] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.
[20] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitionning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, May 1996.
[21] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Process.*, vol. 82, no. 11, pp. 1595–1617, 2002.
[22] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *The Digital Signal Processing Handbook*, V. K. Madisetti and D. Williams, Eds. Boca Raton, FL/Piscataway, NJ: CRC/IEEE Press, 1998.
[23] F. Meyer, A. Averbuch, and R. R. Coifman, "Multilayered image representation: Application to image compression," *IEEE Trans. Image Process.*, vol. 9, no. 11, pp. 1072–1080, Nov. 2002.
[24] M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*. Norwell, MA: Kluwer, 2003.
[25] S. Merdjani and L. Daudet, "Direct estimation of frequency from MDCT-encoded files," in *Proc. DAFx Digital Audio Effects Workshop*, London, U.K., 2003.
[26] L. Daudet, S. Molla, and B. Torrésani, "Transient detection and encoding using wavelet coefficient trees," in *Proc. 18th Symp. GRETSI'01 on Signal and Image Processing*, Toulouse, France, 2001.
[27] S. Molla, "Signaux audiophoniques: Modélisation hybride et schéma de codage," Ph.D. dissertation, Univ. de Provence, France, 2003.
[28] L. Daudet, "Sound Separation in {tones + transients + residual} Using the Molecular Matching Pursuit," Laboratoire d'Acoustique Musicale, Paris, France, http://www.lam.jussieu.fr/src/Membres/Daudet/Separation.html, 2004.

Laurent Daudet (M'03) studied at the Ecole Normale Supérieure, Paris, France, from 1993 to 1997, where he received the degree in statistical and nonlinear physics. In 2000, he received the Ph.D. degree in mathematical modeling from the Université de Provence, Marseille, France, on audio coding and physical modeling of piano strings.

In 2001 and 2002, he was a Marie Curie post-doctoral fellow at the Department of Electronic Engineering, Queen Mary, University of London, London, U.K. Since 2002, he has been a Lecturer at the Université Pierre et Marie Curie (Paris 6), Paris, France, in the Laboratoire d'Acoustique Musicale. His research interests include audio coding, time-frequency and time-scale transforms, analysis of transient signals, and sparse representations for audio.