Hybrid representations for audiophonic signal encoding

Laurent Daudet, B. Torrésani

HYBRID REPRESENTATIONS FOR AUDIOPHONIC SIGNAL ENCODING

L. DAUDET AND B. TORRÉSANI

ABSTRACT. In this paper, we discuss a new approach for signal models in the context of audio signal encoding. The method is based upon hybrid models featuring *simultaneously* transient, tonal and stochastic components in the signal. Contrary to several existing approaches, our method does not rely on any prior segmentation of the signal. The three components are estimated and encoded using a strategy very much in the spirit of transform coding. While the details of the method described here are tailored to audio signals, the general strategy should also apply to other types of signals exhibiting significantly different features, for example images.

Date: October 2001.

A large number of modern signal coding strategies are based on approaches that combine transform coding with suitable quantization and entropy coding of the corresponding coefficients. The transform step is generally achieved via an expansion with respect to a suitably chosen basis, i.e. a basis of some underlying space of signals, with respect to which signals in the considered class are expected to have a "short" expansion. Such approaches have been especially successful for still image compression, for which it has been shown that wavelet (or sub-band) expansions are extremely well adapted, and allow one to achieve high compression rates. This seems to be due to the very structure of images, which often feature slowly varying regions, separated by sharp edges. It turns out that wavelet decompositions are very well adapted to such situations, in the sense that corresponding expansions may be extremely sparse, and that the significant wavelet coefficients generally group themselves into clusters, a property which is exploited in several coders [26, 28].

In this paper, we shall be mainly concerned with the case of audiophonic signals. Transform coding strategies have been successfully applied to audio signals (see, for example, the various versions of MPEG audio coders). However, the structure of audio signals makes the signal encoding problem significantly different from the problem of image encoding, because of the possible *superposition* of different kinds of components: whereas different objects tend to occlude each other in images, they rather superimpose to each other in the case of (polyphonic) audio signals[1]. Since these components may have significantly different behaviors (for example, fast-varying transients, or slower "partials") it is difficult to derive a transform that can be simultaneously matched for all of them. It is the purpose of this paper to describe and analyze the main features of a new approach for modelling and encoding such "composite signals". The approach we propose is based on a signal model of the form

Tonals + transients + residual

and a simultaneous transform coding of the tonal and transient components, using different transforms [12] (we refer to [11] for a more detailed analysis.) The estimation and removal of transients from the signal plays the role of a "stationarization" of the signal; the estimation and removal of the tonal component plays a role similar to an estimation of peaks in the power spectrum. Hence, if the two components above are successfully estimated and removed from the signal, the residual should take the form of a stationary random signal with smooth power spectrum, and should be easy to describe by simple autoregressive models with short memory. This is the goal we shall follow in the present article. Even though we shall not describe here a complete coding scheme (much work remains to be done on optimizing some aspects of our approach), the main ingredients will be analyzed in detail.

It is important to notice that such a strategy automatically introduces redundancy into the representation, which results in extra cost in terms of the amount of data to encode. It is hoped that the adequacy of the models for the three considered components will result in significantly better compression rates for each of them, and a competitive global compression rate.

Such hybrid signal models are quite flexible, which opens interesting possibilities of signal processing directly in the transform domain. Among these, let us mention psychoacoustic processing such as masking (see e.g. [38] for a detailed presentation). Masking is the ability of human auditory system to ignore certain signals when they are received simultaneously with "close" signals with higher amplitude. Masking is actually used in modern audio coders: instead of minimizing the decoding error, the coded tries to make it "transparent" from a perceptive point of view (see for example [4, 25]). It is known that several types of

[1] A similar effect is also present in the case of transparent objects in images, but the effect is much less significant.

masking (for which several models have been proposed) should be taken into account: temporal masking and frequency masking (including different phenomena for masking tonal and wide band signals). A hybrid representation for audio signals should allow one to implement different masking algorithms on different components, and therefore improve coding schemes significantly. Other examples of transform domain processing include several types of signal modifications, such as pitch transposition or speed modification. Such modifications have been discussed in [34].

This paper is organized as follows. We first describe in Section 2 the main ingredients of the transform coding strategies we will use in the sequel. Section 3 is devoted to the description of the hybrid representations on which our work is based, and the improvements introduced by structured representations are described in Section 4. We summarize the signal coding aspects and discuss potential applications to other signal processing problems in Section 5, and conclude in Section 6.

Acknowledgements: L. Daudet is supported by the European Community programme *Human Potential* under contract number HPMF-CT-2000-00917. The authors wish to thank Ph. Guillemain, K. Jensen, R. Kronland-Martinet, S. Molla and M. Sandler for stimulating discussions. The illustrations presented in this paper have been generated using the MATLAB software, in particular the `WaveLab` package developed by the statistics group at Stanford University[2], the `runlength` function written by O. Kelly[3] and the arithmetic coding package written by K. Skretting[4].

[2]available at `http://www-stat.stanford.edu/wavelab`

[3]available at `http://www.mathworks.com/support/ftp/miscv5.shtml`

[4]available at `http://www.ux.his.no/karlsk`

2. Adapted representations for transform coding

In transform coding strategies, the transform step is equivalent to the expansion of the signal with respect to a suitably chosen basis. Among the most popular transforms for transform coding, the local cosine bases and wavelet bases have received particular attention. Local cosine bases have been shown to provide adequate approximations of Karhunen-Loève bases for locally stationary signals, and wavelet bases (or the equivalent subband decompositions in the discrete case) have been extremely successful for coding signals exhibiting abrupt changes. We briefly sketch our notations here, before discussing the interest of each transform for specific kinds of signals.

2.1. **Transform coding, adapted transforms, non-linear expansions.** Transform coding [32, 33] amounts to performing a linear transformation prior to quantization and encoding. Such a linear transformation may be understood as a change of basis, or an expansion with respect to a specific basis of the considered space of signals. The goal of the transform is to "factor out" the redundancies present in the samples, and therefore to reduce the amount of significant coefficients. Transforms which achieve such goals are "adapted" to the signal class. A huge literature has been devoted to adapted transforms in the recent years. We refer to [19] and [33] and references therein for more details.
In the framework of transform coding schemes, when a large number of coefficients are below a given threshold (the accuracy of the encoder), encoding all the coefficients irrespective to their actual significance may become particularly inefficient, and it makes sense to encode only the significant coefficients. The resulting encoded signal is then a *non linear approximation* of the original signal, since the choice of the coefficients to be retained depends on the signal itself[5]. Obviously, such a strategy becomes really useful only when the necessary encoding of the "significance map" (i.e. the addresses of significant coefficients) may be achieved efficiently.

Remark 1. One may also mention at this point the *best basis* strategies advocated in [37], which seek the transform optimizing the sparsity of the representation in a library of possible transforms. The resulting (signal dependent) basis is an *adaptive basis*, and has to be encoded. □

For the present discussion, we shall limit ourselves to two special cases of transforms using bases adapted to specific signal features (and not adaptive bases), namely wavelet and local trigonometric bases. We shall also formulate the problems mainly in terms of expansion of analog signals (i.e. we first describe the transform step in terms of choice of a basis in spaces of functions) before turning to a formulation adapted to discrete signals.

2.2. **Wavelets and subband coding.** We only sketch here the aspects which are relevant for our purpose, and limit ourselves to the simplest version of wavelet theory. More details on wavelet expansions (construction of filters, examples...) and generalizations (biorthogonal decompositions, wavelets on bounded domains,...) may be found for example in [9, 21].
Wavelets introduce themselves naturally in the framework of a multiresolution analysis (MRA for short). A MRA is associated with a pair of L^2 functions ϕ (scaling function) and ψ (wavelet), satisfying the *two-scale difference equations* (or *refinement equations*)

$$\phi(t) = \sqrt{2}\sum_k h_k \phi(2t+k)\ , \quad \psi(t) = \sqrt{2}\sum_k g_k \phi(2t+k), \qquad t \in \mathbb{R}\ , \tag{1}$$

[5]Indeed, the corresponding expansion of the sum of two signals will not be the sum of the expansions of the signals.

where h and g are ℓ^1 sequences (generally with finite support), whose discrete Fourier transforms H and G satisfy the "perfect reconstruction conditions":

$$|H(\omega)|^2 + |G(\omega)|^2 = 2 , \tag{2}$$

$$H(\omega+\pi)\overline{H}(\omega) + G(\omega+\pi)\overline{G}(\omega) = 0 . \tag{3}$$

With ϕ and ψ, associate the family of scaled and shifted copies ϕ_{jk} and ψ_{jk}, defined by

$$\psi_{jk}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right) , \quad \phi_{jk}(t) = 2^{-j/2}\phi\left(2^{-j}t - k\right) . \tag{4}$$

Remarkably enough, given filters h and g satisfying equations (2), and additional technical conditions (the so-called Cohen-Lawton conditions, see for example [9]), there exist a corresponding MRA, such that the wavelets $\{\psi_{jk}, j, k \in \mathbb{Z}\}$ form an orthonormal basis of $L^2(\mathbb{R})$. Therefore, any $x \in L^2(\mathbb{R})$ may be decomposed as the scalar products

$$x = \sum_k s_{j_0 k}\phi_{j_0 k} + \sum_{j \le j_0}\sum_k d_{jk}\psi_{jk} , \tag{5}$$

where j_0 is a fixed reference scale index, and the coefficients are defined as

$$s_{jk} = \langle x, \phi_{jk}\rangle , \quad \text{and} \quad d_{jk} = \langle x, \psi_{jk}\rangle . \tag{6}$$

An important feature of wavelet expansions is the fact that the computation of the s and d coefficients may be done using a fast recursive algorithm, identical to the subband coding schemes [36]:

$$s_{j-1\,k} = \sum_\ell \overline{h}_{2k-\ell} s_{j\ell} , \quad \text{and} \quad d_{j-1\,k} = \sum_\ell \overline{g}_{2k-\ell} s_{j\ell} . \tag{7}$$

$$s_{j+1\,\ell} = \sum_k \left(h_{2k-\ell} s_{jk} + g_{2k-\ell} d_{jk}\right) . \tag{8}$$

The connection between wavelet expansions and subband coding allows the extension of wavelet transforms to the case of discrete signals. When only samples are available, the latter are generally identified with the scaling function coefficients s at the finest scale, say, $j = 0$ for simplicity (alternatives are discussed in [13]), and the corresponding wavelet coefficients at coarser scales $j > 0$ are just computed[6] using the relations (7).

REMARK 2. Unlike several authors (see for example [29]), we shall not consider in the present article the more general subband decompositions (wavelet packets, see [37]), yielding different partitions of the time-frequency plane. Such expansions may be tailored specially to match the critical frequency bands of human auditory system, and therefore become quite efficient in the audio coding context. However, since we shall be only interested in subband coding of transients, the choice of classical wavelets is a natural one. In addition, since our goal is to characterize transients, we limit ourselves to short filters, emphasizing time localization. □

Signals with finite support may be treated in many different ways. The simplest one amounts to considering a signal defined on an interval, say $[0, 1]$, as the restriction to $[0, 1]$ of a 1-periodic function, and use corresponding wavelet expansions. The alternative consists of changing the construction rule for wavelets whose support intersect the boundaries of the interval, still preserving the very properties of wavelet bases (orthogonality, completeness, vanishing moments,...) We shall limit ourselves to the first solution, and refer to [8] for a discussion of wavelet bases on intervals.

[6] An alternative algorithm, based on an initial polyphase decomposition and a factorization into a series of simple lifting steps [30, 10] may also be used, resulting in more efficient implementations.

A remarkable feature of wavelet bases lies in their behavior with respect to non linear approximation in a variety of functional spaces. The non linear approximation problem, which turns out to be very close to the algorithms which are used in practice, may be described as follows: given a signal x and a fixed integer N, seek the best possible approximation of x as a linear combination

$$S_N = \sum_{n=0}^{N-1} c_n u_{i(n)}$$

of N basis vectors $u_{i(0)}, \dots u_{i(N-1)}$ in a given basis: the best approximation is the one which minimizes a given norm of the residual:

$$\min_i \|x - S_N\|$$

For general choices of the basis (and the norm), the selection of the best non linear approximation turns out to be a difficult problem. However, in the case of wavelet bases, the situation becomes much simpler. For example, assuming that $x \in L^p([0,1])$, set $\delta_{jk} = 2^{-(1/p-1/2)j} d_{jk}$, pick the N largest δ_{jk} coefficients and denote by Σ_N the corresponding N terms wavelet expansion. Then the striking result is that Σ_N achieves the same asymptotic rate of convergence as S_N: there exists a constant C_p such that for all $x \in L^p([0,1])$, $\|x - \Sigma_N\|_p \leq C_p \inf \|x - S_N\|_p$. In addition, several families of functional spaces may be characterized by the speed of convergence of non linear wavelet approximations (we refer to [7, 14] for a detailed mathematical presentation). Among them, the Besov spaces have received a particular attention in the context of image modeling, since they provide good models for functions with controllable "density" of singularities of a given strength. For the same reasons, such models also seem adequate for describing transient signals.

2.3. **Local trigonometric bases.** It is well known that any square-integrable function on an interval may be represented by a Fourier series, or a cosine series, and that functions on the real line also possess such trigonometric expansions, obtained by first segmenting the real axis into blocks, and using classical trigonometric expansions within each block. More recently, smooth versions of such bases (avoiding brutal "block" segmentation) have been proposed (see [37] or [21] for a review). The construction (in the case of the real line; the adaptation to bounded domains is straightforward) goes as follows. Consider an increasing sequence of numbers $\{a_k\ , k \in \mathbb{Z}\}$, and further numbers η_k such that

$$a_k + \eta_k < a_{k+1} - \eta_{k+1}\ , \tag{9}$$

and consider a set of functions w_k such that

$$\begin{cases} w_k(t) &= 1 \text{ if } t \in [a_k + \eta_k, a_{k+1} - \eta_{k+1}]\ , \\ w_k(t) &= 0 \text{ if } t \notin [a_k - \eta_k, a_{k+1} + \eta_{k+1}]\ . \end{cases} \tag{10}$$

Set

$$u_{kn}(t) = \sqrt{\frac{2}{\ell_k}}\, w_k(t) \cos\left[\frac{\pi}{\ell_k}\left(n + \frac{1}{2}\right)(t - a_k)\right]\ , \quad n = 0, 1, 2, \dots \tag{11}$$

where $\ell_k = a_{k+1} - a_k$ is the length of the k-th interval. Notice that (9) becomes $\eta_k + \eta_{k+1} \leq \ell_k$. Then, if the windows w_k satisfy the compatibility relations: $\forall \tau$ such that $0 \leq |\tau| \leq \eta_k$,

$$\begin{cases} w_{k-1}(a_k + \tau) &= w_k(a_k - \tau)\ , \\ w_k(a_k + \tau)^2 &+ w_{k-1}(a_k + \tau)^2 = 1\ , \end{cases} \tag{12}$$

then the functions $\{u_{kn}\ , k \in \mathbb{Z}, n \in \mathbb{Z}^+\}$ form an orthonormal basis of $L^2(\mathbb{R})$. Classical choices for the bell functions include sine functions (actually implemented in MPEG coders) and modified Kaiser-Bessel windows.

Again, the bases of L^2 spaces have their discrete counterparts. Discrete local cosine bases[7] (based upon similar constructions) are described in [37] or [21], and have been used in several instances in signal coding, for example in the MPEG audio coders (see e.g. [4]).

REMARK 3. Again, we shall not use these expansions in full generality, and limit ourselves to expansions involving fixed size windows: $w_k(t) = w_0(t - a_k)$. The more general solution offers the possibility of a better match with the tonal component of the signal, at the price of higher complexity. □

The choice of local cosine bases is motivated by the will of describing tonal signals, i.e. signals which may be reasonably modeled as stationary signals within a given time frame, and may, in addition, be characterized by a few coefficients (which excludes "wide band" stationary or locally signals). Therefore, this opens the problem of non linear approximation by local cosine bases. This problem has been addressed by Gröchenig and Samarah [16], who have shown that the so-called modulation spaces introduced by Feichtinger in the eighties provide an appropriate setting for non linear approximation with local cosine bases. A function x is well approximated by local cosine bases if and only if it belongs to a certain modulation space, i.e. if its continuous Gabor (or short time Fourier) transform G_x (see [5] for notations) belongs to some $L^p(\mathbb{R}^2)$ space, with $1 \leq p < 2$.

2.4. **Overcomplete systems.** Besides the expansions with respect to orthonormal (or biorthogonal) bases, the alternative offered by overcomplete decompositions has received much attention in recent years. The main ingredient in such approaches is a large library of elementary waveforms, from which an optimal expansion is seeked. Several methods have been proposed for achieving such a program, implementing various libraries, optimality criteria and optimization techniques. Among them, let us quote in particular

- The "best basis" algorithm (already mentioned above, see [37]), which looks for an optimal basis within a library of orthonormal bases, based on a sparsity criterion and a dynamic programming search algorithm. The resulting optimal expansion involves orthonormal waveforms.
- Matching pursuits (see [21, 15]), and several cognates (basis pursuit [6], and more general greedy and weak greedy algorithms), which result in an expansion which is generally not orthonormal, through algorithms which are often quite time consuming.

The approach we are about to describe shares some of the features of the above mentioned methods (overcompleteness, search for sparsity), but does not rely on any greedy algorithm. In particular, our approach indeed uses different bases for describing different features of the signal, but avoids optimization of the bases (mainly to stick to simple and fast decomposition algorithms).

[7] Expansions with respect to these bases are sometimes called LOT (lapped orthogonal transforms) or - in the case of a sine window - MDCT (modulated discrete cosine transforms).

3. An hybrid model, projections

In this paper, we are concerned with hybrid signal models that include components of different kinds. More specifically, we will limit our investigations to additive models of the form *"Tonal + Transient + Noise"*. Such models have been considered in the literature in various contexts, for example for sound modeling and transformation (see for example [27]), or for encoding and compression (see [18, 20, 34, 35] and references therein). However, the models developed in a signal encoding context are generally based upon a segmentation of the signal into its tonal and transient components. Transients and tonals, as we shall term them from now on, are then encoded using different transforms.
Here, we develop an approach avoiding such a segmentation, by considering superpositions of transient, tonal and stochastic[8] components:

$$x(t) = x_{ton}(t) + x_{tr}(t) + x_s(t) \ .$$

The main difficulty lies in the joint estimation of the three components. Ideally, a simultaneous estimation would be desirable, as it is difficult to predict in advance the respective bit rates to be assigned for each component. However such a procedure seems difficult to implement (except maybe using matching pursuit strategies [23], at the price of a very high computational cost), and we shall limit ourselves to individual estimates for the transients and the tonals, defining the stochastic part as a residual. The estimation of tonal and transient components follow similar lines, i.e. are in the spirit of transform coding schemes, using local cosine and wavelet bases respectively. However, in each case, only the largest coefficients in each time frame are retained. The algorithm therefore depends on a pair of threshold values, which are estimated adaptively within larger time frames. The choice of these threshold values is related to the step used in the quantization stage. For the sake of simplicity, we shall limit ourselves to (mid-tread) uniform quantization in this paper, but more sophisticated alternatives will be briefly mentioned and discussed.

As an illustration of the potential benefits of such hybrid schemes, let us consider a simple example: we take a small portion of the `MPEG` test signal gspi35.1[9]. In Figure 1, we display the original signal, MDCT approximations of the tonal part, and the residuals corresponding to two decompositions of the signal using MDCT only and MDCT and wavelets respectively. In the first case (second and third plots), we only used MDCT expansion: the signal was expanded with respect to a MDCT basis, and the 950 largest coefficients (1.45 percent of the total number of coefficients) were selected. One may clearly see in the non-tonal residual (plot 3) that the attack has been poorly captured by the MDCT basis, and is still present in the residual. The energy of the residual represents 9.97 percent of the total energy of the signal.
In the second case (plots 4 and 5), only the 850 largest MDCT coefficients (1.3 percent of the total number of coefficients) were selected, the corresponding tonal component was subtracted, and the 100 top wavelet coefficients were selected to estimate the transient part. The corresponding residual (whose energy represents 6.58 percent of the total energy of the signal) is shown in the bottom plot. The residual has clearly a much smaller dynamical range in this case, and is much easier to model as a wide sense stationary process. In addition, the description of the transient (the attack here) using wavelets is much more parsimonious (ie. it requires much less non-zero coefficients). More complex examples are discussed below.

[8] The term "stochastic" refers to the terminology used in [20, 34], and to the model used to describe such a component; "residual" would be more appropriate.

[9] Available at `http://www.tnt.uni-hannover.de/project/mpeg/audio/sqam/`

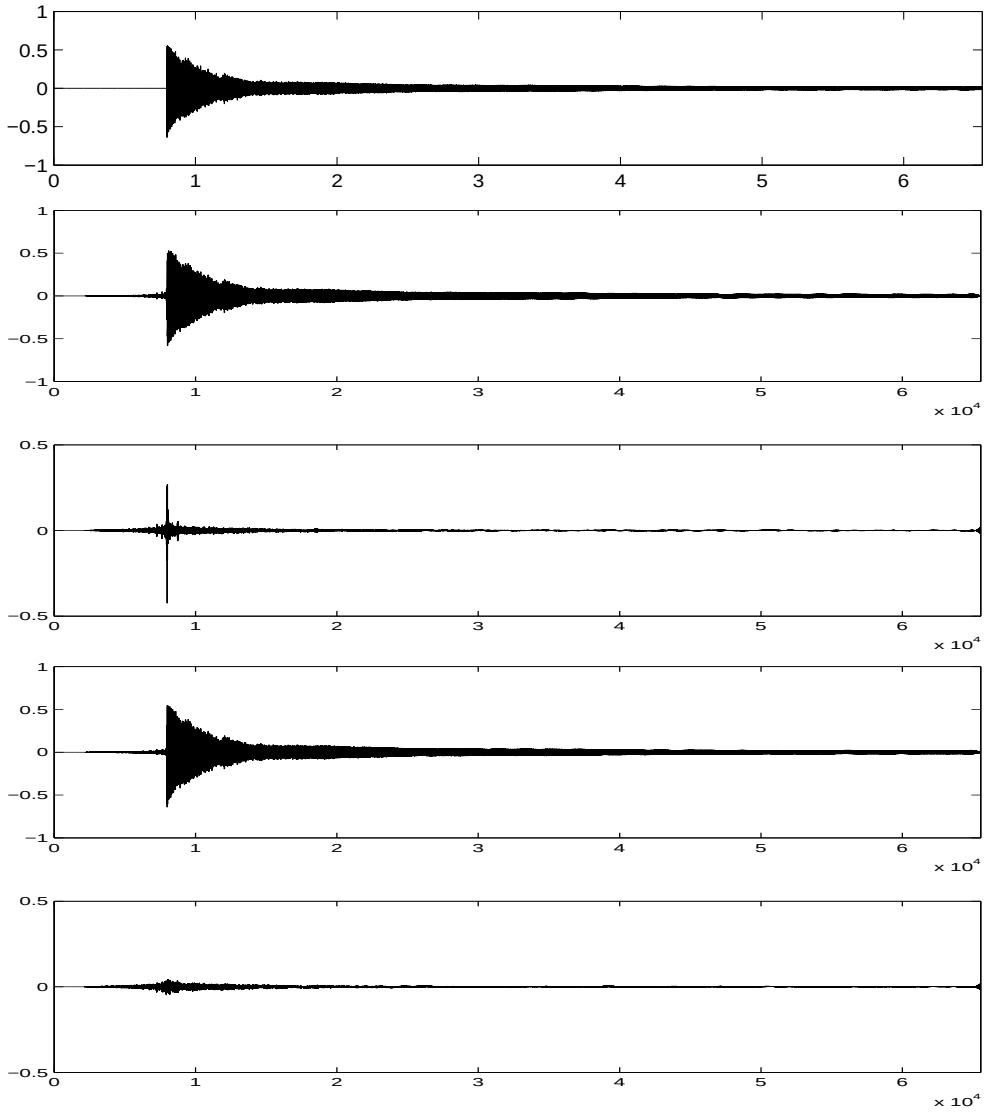

FIGURE 1. MDCT and MDCT+wavelet expansions: from top to bottom: a 65536 samples segment of the MPEG glockenspiel signal: approximation of a tonal part from 950 MDCT coefficients; corresponding residual; approximation of a tonal part from 850 MDCT coefficients and 100 wavelet coefficients; corresponding residual.

3.1. **Tonal component modeling and estimation.**

3.1.1. *Tonals and local cosines.* Estimation and modeling of "tonals" (or "sinusoidal partials") is a fairly classical topic in the sound and speech signal processing literature. Examples may be found for example in [24, 27], in different contexts. We stick here to transform coding schemes, and use expansions with respect to adapted bases. A "tonal" may be modeled as a locally stationary signal in the sense of [22], i.e. well (in other words, sparsely) represented in local cosine bases. The following example provides an illustration of the expected situation.

REMARK 4. Let us assume a model of the form

$$x_{ton}(t) = A(t)Y(t) \ ,$$

where A and Y are uncorrelated second order random processes, such that the correlation function $C_A(t,s) := \mathbb{E}\{A(t)A(s)\}$ of A is continuously differentiable and slowly varying, and Y is wide sense stationary (hence $C_Y(t,s) = C_Y(t-s)$) and zero mean. Denote by $\mathcal{S}_Y$ the spectral density of Y. Given a local cosine basis $\{u_{kn}\}$ as defined in the previous section, we easily obtain the following approximate expression for the variance of the coefficients $\langle x_{ton}, u_{kn}\rangle$:

$$\begin{aligned}\mathbb{E}\left\{|\langle x_{ton}, u_{kn}\rangle|^2\right\} &= \int C_A(t,s)C_Y(t-s)u_{kn}(t)u_{kn}(s)\,dt\,ds \\ &= \frac{1}{2\pi}\,C_A(c_k,c_k)\int |\widehat{u_{kn}}(\omega)|^2\,\mathcal{S}_Y(\omega)\,d\omega + r \ ,\end{aligned}$$

where we have denoted by $c_k = (a_k + a_{k+1})/2$ the center of the window w_k. In the latter approximation, r is a remainder term which depends on the gradient of C_A:

$$|r| \leq K \sup_{t,s\in \mathrm{supp}(w_k)} |\partial_t C_A(t,s)| \ .$$

Hence, the more slowly varying C_A (within the support of w_k), the better the approximation.
Taking into account the expression of u_{kn}, and assuming that the remainder r can be neglected[10], we finally obtain

$$\mathbb{E}\left\{|\langle x_{ton}, u_{kn}\rangle|^2\right\} \approx \frac{1}{8\pi}\,C_A(c_k,c_k)\int |\hat{w}_k(\omega - \pi(n+1/2)/\ell_k) + \hat{w}_k(\omega + \pi(n+1/2)/\ell_k)|^2\,\mathcal{S}_Y(\omega)\,d\omega \ .$$

Since the windows w_k are generally localized near the origin in the Fourier domain, the main contributions to such an expression are provided by the neighborhoods of $\omega = \pm\pi(n+1/2)/\ell_k$. Assuming that the power spectrum $\mathcal{S}_Y$ contains sharp peaks at frequencies $\omega_1, \omega_2, \ldots$, the coefficients $\langle x_{ton}, u_{kn}\rangle$ such that $n + 1/2 \approx \omega_1\ell_k/\pi, \ldots$ may be expected to be significantly large, so that they will be selected by an appropriate thresholding.
Similar results may be obtained using a slightly more sophisticated model allowing for frequency modulation: $x(t) = A(t)Y(t)e^{i\varphi(t)}$ with A and Y as before, and φ is a (smooth) local phase function whose derivative φ' is slowly varying. The conclusion is similar, as long as the windows w_k are such that the variations of φ' within the support of each w_k are small enough. □

Summarizing, a tonal is expected to be adequately represented by (a small number of) significant local cosine coefficients.

3.1.2. *Estimation of tonals.* Following ideas developed by the Yale group [3], we then "model" the tonal component as the component of the input signal which is "best described" by local cosines. More precisely, let $\{u_{kn}\}$ denote an MDCT basis as given in (11), associated with a family of identical (sine) windows $w_k(t) = w(t - a_k)$ and an uniform segmentation of the time domain ($a_{k+1} = a_k + \ell$ and $\eta_k = \eta$ for all k). The windows are chosen in such a way that the tonal components may reasonably be assumed stationary within the window. The windows' overlap is maximised (i.e. taken such that $\eta = \ell/2$), so as to optimize

[10] Such an assumption implies restrictions on the choice of the windows w_k, in particular on their length ℓ_k: the windows w_k have to be short enough to ensure that C_A is slowly varying enough within w_k; on the other hand it has to be large enough to ensure a sufficient frequency resolution.

frequency localization. A typical value for the window half-length ℓ is about 20 milliseconds (ie. 1024 coefficients at 44.1 kHz sampling rate). Given the input signal x we consider the MDCT coefficients

$$\alpha_{kn} = \alpha_\lambda = \langle x, u_{kn} \rangle \ , \tag{13}$$

where $\lambda = (k, n)$ denote a joint time-frequency index, and pick the following first (non linear) estimate of the tonal part

$$x^0_{ton} = \sum_{i=0}^{N_{ton}-1} \alpha_{\lambda(i)} u_{\lambda(i)} \ , \tag{14}$$

where the coefficients $\alpha_{\lambda(i)}$ are the N_{ton} coefficients α_{kn} whose modulus exceeds a given threshold τ_{ton}. The determination of the threshold τ_{ton} (or the number of atoms N_{ton}) is a crucial aspect of the algorithm. τ_{ton} must be small enough to capture most of the tonal information, but large enough to allow efficient compression. Since we work in the context of uniform quantization (mid-tread uniform quantizer, see [19]), the threshold τ_{ton} is also intimately related to the quantization intervals. More precisely, we take τ_{ton} in the vicinity of the quantization step (an analogy of the case of low-resolution transform coding strategies [21]).

Furthermore, in order to capture the dynamic (amplitude) variations of the signal, τ_{ton} must be adapted locally. In this work, we use the following strategy. The distribution of the α coefficients is estimated within a "large" time frame (consisting of several "small" time frames), and the critical value z_p corresponding to a given p-value (i.e. such that $\mathbb{P}\{|\alpha| \geq z_p\} = p$) is estimated. The threshold τ_{ton} is a fixed multiple of the critical value:

$$\tau_{ton} = \rho z_p \ . \tag{15}$$

Within a "small" time frame all coefficients α larger than the threshold are retained. Typical values for large and small frame lengths are about 100 milliseconds and 20 milliseconds respectively.

REMARK 5. The tonal component estimation may also be understood as a way of selecting locally the discrete component of the power spectrum. However, since our approach is based upon a transform followed by a thresholding (and not a simple "peak picking" as in sinusoidal coders), it may be necessary to correct for the overall behavior of the latter, which may vary by several orders of magnitude. Such a correction may be done by an appropriate normalization, as follows.

The selection of the significant coefficients α_λ may be performed on appropriately weighted coefficients (where the weighting depends on the frequency variable only): given a family of weights $\mathbf{w}_0, \mathbf{w}_1, \ldots$, consider the coefficients

$$\tilde{\alpha}_{kn} = \mathbf{w}_n \alpha_{kn} \ . \tag{16}$$

Denote by $\lambda(i)$, $i = 0, 2, \ldots N_{ton} - 1$ the indices corresponding to the N_{ton} weighted coefficients $\tilde{\alpha}_{kn}$ whose absolute value exceeds a given threshold τ_{ton} (and where we have set $\mathbf{w}_{k\ell} = \mathbf{w}_k$), and consider the approximation (14) obtained with this new set of coefficients. Typical choices for the weighting functions are of the form

$$\mathbf{w}_n = \frac{1}{(\nu_0 + \nu(n))^\alpha} \ , \tag{17}$$

where $\nu(n)$ is the value of the frequency corresponding to the frequency index n, and the constants ν_0 and α are adjusted so as to tune the behavior at low frequencies (typically, frequencies much smaller than ν_0 are not affected by the weighting) and the decay at high frequencies. The effect on such a normalization appears clearly on FIGURE 5 below. □

3.1.3. *Quantization and encoding.* The encoding of the tonal component involves encoding of the addresses of retained coefficients, and quantization and encoding of their values. This information is sufficient for the decoder to reproduce the tonal component, according to equation (14).
The encoding of the significance map is a classical issue. We use standard run-length coding of the significance map (see e.g. [19] for details), followed by entropy coding of the lengths.
Let us first focus on the quantization of the coefficients. Consider a uniform quantizer Q, and denote by $Q(\alpha)$ the quantized coefficients. The simplest approach consists in quantizing the N_{ton} coefficients $\alpha_{\lambda(0)}, \dots \alpha_{\lambda(N_{ton}-1)}$, which yields the corresponding tonal component

$$x_{ton} = \sum_{i=0}^{N_{ton}-1} Q(\alpha_{\lambda(i)}) u_{\lambda(i)} \ , \tag{18}$$

The quantized coefficients $Q(\alpha_{\lambda(0)}), \dots Q(\alpha_{\lambda(N_{ton}-1)})$, are then entropy coded.

REMARK 6. The simple quantization above gives the same precision to all the frequency bands of the signal. However, it is known that different frequencies are not treated equally in the auditory system. In particular, high frequencies are much less audible than intermediate frequencies. In our case, a simple uniform quantization indeed leads us to spend a significant part of the allowed bit budget for coding domains of the frequency scale which are poorly relevant from a perceptive point of view. Without going into sophisticated modeling of the auditory system (e.g. masking effects), it is nevertheless possible to take simple information into account by simple normalization of the coefficients, involving a modeling of the listening threshold. For the latter, we use the following expression, taken from [31]

$$T_q(\nu) \approx 3.64\nu^{-0.8} - 6.5e^{-0.6(\nu-3.3)^2} + 0.001\nu^4 \ , \tag{19}$$

where the frequency ν is expressed in kHz, and the threshold T_q is in dB. Setting

$$A(n) = 10^{T_q(n)/20} \ ,$$

and introducing a family of (possibly) frequency dependent uniform quantizers Q_n, we finally consider the following quantizer

$$\tilde{Q}(\alpha_{kn}) = Q_n\left(\frac{\alpha_{kn}}{A(\nu(n))}\right) \tag{20}$$

(where again $\nu(n)$ is the frequency associated with the index n) and the corresponding estimate of the tonal component,

$$x_{ton} = \sum_{i=0}^{N_{ton}-1} A(\nu[i]) \tilde{Q}(\alpha_{\lambda(i)}) u_{\lambda(i)} \ , \tag{21}$$

with the obvious definition: $\nu[i] = \nu(n)$ for $\lambda(i) = (k, n)$). The quantized coefficients $\tilde{Q}(\alpha_{\lambda(i)})$ are finally entropy coded.
Several variations around such choices are of course possible (implementing for example scale factors as in the MPEG coders). Those issues will be discussed in more detail elsewhere. □

3.2. **Transient component modeling and estimation.** After the estimation of a tonal component (21), the "non-tonal" signal is then defined as

$$x_{nt} = x - x^0_{ton} \ , \tag{22}$$

and has energy

$$\|x_{nt}\|^2 = \|x\|^2 - \sum_{n=0}^{N_{ton}-1} |\alpha_{\lambda(n)}|^2 \ .$$

3.2.1. *Transients.* According to the discussion of Section 2.2, transients are components of the signal which exhibit fast variations, and are therefore well represented in a wavelet basis. This means that transients should be characterized by a small number of large wavelet coefficients. An expansion of the signal (with the tonal component removed) with respect to a wavelet basis (using a wavelet ψ with sharp time localization, i.e. filters with short impulse response) followed by a suitable thresholding is therefore expected to provide an estimate for the transient component.
In order to reduce boundary effects (which would manifest themselves as "transients"), the wavelet coefficients corresponding to a given time frame are computed with a periodized wavelet transform on a larger (say, 3 times bigger) time frame. Within such a framework, it is possible to adjust the depth (i.e. the maximum number of scales) of the expansion to the frame length and the filter length so as to keep perfect reconstruction [11].

3.2.2. *Transient estimation.* The procedure for transient estimation (in its simplest form, see Section 4 below for the "structured" version) is very similar to the previous one. Starting from the estimated non-tonal signal within a "small" time frame, we seek the part corresponding to the largest wavelet coefficients. Introducing a joint "time-scale" index $\mu = (j,k)$ and a family of scale-dependent weights $\mathbf{w}_0, \mathbf{w}_1, \ldots$ (introduced in order to take into account some overall behavior of the coefficients across scales, as in REMARK 5), set

$$\beta_{jk} = \beta_\mu = \langle x_{nt}, \psi_{jk} \rangle \ , \quad \text{and} \quad \tilde{\beta}_{jk} = \tilde{\beta}_\mu = \mathbf{w}_j \beta_{jk} \ , \tag{23}$$

and define

$$x_{tr}^0 = \sum_{m=0}^{N_{tr}-1} \beta_{\mu(m)} \, \psi_{\mu(m)} \ , \tag{24}$$

where the sum is taken over the set of indices $\mu(m)$ corresponding to the N_{tr} largest coefficients $\tilde{\beta}_{jk}$, whose modulus exceeds a given threshold τ_{tr}.
The threshold τ_{tr} may be determined along lines similar to the estimation of τ_{ton} in (15): the empirical distribution of the β coefficients is estimated within a "large" window, and the threshold τ_{tr} is set to a fixed multiple of a given percentile.

3.2.3. *Quantization and encoding.* After the selection of significant coefficients, the resulting coefficients are (uniformly) quantized, and entropy coded. Of course, the discussion of REMARK 5 still holds true: further perceptive arguments (such as listening threshold, or more involved masking models) may be invoked to further refine the quantizer. Here, we stick to the simple version. Denoting by Q_{tr} the quantizer used for the α coefficients,

$$x_{tr} = \sum_{m=0}^{N_{tr}-1} Q_{tr}(\beta_{\mu(m)}) \, \psi_{\mu(m)} \ , \tag{25}$$

3.3. **Modeling the residual.** The residual component of the signal is defined by

$$x_{res} = x_{nt} - x^0_{tr} \ , \tag{26}$$

and has energy

$$\|x_{res}\|^2 = \|x\|^2 - \sum_{n=0}^{N_{ton}-1} |\alpha_{\lambda(n)}|^2 - \sum_{n=0}^{N_{tr}-1} |\beta_{\mu(n)}|^2 \ .$$

If the estimation of the transient and the tonal parts is successful, the residual should behave as a wide band (locally) stationary signal, and is modeled as such. More precisely, x_{res} is modeled within each time frame (typically 1024 samples) using an autoregressive model of fixed length (a typical value for the filter length is 20 samples). The model parameters are estimated using a standard Levinson algorithm. This procedure is essentially the same as the one used in LPC coders used for speech coding. The filter coefficients within each time frame are quantized and encoded. In the present stage, we use uniform quantization with 16 bits per coefficient.
In the decoding stage, the residual is simulated as a stationary stochastic process, using the AR model with the prescribed coefficients, the input being a pseudo random number generator.

3.4. **A simple hybrid encoding scheme.** In this paragraph, we describe the main steps of a simple scheme implementing the above approach (see also FIGURE 2), and exhibit corresponding numerical results for illustration.

(1) Tonal component estimation:
- Expand the signal x with respect to a local cosine basis.
- Pick the most significant coefficients and form the tonal part x_{ton} in (18).
- Quantize the significant coefficients (uniform quantization).
- Encode the significance map (run length coding).
- Subtract the tonal part.

(2) Transient component estimation:
- Expand the non-tonal signal x_{nt} with respect to a wavelet basis.
- Pick the most significant coefficients, and form the transient part x_{tr}.
- Quantize the significant coefficients (uniform quantization).
- Encode the significance map (run length coding).
- Subtract the transient part.

(3) Residual estimation:
- Fit an autoregressive model of given order to x_{res}.
- Quantize the filter coefficients.

(4) Entropy coding, multiplexing

The interest of such hybrid representations is best illustrated by audio signals containing both tonals and transients. We display here results obtained on a small segment of the glockenspiel *gspi35.2*, (65536 samples, 44.1 kHz, 16 bit/sample) taken from the MPEG sound examples[11]. The signal is displayed in FIGURE 3, together with its spectrogram -which indeed exhibits tonals (i.e. "horizontal" features) and transients ("vertical" features).
The hybrid decomposition (without quantization) is shown in FIGURE 4. We used windows of length 2048 (about 45 milliseconds), and the tonal estimate corresponds to 606 (weighted) MDCT coefficients (0.92%). The weighting parameters in Eq. (17) were set to $\alpha = 2$, and $\nu_0 \approx 3000Hz$. The tonal estimate still

[11] Available at http://www.tnt.uni-hannover.de/project/mpeg/audio/sqam/

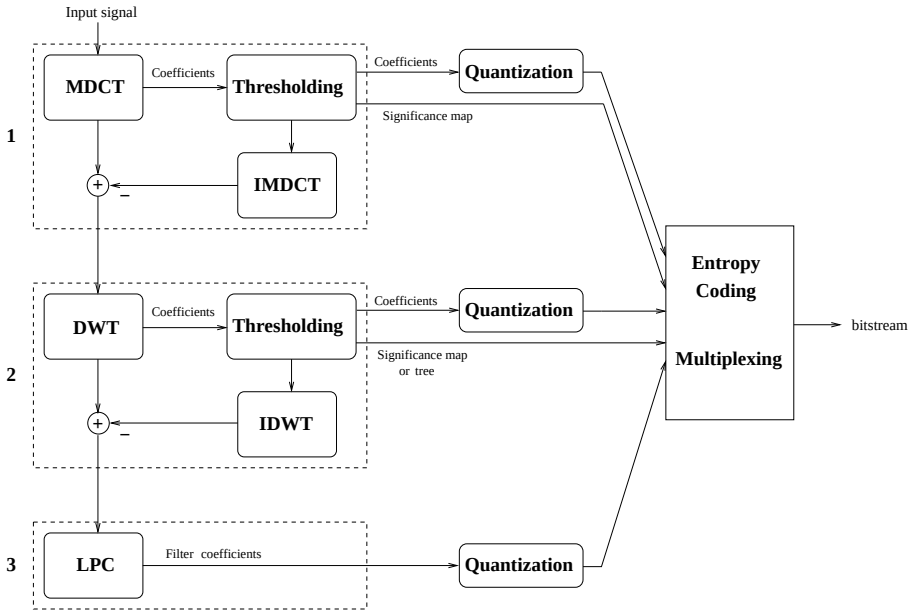

FIGURE 2. The prototype hybrid coder; MDCT and IMDCT stand for direct and inverse modulated cosine transforms respectively; DWT and IDWT stand for direct and inverse discrete wavelet transform respectively; LPC stands for linear predictive coding.

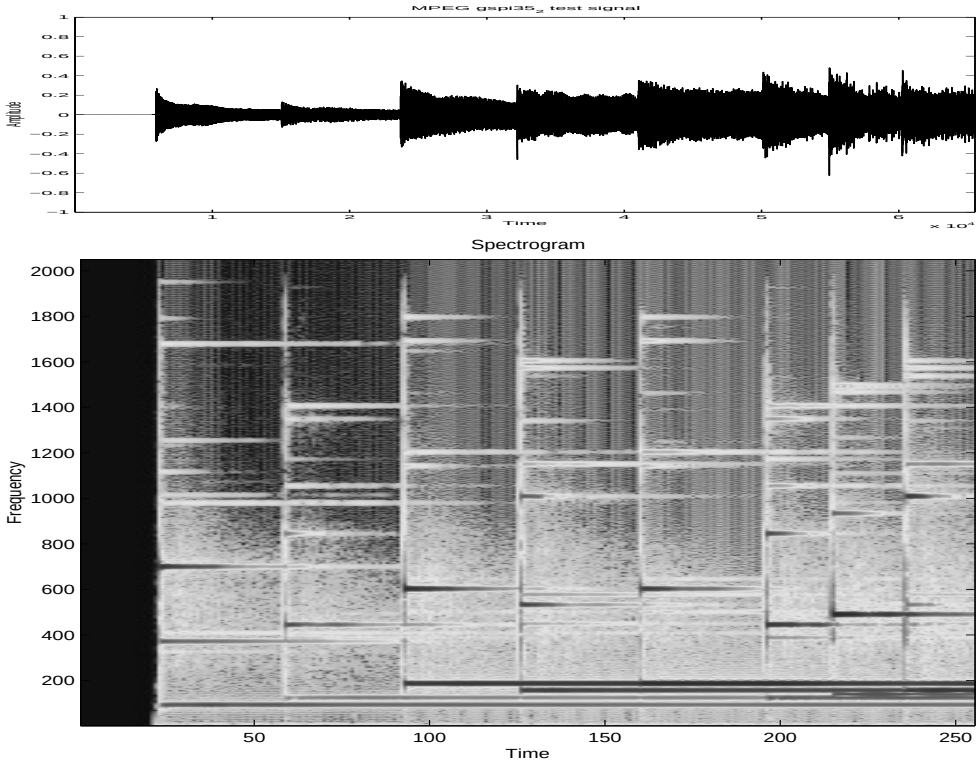

FIGURE 3. MPEG test signal (*Glockenspiel*) and its spectrogram.

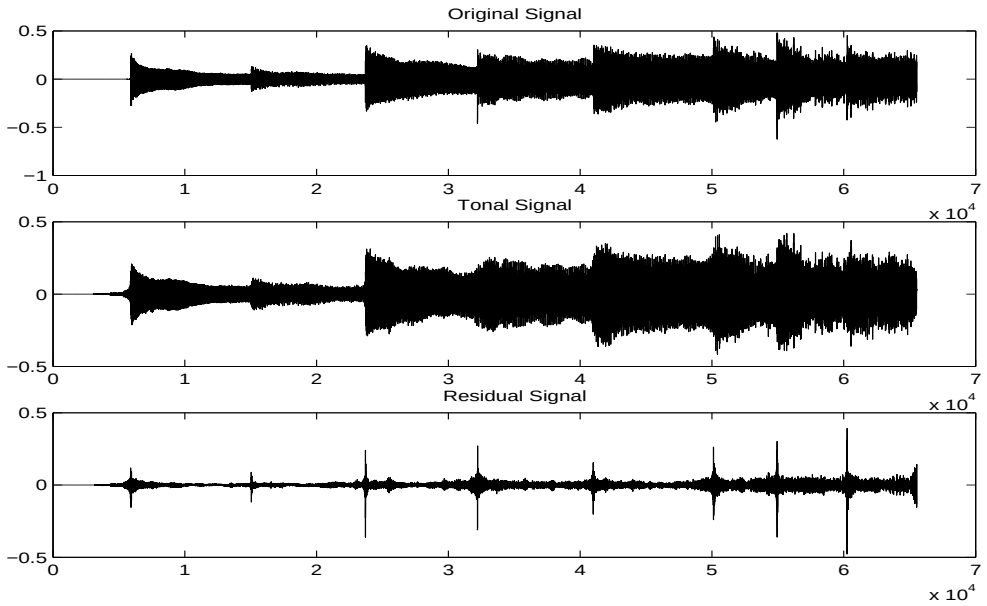

FIGURE 4. Hybrid decomposition MPEG gspi35.2 test signal. From top to bottom: original signal, tonal and non tonal parts, parts, estimated using weighted MDCT.

exhibits transitions which appear sharp on the plot, but turn out to be "slower" in a closer examination; it also features an expected "pre-echo" effect. As may be seen, the removal of the tonal estimate yields sharp attacks, together with a small "noise-like" component. Listening to the corresponding sounds[12] shows a satisfactory separation between what one would intuitively call attacks and tonals (even though the non-tonal part seems to contain tonal information, which appears difficult to avoid.)

The weighting (see REMARK 5) appears to play a significant role in the resolution of the tonal component. This is best illustrated by FIGURE 5, where we display the power spectra of the different components using unweighted and weighted tonal estimates. The power spectra were estimated using a simple periodogram, without any smoothing since here we want to emphasize the behavior of the discrete part (the peaks) of the spectrum. The second and third plots (starting from top) show that the unweighted MDCT procedure fails at picking all the peaks in the power spectrum: the high frequency components are not well resolved in the tonal estimate, and are still significantly present in the non tonal one. Such an effect is clearly corrected by an appropriate weighting, such as the one proposed in (16) and (17). In the weighted tonal estimate, all the peaks are satisfactorily resolved.

In that particular example, after quantization (using a uniform quantizer, as in (18), without any weighting, and quantization step equal to the threshold value, which yields 11 bits per coefficient), we obtained quite a small number (606, i.e. less than 1%) of significant coefficients. After run-length coding (maximal length=256) followed by Huffman coding, the significance map yields a cost of about 0.074 bits per coefficient (0.088 bits per coefficient if Huffman coding is not employed). Huffman encoding of the coefficients yields a cost of about 0.093 bits per coefficient. All together, the encoding of the tonal part requires about 0.167 bits per sample. More precise values are given in TABLE 1.

[12]Corresponding `wav` files are available at `http://www.cmi.univ-mrs.fr/~torresan/SP.html`, together with other examples.

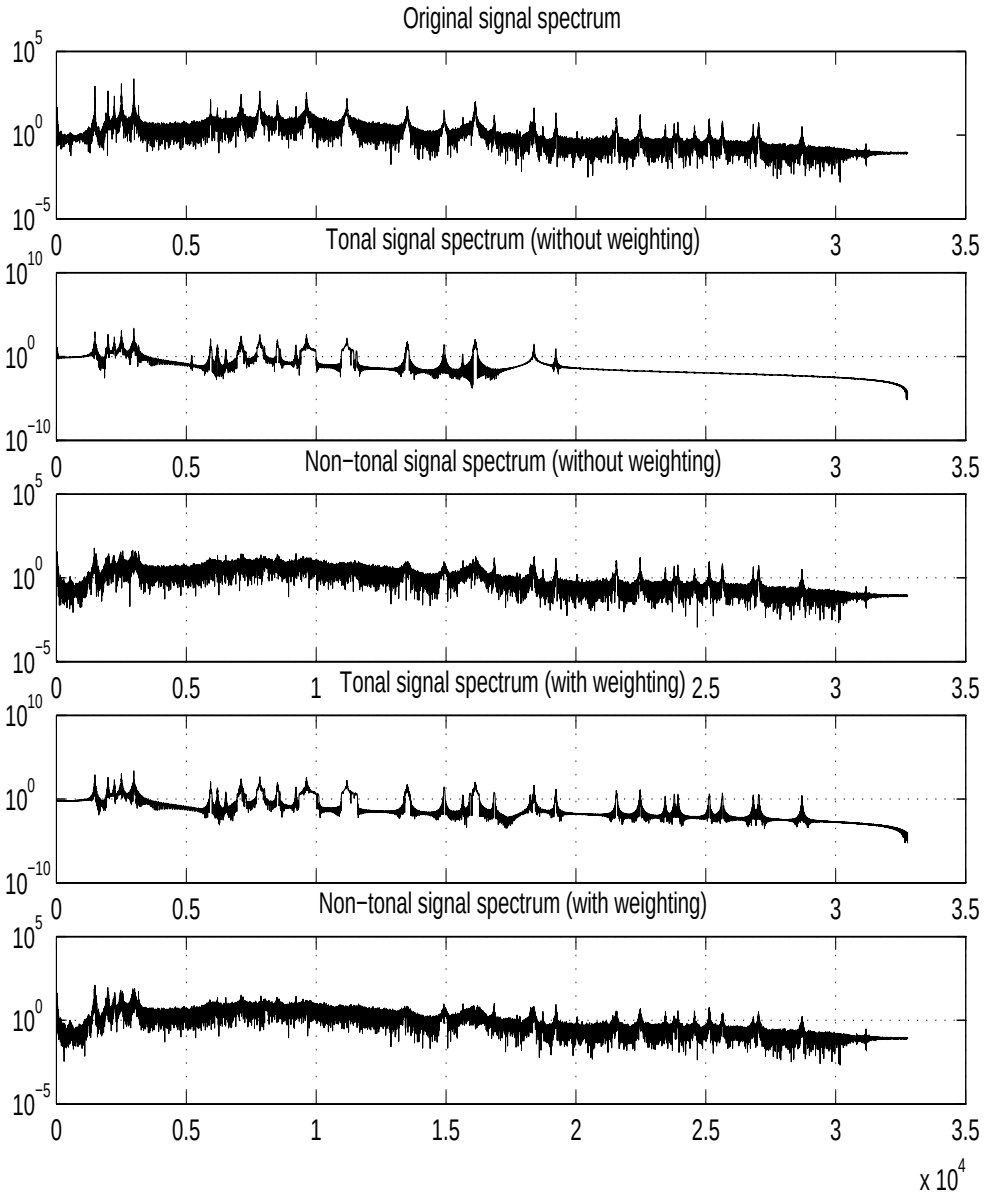

FIGURE 5. Power spectra for the MPEG gspi35.2 test signal. From top to bottom: original signal, tonal and non tonal parts, estimated without weighting the MDCT coefficients, and tonal and non tonal parts, estimated with weighting the MDCT coefficients.

In FIGURE. 6, we display the estimation of the transient component via the selection of the largest coefficients. In that particular experiment, we used the 3591 largest(unweighted) wavelet coefficients (i.e. the 5.48% largest). As may be seen, the most important part of the transient features of the non tonal signal have been captured, and the residual has much slower variations.

As before, we use entropy coded quantized coefficients and significance map. After Huffman coding, the bit rate used to encode the quantized was about 0.54 bits per coefficient. Huffman coded run-length representation of the significance map (particularly efficient here) yields a cost of about 0.2 bits per coefficient (0.51 bits per coefficient without entropy coding). All together, the encoding of the transient

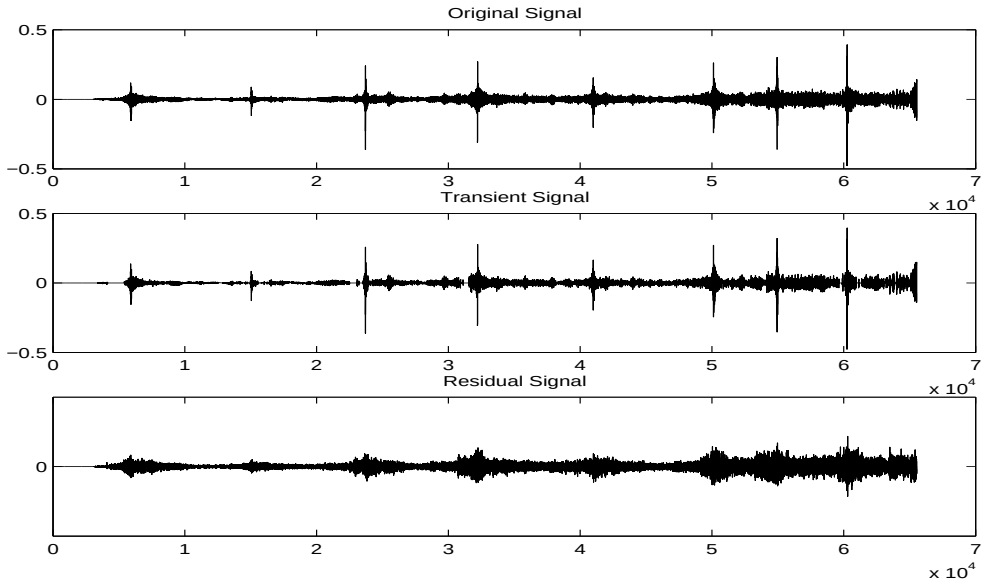

FIGURE 6. Hybrid decomposition of the MPEG `gspi35.2` test signal (followed). From top to bottom: non-tonal signal, transient component and residual, estimated using FWT.

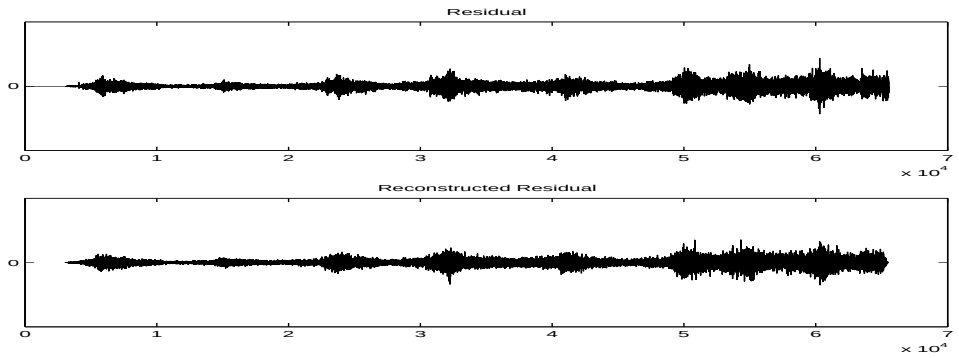

FIGURE 7. Hybrid expansion: residual (top), and corresponding LPC synthetized residual.

component requires 0.8 bits per sample, which is quite a large number. More precise figures are given in TABLE 1.

The residual is encoded using a traditional LPC procedure: filter coefficients are estimated within windows of fixed length (using smooth overlaping window functions), and uniformly quantized. In our experiment, we used Hanning windows of length 1024 to estimate 16 filter coefficients (uniformly quantized using 16 bits per coefficient). The result is displayed in FIGURE 7.

The overall bitrate for this experiment is of the order of 1.22 bits per sample, i.e. 53.84 kbits/sec.

As appears on the considered examples, the encoding of the transient component often involves a large number of significant wavelet coefficients. Meanwhile, the description is not completely satisfactory, as the estimated transient part is often difficult to separate from the residual (in other words, the optimization of

Component		number of bits	bits/sample
Tonal	coefficients	6072	0.0927
	signif. map	4840	0.0739
	global	10912	**0.1665**
Transient	coefficients	39501	0.6027
	signif. map	13208	0.2015
	global	52709	**0.8043**
Residual		16384	**0.25**
Global bit rate		80005	**1.2208**

TABLE 1. Hybrid decomposition of `gspi2.35` signal.

the coefficients is a very difficult task). We shall see in the next section how to overcome these drawbacks (i.e. reduce the data volume and improving the model) by introducing structure in the wavelet coefficient domain.

3.5. **Alternate projections.** The approach above may be criticized in several respects. A potential shortcoming of the "two-steps" estimation of the tonal and transient components is that the estimation of tonals is biased by the presence of transients and vice versa. To avoid such drawbacks, an alternative strategy would be to use alternate projections, as suggested in [3]: first estimate a tonal component using MDCT, with a large value of the threshold τ_{ton} (or a small value of N_{ton}), and subtract it from the signal. One then obtains a very small number of "tonal atoms", which are subtracted from the signal. Then estimate a transient component, still with a large τ_{tr}, which results in a small number of wavelets. Then, iterate the procedure, until the residual may be satisfactorily encoded using LPC. Referring to FIGURE 2, steps **1** and **2** are iterated several times before moving to step **3**.
Such an approach is very much in the spirit of additive models and matching pursuit (see [15]), alluded to in the previous section. One expects a priori more precise estimates for the tonals and transients, at the price of an increased computational burden. Notice that such a procedure only modifies the estimation part, the decoding stage is not affected.

However, in our experiments, implementing such a two step estimation procedure does not seem to improve the results significantly in terms of the sparseness of the representation. Likewise, matching pursuit strategies, while being much more computer intensive, do not seem to yield much sparser expansions. This seems to indicate that the considered bases are sufficiently different to separate transient and tonal components without iterative strategies. In addition, a complete implementation of such methods requires updating the MDCT and wavelet coefficients at each stage, which yields a large computational cost.
In this respect, it is worth mentioning the recent work [2] which addresses the problem of blind source separation from a single captor, a problem similar to the one we address here. The authors propose a criterion for testing the discriminating power of different bases, and it would be interesting to study this criterion in the case of the bases considered here.

4.1. **The interest of structured expansions.** As stressed already, transform coding based on non linear expansions opens up the problem of encoding the addresses of significant coefficients. In the absence of additional information, the cost of significance map encoding is one bit per sample if a single basis is used, and two bits per sample if two different bases are used as in our case. If the representation of the signal is sparse enough, direct encoding of addresses may nevertheless become efficient: for example, if in a given time frame of length say $L = 2^J$ samples, only N coefficients out of L are significant, encoding their address results in a cost of NJ bits, which may be much less than L (using typical values, $N = 50$ significant coefficients out of $L = 1024$ results in a cost of approximately half a bit per coefficient for the significance map).
However, this cost may be significantly reduced if the significant coefficients are "structured", i.e. if they are not distributed arbitrarily. When the significant coefficients are structured, the structure information may be used to reduce the address encoding cost. A simple example is provided by run-length coding of significance maps: if significant and insignificant coefficients are not distributed in an homogeneous way, i.e. if they have a tendency to group into clusters of significant or insignificant coefficients, it becomes "cheaper" to encode the lengths of the clusters (using entropy coding if necessary) rather than the significance map itself.

4.2. **Trees of significant wavelet coefficients.** Here, we shall discuss another kind of (deterministic) structure, and limit our discussion to the coding of transient components (similar ideas might be developed in the case of the tonal component, along lines similar to the sinusoidal models discussed in [24], or using adapted matching pursuits [15], but we shall not follow those lines here). Our starting point is the fact that transients not only manifest themselves by large wavelet coefficients (in particular at small scales), but rather by clusters of significant wavelet coefficients [11].

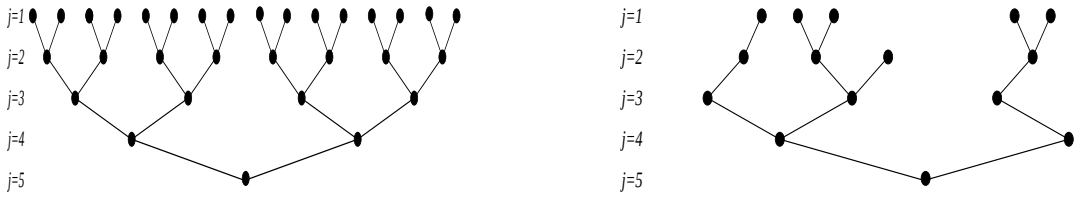

Figure 8. Wavelet coefficients grid, and a subtree of significant coefficients.

4.2.1. *A tree model for transients.* Because of the natural dyadic grid structure of the wavelet coefficients[13] (see Figure 8), the latter often take the structure of a dyadic tree, which suggests the use of dyadic trees of wavelet coefficients as a model for transients. In a structural model for transients, a transient is defined by a dyadic tree of significant wavelet coefficients, connected in the following sense: a wavelet coefficient $\alpha_{j+1,2k}$ or $\alpha_{j+1,2k+1}$ can be considered significant only if the parent coefficient $\alpha_{j,k}$ is significant. Such an approach is very much in the spirit of the the image coders based upon trees of (significant or insignificant) wavelet coefficients, for example the EZW [28] of the SPIHT [26] algorithms - however in our case we only consider trees whose roots are at the largest scale.

[13] Assuming for the sake of simplicity that the function ψ is centered at the origin $t = 0$, then the wavelet ψ_{jk} is centered at $t = k2^j$, which yields a dyadic grid.

Such a connected tree of wavelet coefficients is considered relevant if, for all connected branches $\mathcal{B}$, the corresponding wavelet coefficients are significant *in some average*. This is mathematically expressed by the fact that the following modulus of regularity

$$\kappa_{q,s}[\mathcal{B}] = \frac{1}{|\mathcal{B}|} \sum_{(j,k)\in\mathcal{B}} 2^{js} |\alpha_{j,k}|^q \ , \tag{27}$$

exceeds a given maximum value (here, $|\mathcal{B}|$ denotes the length of the branch $\mathcal{B}$). A branch $\mathcal{B}$ is called a *full length* branch if its leaves always correspond to the finest considered scales. A tree is a *complete tree* if all its branches are full length branches.

The choice of assigning a cost to the branches of the trees (and not to the trees themselves) is motivated by the will of assigning a *local* feature to a transient: the leaves of the branches are naturally associated with samples in the signals.

The constants s, q characterize the considered type of transients, in the sense that they weight coefficients corresponding to different scales (here, 2^{js} plays the same role as the weight $\mathbf{w}_j$ in section 3.2.) For example, large (positive) values of s emphasize large scales, and favor trees with short branches, whereas smaller values favor longer branches. A choice of s therefore represents an "a priori" model for the transients to be considered. It can be mentioned that higher-complexity models with more general scale-dependent weights can be considered. The choice of q also influences the type of transients to be estimated, in the sense that small values of q enforce sparsity.

4.2.2. *Tree estimation.* In practice, the tree (i.e. the transient associated with the tree) has to be estimated from the signal, and encoded. For the estimation of transients, we use a dual "top-down" search algorithm: first a tree containing full-length branches only (ie. maximum length) is selected, and secondly these branches are pruned from insignificant sub-branches.

Starting from a leaf ℓ of the complete tree, consider the branch of all its ancestors, denoted by $\mathcal{D}_\ell$, and form the following quantity:

$$\kappa_{q,s}[\ell] = \sum_{(j,k)\in\mathcal{D}_\ell} 2^{js} |\alpha_{j,k}|^q \ , \tag{28}$$

where s, q characterize the type of transients which are to be retained. In the numerical examples given below, we limit ourselves to the simplest choices $s = 0$ and $q = 1$ or 2. The selection of full branches is done by retaining the leaves of the tree ℓ such that $\kappa_{q,s}[\ell]$ exceeds a threshold value $\tilde{\kappa}$.

In a second top-down pass, sub-branches of insignificant coefficients are pruned down: starting from scale $j = 1$, one prunes down *leaves* that are insignificant with respect to a certain (fixed) threshold ρ. Note that only leaves are considered in this stage, which ensures that after pruning the trees remains connected. A good choice for ρ in the case $s = 0$ and $q = 1$ is $\rho = \tilde{\kappa}/J$, which represents the mean of the absolute value of the coefficients on a branch that has just been selected during the first stage.

In practice, this second pass often reduces the number of coefficients by a very significant amount (typically by a factor of 2 to 3), as small-scales coefficients usually have very small amplitudes. All the wavelet coefficients belonging to the retained tree are then quantized and encoded. The corresponding "significance tree" also has to be encoded, which requires 1 bit to encode the significance of the root, 2 bits per node at scales $j = 2 \dots J$ (to specify the significance of the children of the node), and nothing for leaves at the smallest scale $j = 1$ (they can't have significant children). For example, the tree exhibited in FIGURE 8 (right) requires 21 bits for encoding the significance tree (15 coefficients).

REMARK 7. Again, the selection of the threshold $\tilde{\kappa}$ is a crucial ingredient. In order to capture the whole dynamics of the signal, a good alternative is to adapt it locally to the signal. The value $\tilde{\kappa}[\ell]$ has to be estimated within a time frame larger than the time frame defined by the complete tree. This may be done using a comparison of the κ function with a smoothed version $\tilde{\kappa}$. Given such a smoothed version, consider all the values ℓ such that

$$\frac{|\kappa[\ell] - \tilde{\kappa}[\ell]|}{\kappa[\ell]} \geq p$$

where p is a reference percentile.

REMARK 8. There exist many alternatives to the method presented here for tree estimation. Let us mention for the record the approach based upon hidden Markov trees of significant wavelet coefficients developed in [12] following [1], which models the signal with a mixture random model consisting of "transient" wavelet coefficients and "non transient" coefficients.

Besides being well adapted to the description of transient signals, we have seen in this section that tree structures are also quite efficient in terms of coding, as one can easily construct strategies that require a maximum of 2 bits per *retained* coefficient for encoding the significance map of a binary tree. Therefore, if the number of retained coefficients is small enough (which is one of the objectives of transform coding strategies), the cost of significance map encoding is expected to be quite low.
However, the underlying assumption that significant coefficients have a significant parent is sometimes not verified, which results in cases where we have to encode insignificant coefficients. Statistics based on a wide variety of sounds show that the overall gain is usually very much in favor of this structured approach.

4.3. **Results, parameter estimation.** Let us describe in this section the coding costs associated with the transient part of the sound example presented in the previous section (first 2^{16} samples of MPEG gspi35.2 test signal). The top plot in fig. 9 shows the non tonal part, which is the residual of the signal after tonal part extraction. The corresponding modulus of regularity $\kappa_{0,1}[\ell]$, renormalized to a maximum value of 1 is shown below. The threshold $\tilde{\kappa}$ for branches selection has here been taken as $\tilde{\kappa} = .16$. As we only use $J = 8$ scales in our wavelet decompositions, we have to consider one tree every 256 samples. The last two plots in fig. 9 show the corresponding trees, before and after the subbranches pruning, respectively. The threshold for pruning has been taken as $\rho = \tilde{\kappa}/J = .02$.
The cost of coding the "significance tree" (ie. the structure of the tree) is in this case 256 bits to encode the significance of the roots, and 3962 bits for the intermediate nodes (2 bits per node at scales $j = 2 \dots 8$), hence a total of 4218 bits.
The 2313 coefficients at scales $j = 2 \dots 8$ are uniformly quantized with a quantization step equal to the threshold ρ. The cost for each coefficient is in this case 6 bits (5 bits for the absolute value, and 1 bit for the sign). Using entropy coding, this bit requirement can be decreased by a substantial amount, in our case the cost is as low as 3.04 bit / coefficient.
The total cost (significance map + quantized value of the coefficients) is in this case 10666 bits, which amounts to an average of .16 bit / sample. It is interesting to note that this value is similar to the bitrates required for encoding the tonal part.

The residual, defined as before, is encoded using the same method (using Hanning windows of size 1024, and 16 coefficients long filters). The corresponding graphs are displayed in FIGURE 11.

The total bitrate, after adding the cost of encoding the tonal and stochastic parts, amounts to about .58 bit / sample, which, at our sampling rate (44.1 kHz), is about 25.5 kbits/sec. As expected at such low

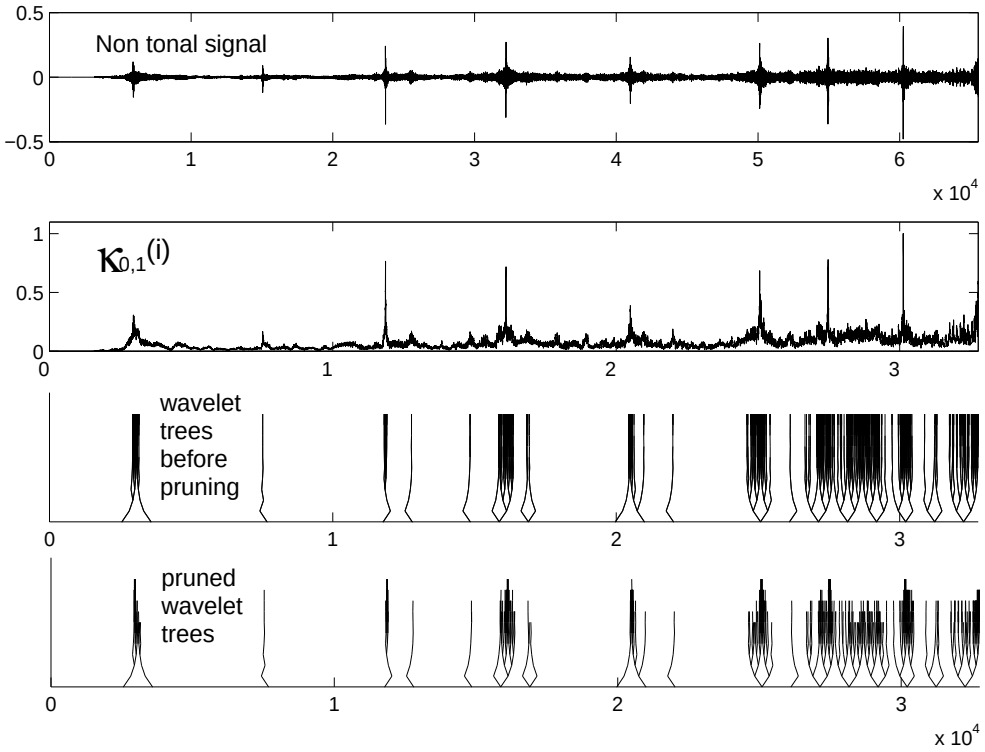

FIGURE 9. Extraction of transient part with structured wavelet expansions, MPEG `gspi35.2` test signal. From top to bottom: non-tonal signal, modulus of regularity $\kappa_{0,1}[\ell]$, trees of wavelet coefficients before pruning, pruned trees.

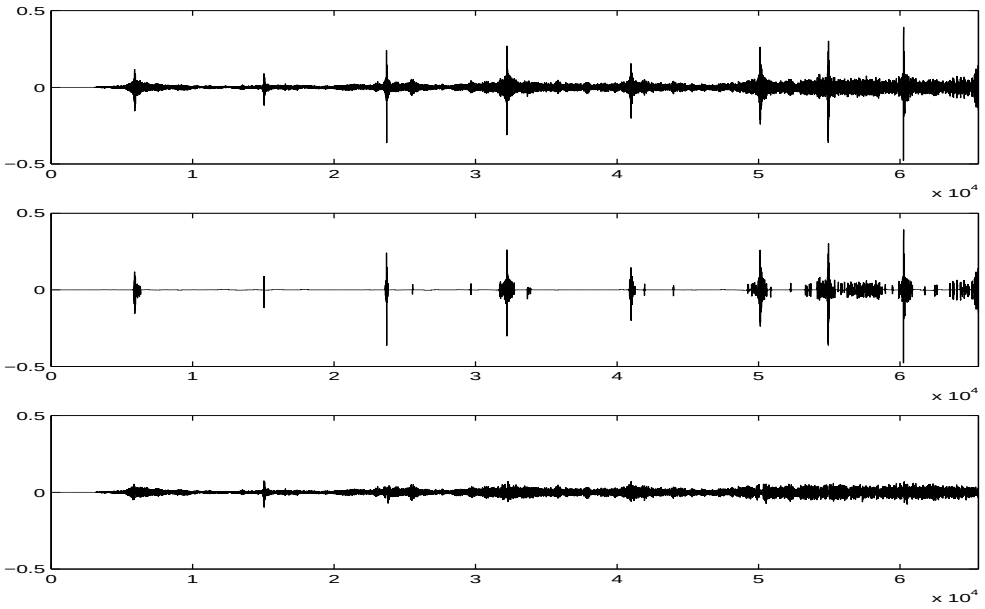

FIGURE 10. Extraction of transient part with structured wavelet expansions, MPEG gspi35.2 test signal. From top to bottom: non-tonal signal, reconstructed transient part (from tree shown on the bottom plot of fig. 9) and residual (stochastic part).

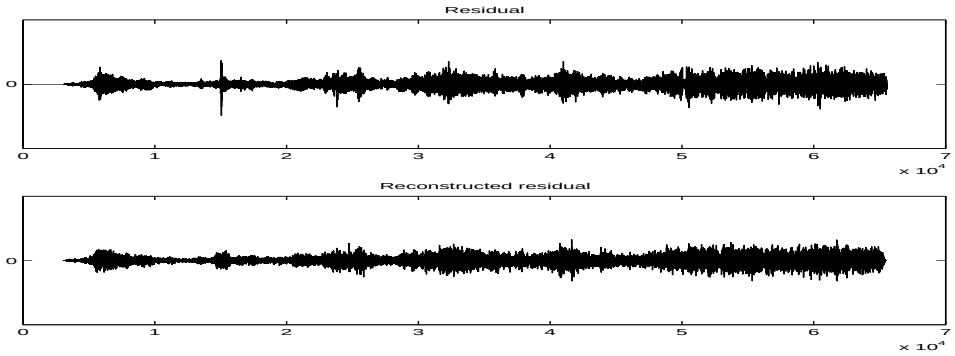

FIGURE 11. Hybrid expansion using structured trees of wavelet coefficients: residual (top), and corresponding LPC synthetized residual.

Component		number of bits	bits/sample
Unstructured transients	coefficients	39501	0.6027
	signif. map	13208	0.2015
	global	52709	**0.8043**
Total (tonal+transients+residual)		80005	**1.2208**
Tree-structured transients	coefficients	7031	0.1073
	signif. tree	4218	0.0644
	global	10666	**0.1628**
Total (tonal+transients+residual)		37962	**0.5793**

TABLE 2. Comparison of coding cost associated with the transient part, in the unstructured and tree-structured cases (`gspi2.35` test signal).

bit-rates, some distortion is audible, due partly from the unperfect separation between different features (some tonal components and transients can still be heard in the residual), and from the quantization the introduces distortion.

In order to validate our method, similar experiments have to be made on a wider rage of musical sounds. it is very likely that some of them (for instance orchestral pieces, with a very high sinusoidal content) would lead to higher bitrates. Corresponding results will be given in a forthcoming publication.

The interest of hybrid representations for audio coding is clearly illustrated by the results of the previous sections, even though further investigations are still needed in various aspects of the proposed scheme. Among the aspects which deserve particular attention, let us quote in particular the following issues:

- The quantization of the transformed coefficients: for the sake of simplicity, we have here only displayed results obtained using uniform quantization of the transformed coefficients. The use of more elaborate quantization schemes, involving the weighting alluded to in Remark 6 of Section 3, or more elaborate perceptual models, are likely to improve the performances of the coding scheme.
- The encoding of the residual: at the bit rates obtained here, the cost of encoding the residual (i.e. the LPC coefficients) is fairly important, and further optimization (including an analysis of the optimal filter length and the precision required for quantizing the filter coefficients) is needed here.
- The optimization of parameters (thresholds, weights) as function of time. The implementation of high level perceptive characteristics of the signals (for a posteriori correction) is another one of the lines we plan to follow in the near future.

In addition to those important issues, and in view of the improvement introduced by structured expansions in the case of the transient components, we also plan to investigate possibly structured MDCT expansions for describing tonals.

However, besides the aforementioned application to audio coding, hybrid schemes as presented in this paper are a good framework for numerous applications in the field of audio analysis and synthesis. Let us review three of these, namely masking modeling, audio effects and high-level audio representations.

- **Masking modeling** One of the shortcomings of subband-based audio coding schemes, like those derived from the MPEG family, is that they rely heavily on the psychoacoustic model. Masking models constructed from the spectrum and based on psychoacoustic models, tell us which amount of quantization noise is tolerable within each subband, ie. what is the minimum number of bits that we are allowed to use in the quantization stage in order to ensure "transparency". Because all coefficients in a given subband are considered and subject to quantization (on a small or even zero number of bits, hence the data reduction), the estimation of a relevant masking curve is a key point. One main objection is that the validity of the models for signals exhibiting a large number of spectral components is highly questionable, as it is for non-stationary sources (for instance the only value of the spectrum my lead us to misinterpret a sharp but well-defined transient as some wide-band noise).

  On the other hand, transform coding (where the large majority of coefficients is set to zero and only the coefficients that are selected as significant are quantized), one is less sensitive to the design of the (non-uniform) quantizer. Nonetheless, psychoacoustic principles can still be applied to improve the perceptual quality of the coding scheme. In this respect, an efficient separation in tonal, transient and stochastic components is a good framework for an enhanced psychoacoustic model. The tonal and stochastic contributions in each frequency channel are a natural estimate for the *tonality index* that appears in the computation of thresholds in the MPEG psychoacoustic model. Moreover, the information on sharp transients can be used to explicitly implement some temporal masking.

Is is believed that psychoacoustic modeling can be significantly improved by the use of redundant (overcomplete) representations, in the same way as redundancy itself plays an important role in the neuro-cognitive processes of human perception [38].

- **Signal modifications** Following the same lines, we can also discuss the importance of separating the tonal, transient and stochastic components of a soundfile for a wide range of sound effects. Amongst them, let us focus specifically of time stretching without pitch modification [17] - or its dual problem (up to a global resampling) pitch shifting without time modifications. This problem - ill-posed from a strict mathematical point of view, as many "real life" signal processing problems - is one of the most difficult to implement in order to handle the largest variety of sounds.

  Most popular methods are based on the now classical phase vocoder (frequency-domain methods), or waveform duplication (time-domain methods). However, when sharp transients are present in the signal, none of these approaches gives satisfactory results, as frequency methods tend to smear transients, whereas time-domain methods would duplicate them. Indeed, a perceptually good algorithm needs to leave the transients unchanged. Our method can therefore be used as a transient detection scheme, and a segmentation can be performed between steady-state (processed using standard schemes) and transient regions (left unchanged). Spectral envelope modeling can also be used in the tonal and stochastic parts in order to preserve formantic structures.
- **High-level signal representations** Another promising aspect in this framework of representations based on an explicit model of the structure of musical signals is that it can lead to higher-level representations, such as single notes (or more generally *objets sonores*). The different classes of such sound objects could be parametrized with features issued from our scheme. For instance a single organ note could be described as an attack (made of transients) and a sustained part (composite with a tonal part - harmonic - and stochastic components - the typical turbulence noise associated with the reed-jet interaction). This is very much in the spirit of object-oriented coders, such as MPEG-4. However, the transcription algorithms (ie. in this case the passage from soundfiles to "object" files, typically a pattern recognition algorithm) will obviously become of extreme complexity in the most general polyphonic multi-instrumental case.

6. Conclusion

In this paper we have discussed some advantages of overcomplete representations for audio signals. Amongst representation spaces, a union of local cosine and dyadic wavelets have proved good separation properties for structural features of audio signals, their tonal part and transient part respectively. The separation of these two layers is improved by the use of structured representations, that also greatly reduce the cost of encoding significance maps (these structures can be seen as a way to enforce constraints that have been released by overcompleteness). The residual of these models can generally be modeled without perceivable difference as stochastic processes with slowly-varying parameters. Preliminary results indicate that good sound quality can be obtained at bitrates as low as about 20 kbits/s for some test mono files.

It is important to note that the algorithms presented here are all based on "fast" numerical schemes and do not require any prior training. Furthermore, their small number of parameters ensure that they can be applied with comparable results to most types of audio materials.

It is believed that these schemes could have numerous applications, particularly for the coding of "very high-quality" sounds (like those issued from Super Audio CD, or DVD-Audio sources), for whom traditional coders are clearly sub-optimal. One of the main advantages of this method (transform coding associated with overcomplete representation spaces) is that they can retain a high sound quality even at high compression ratios without the need of complicated psychoacoustic models.

Many open questions have been mentioned in the body of the article and in the previous section. Future research will primarily focus on the choice of time-varying parameters, and on the construction of structures across adjacent windows for the selection and encoding of tonal components.

References

[1] R. Baraniuk. Optimal tree approximation using wavelets. In A. J. Aldroubi and M. Unser, editors, *Wavelet Applications in Signal Processing*, volume VII, pages 196–207. SPIE, 1999.

[2] L. Benaroya, R. Gribonval, and F. Bimbot. Représentations parcimonieuses pour la séparation de sources avec un seul capteur. In *Proc. 18th Symposium GRETSI'01 on Signal and Image Processing, Toulouse*, 2001.

[3] J. Berger, R. Coifman, and M. Goldberg. Removing noise from music using local trigonometric bases and wavelet packets. *J. Audio Eng. Soc.*, 42(10):808–818, 1994.

[4] K. Brandenburg. MP3 and AAC explained. In *Proc. of AES 17th international conference on High Quality Audio Coding (Florence)*, september 1999.

[5] R. Carmona, W.L. Hwang, and B. Torrésani. *Practical Time-Frequency Analysis: continuous wavelet and Gabor transforms, with an implementation in S*, volume 9 of *Wavelet Analysis and its Applications*. Academic Press, San Diego, 1998.

[6] S. Shaobing Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[7] A. Cohen, W. Dahmen, I. Daubechies, and R DeVore. Tree approximation and optimal encoding. *IGPM Report, RWTH Aachen*, sept 1999.

[8] A. Cohen, I. Daubechies, B. Jawerth, and P. Vial. Multiresolution analysis, wavelets and fast algorithms on an interval. *Comptes Rendus Acad. Sc. Paris*, 1992.

[9] I. Daubechies. *Ten lectures on wavelets*. SIAM, Philadelphia, PA, 1992.

[10] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.

[11] L. Daudet. *Représentations structurelles de signaux audiophoniques. Méthodes hybrides pour des applications à la compression*. PhD thesis, Université de Provence, 2000.

[12] L. Daudet, S. Molla, and B. Torrésani. Transient modeling and encoding using trees of wavelet coefficients. In *Proc. 18th Symposium GRETSI'01 on Signal and Image Processing, Toulouse*, September 2001.

[13] B. Delyon and A. Juditsky. On the computation of wavelet coefficients. *Journal of Approximation Theory*, 88(1):47–79, 1997.

[14] R.A. DeVore, B. Jawerth, and B.J. Lucier. Image compression through wavelet transform coding. *IEEE Trans. on Information Theory*, 2:719–746, 1992.

[15] R. Gribonval. *Approximations non-linéaires pour l'analyse des signaux sonores*. PhD thesis, Université de Paris IX Dauphine, 1999.

[16] K. Grochenig and S. Samarah. Non-linear approximation with local fourier bases. *Constr. Approx.*, 16:317–332, 2000.

[17] K. Hamdy, A. Tewfik, T. Chen, and S. Takagi. Time-scale modification of audio signals with combined harmonic and wavelet representations. In *Proc. IEEE Intern. Conf. Acoust., Speech, and Sig. Processing (ICASSP)*, 1997.

[18] K. N. Hamdy, A. Ali, and A. H. Tewfik. Low bit rate high quality audio coding with combined harmonic and wavelet representations. In *Proc. IEEE Intern. Conf. Acoust., Speech, and Sig. Processing (ICASSP)*, volume 2, pages 1045–1048, 1996.

[19] N. S. Jayant and P. Noll. *Digital coding of waveforms*. Prentice-Hall, 1984.

[20] S. Levine. *Audio Representations for Data Compression and Compressed Domain Processing*. PhD thesis, Stanford University, 1998.

[21] S. Mallat. *A wavelet tour on signal processing*. Academic Press, 1998.

[22] S. Mallat, G. Papanicolaou, and Z. Zhang. Adaptive covariance estimation of locally stationary processes. *Ann. Stat.*, 26(1):1–47, 1998.

[23] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[24] R.J. McAulay and Th.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Acoust., Speech and Signal Proc.*, 34:744–754, 1986.

[25] T. Painter and A. Spanias. Perceptual coding of digital audio. *Proc. IEEE*, 88(4), april 2000.

[26] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitionning in hierarchical trees. *IEEE Trans. on Circ. and Syst. for Video Tech*, 6(3):243–250, 1996.

[27] X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

[28] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing*, 41(12):3445–3462, 1993.

[29] D. Sinha and A. Tewfik. Low bit rate transparent audio compression using adapted wavelets. *IEEE Trans. Signal Processing*, 41(12):3463–3479, 1993.

[30] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.

[31] E. Terhardt. Calculating virtual pitch. *Hearing Research*, 1:155–182, 1979.

[32] P.P. Vaidyanathan. *Multirate systems and filter banks*. Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[33] P.P. Vaidyanathan and S. Akkrarakaran. A review of the theory and applications of optimal subband and transform coders. *Appl. and Comp. Harm. Anal.*, 10:254–289, 2001.

[34] T. Verma, S. Levine, and T. Meng. Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals. In *Proc. of the International Computer Music Conference*, Greece, 1997.

[35] T. Verma and T. Meng. Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2):47–59, 2000.

[36] M. Vetterli and J. Kovacevic. *Wavelets and subband coding*. Prentice Hall, Englewood Cliffs, NJ, USA, 1995.

[37] M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. AK Peters, Boston, MA, USA, 1994.

[38] E. Zwicker and H. Fastl. *Psychoacoustics, facts and models*. Springer Verlag, 1990.

Addresses:

L. Daudet
Dept. of Electronic Engineering,
Queen Mary, University of London,
Mile End Road,
London E1 4NS, UK.

B. Torrésani
LATP,
CMI, Université de Provence,
39 rue Joliot-Curie,
13453 Marseille Cedex 09, F.