Transient detection and encoding using wavelet coefficient trees

Laurent DAUDET[1,2], Stéphane MOLLA[1], Bruno TORRÉSANI[1]

[1]Laboratoire d'Analyse, Topologie et Probabilités
CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13

[2]Department of Electronic Engineering
King's College London, Strand, London WC2R 2LS, UK

{Laurent.Daudet,Stephane.Molla,Bruno.Torresani}@sophia.inria.fr

Résumé – Un grand nombre de problèmes de traitement du signal nécessitent des modèles précis et efficaces de signaux transitoires, et des algorithmes de détection et estimation correspondants. On propose dans cet article un cadre général pour des modèles de transitoires, basé sur des arbres dyadiques de coefficients d'ondelettes. Un modèle déterministe et un modèle stochastique sont présentés, et des algorithmes d'estimation correspondants sont décrits. Les résultats sont illustrés par des exemples numériques, dans un cadre de codage de signaux audio.

Abstract – Many signal processing problems call for accurate and efficient models for transient signals, and corresponding detection/estimation algorithms. This paper proposes a general setting for transient models, based upon dyadic trees of wavelet coefficients. A deterministic and a stochastic model are presented, and corresponding estimation algorithms are described. Numerical results are given in the framework of audio signal encoding.

1 Generalities

In this paper, we address the problem of transient detection and estimation in the context of (audio) signal encoding. This work is a part of a more general program on audio signal encoding, in which the input signal is decomposed into "tonal", "transient" and "stochastic" components, which are estimated and encoded separately. A more complete presentation of the scheme may be found in [5]. For the sake of the present discussion, let us simply stress that the scheme does not rely on any segmentation of the signal, unlike most approaches, for instance the approaches described in [11].

The performances of the global scheme turn out to depend heavily on its capabilities of separating correctly the three components. In the context of transform coding (see e.g. [6, 7]), and following ideas developed by Coifman and collaborators [2], it is tempting to use different transforms for estimating and encoding the different components, namely trigonometric bases (or, rather the smoothed versions of such bases) for the tonal part, and wavelet bases for the transients (see the discussion below): the large coefficients of the expansion with respect to the trigonometric basis are likely to "belong" to a tonal component, while the large wavelet coefficients are more easily interpreted as transients. However, such an approach is generally not sufficient, as a choice of basis alone is not enough to separate the components. A possible approach amounts to impose additional structure on the coefficients which are retained and encoded.

We focus here on the case of the transient part. Our approach relies on the fact that transient signals not only manifest themselves by (a small amount of) significant small scale wavelet coefficients, but the latter coefficients are "structured" in the time-scale space: a significant coefficient is likely to be accompanied by additional significant coefficients at the same location and coarser scales. We use this remark as our definition for transients, which are therefore associated with incomplete dyadic trees of wavelet coefficients.

We describe and compare here two different methods for estimating such transients on 1D signals. We also present numerical illustrations on audio signals.

2 Transients and trees of wavelet coefficients

The notion of transient signal, although heuristically clear, is difficult to define in precise mathematical terms. In general, most transient detection algorithms rely (at least implicitely) on *a priori* models for transients. Note that in an audio processing context, "transients" are usually restricted to note onsets, or "attacks". Our definition will here be broader, as to include any well- localized feature. Wavelet-based methods have often been used for singularity characterization and transient detection and modelling [8]. It is well known that wavelet methods are extremely efficient for characterizing localized features in signals, because of their capability of "zooming" in particular regions in signals. In addition, they provide precise characterization of singularities in functions, essentially through the behavior of the wavelet coefficient magnitude accross scales. Even though is does not make sense to

model transients strictly speaking as mathematical singularities, the rate of decay of wavelet coefficients across scales provides useful information on the "local strength" of the signal. However, such a characteristics is a property of a family of coefficients rather than a property of individual coefficients. Motivated by recent algorithms using trees-structured decompositions (see for example [10] and [4]), we focus our analysis on wavelet coefficients trees.

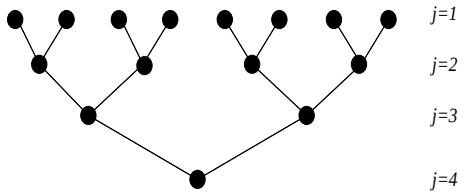

FIG. 1: Wavelet coefficients tree.

To sketch our notations, we work in the framework of 1D multiresolution wavelet transform. Let ψ denote a compactly supported wavelet, and set as usual $\psi_{jk}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right)$. Denote by

$$d_{jk} = \langle f, \psi_{jk} \rangle \ , \quad j = 1, 2, \dots J$$

the wavelet coefficients of a signal $f \in L2(\mathbb{R})$. The latter are obtained via a sub-band coding algorithm (see [7] for details), and are naturally associated with a dyadic tree structure: each coefficient d_{jk} at scale j has two children $d_{j+1,2k}$ and $d_{j+1,2k+1}$ at scale $j+1$ (see FIG. 1). According to the common practice, the samples f_k of the input signal are identified with small scale scaling function coefficients, and we consider wavelet expansions of the form

$$\langle f, \phi_{0k} \rangle \approx f_k \ , \quad f \approx \sum_k s_k \phi_{Jk} + \sum_{j=1}^{J} \sum_k d_{jk} \psi_{jk} \ .$$

A tree of wavelet coefficients is termed *admissible* if for every node of the tree, its parent node belongs to the tree. Given an admissible tree and a leave, the union of edges connecting the root to the considered leave is called a *connected branch*. A *full branch* is a connected branch whose nodes have at least one child. Therefore, the leaves of full branches correspond to the minimal considered scale.

We define transients from associated trees of *relevant* coefficients: *a transient structure is a connected tree of wavelet coefficients which satisfies a given relevance property.* The latter property characterizes the transient model under consideration. Two examples are given below.

2.1 Deterministic model

Our first model is strongly inspired by the discussion of [4], and rests on local regularity estimates from wavelet coefficients. We first consider only trees consisting of full branches. Such a 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

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

exceeds a fixed threshold ($|\mathcal{B}|$denotes the length of the branch $\mathcal{B}$). The choice of assigning a cost to full branches of the complete tree is motivated by the will of assigning a *local* feature to a transient: the leaves of full branches are naturally associated with samples in the signals.

The constants s, p characterize the type of transients considered, in the sense that they weight coefficients corresponding to different scales. 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. The choice of p also influences the type of transients to be estimated.

2.2 Stochastic model

The second approach models wavelet coefficients as random variables, distributed according to a mixture of two different distributions (two states). Transitions from a state to another are governed by (hidden) Markov chains. Such hidden Markov trees have been introduced and studied by Baraniuk and coworkers [1] in the context of signal and image modelling and denoising. The model is adapted to our problem as follows. The wavelet coefficients $d_{j,k}$ are "emitted" by random variables Y_{jk}, whose distribution depends on a hidden state $X_{jk} \in \{T, O\}$ (T stands for "transient", and O for "other"). At each scale j, the T-type coefficients are the ones which belong to a transient structure, and are modelled by a centered distribution with a large variance $\sigma^2_{T,j}$. The O-type coefficients are modelled by a centered distribution with a small variance $\sigma^2_{O,j}$. For the sake of simplicity, we limit ourselves here to normal distributions. Such a choice is compatible with the choice $p = 2$ in the previous section, as argued in [3].

The distribution of hidden states is given by a "bottom-up" Markov chain, characterized by a 2×2 transition matrix, and the distribution of the coarsest scale state. In order to retain only connected trees, we impose a taboo transition: the transition $O \to T$ is forbidden. Therefore, the transition matrix assumes the form

$$\Pi = \begin{pmatrix} \pi & 1-\pi \\ 0 & 1 \end{pmatrix}$$

where p denotes the probability of transition $T \to O$:

$$\pi = \mathbb{P}\left\{X_{j-1,\ell} = T | X_{j,k} = T\right\} \ , \ell = 2k, 2k+1 \ .$$

The hidden Markov process is completely determined by the matrix Π and the "initial" probability distribution, namely the probabilities $\nu = (\nu_T, \nu_O)$ of states at the maximum scale J. The complete model is therefore characterized by Π, ν, and the emission probability densities:

$$\rho_S(y) = \rho(y | X = S) \ , \quad S = T, O \ .$$

According to our choice (centered Gaussian distributions), the latter are completely characterized by their variances $\sigma^2_{T,j}$ and $\sigma^2_{O,j}$.

3 Tree Estimation

We describe two estimation procedures for wavelet coefficient trees, based upon the two above models. Example of corresponding trees may be found in the next Section.

3.1 Deterministic model

The first approach considers only trees with "full length" branches (i.e. the leaves always correspond to the finest considered scales), and is a "top-down" algorithm. Starting from a time index ℓ, consider the branch of all its ancestors, denoted by $\mathcal{D}_\ell$, and form

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

where s, p 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 $p = 2$.

The pruning is done by retaining the leaves ℓ whose modulus of regularity $\kappa_{p,s}[\ell]$ exceeds a threshold value $\tilde{\kappa}[\ell]$, which is to be adapted locally. All the wavelet coefficients belonging to a retained branch are encoded. The threshold value has to be estimated within a time frame larger than the time frame defined by the complete tree. More details on the estimation of threshold values may be found in [5].

3.2 Hidden Markov Trees

As stressed before, we limit our investigations here to the case of Gaussian mixture models. Therefore, the emission probability distributions are completely characterized by the variances $\sigma^2_{T,j}$ and $\sigma^2_{O,j}$. The estimation of the parameters of the model (the variances, and the transition probability matrix Π and the large scale distribution ν for the hidden states) may be performed using a standard Baum-Welsh `EM` algorithm (see chap. 6 of [9] for a comprehensive account of Hidden Markov models, and [1] for the adaptation to the Hidden Markov Tree situation.)

The hidden states are usually estimated via an adapted version of the Viterbi algorithm. Unfortunately, the adaptation of the latter to HMT models is quite difficult, because of an increased complexity (at scale j, the number of configurations goes like $2^{2^{J-j}}$, with $j = 1, \ldots J$ and J typically equals 10), and we have to limit to local estimations. The tree (i.e. the connected set of "T-class" wavelet coefficients to be retained for encoding) is characterized by its leaves. The latter are selected when their posterior probability exceeds a given threshold value.

An important point is the fact that the EM parameter estimation has to be performed on a more global basis than the tree estimation (such a procedure is called "tying" in [1]). Otherwise, the algorithm has a tendancy to detect transients in all time frames, which is quite inadequate. In the simulations presented below, a tree is estimated within sets of 6 time frames (6×1024 samples long), while the parameters are estimated in 10 consecutive frames simultaneously.

4 Results and discussion

4.1 Results

We illustrate our approach with a signal obtained from an audio signal after estimating and removing a tonal component (see [5] for details); therefore the signal contains essentially transients and a residual. In some sense, the goal of transient estimation in such a context is to obtain a residual part as close as possible to a weakly stationary random signal, with the smallest possible variance.

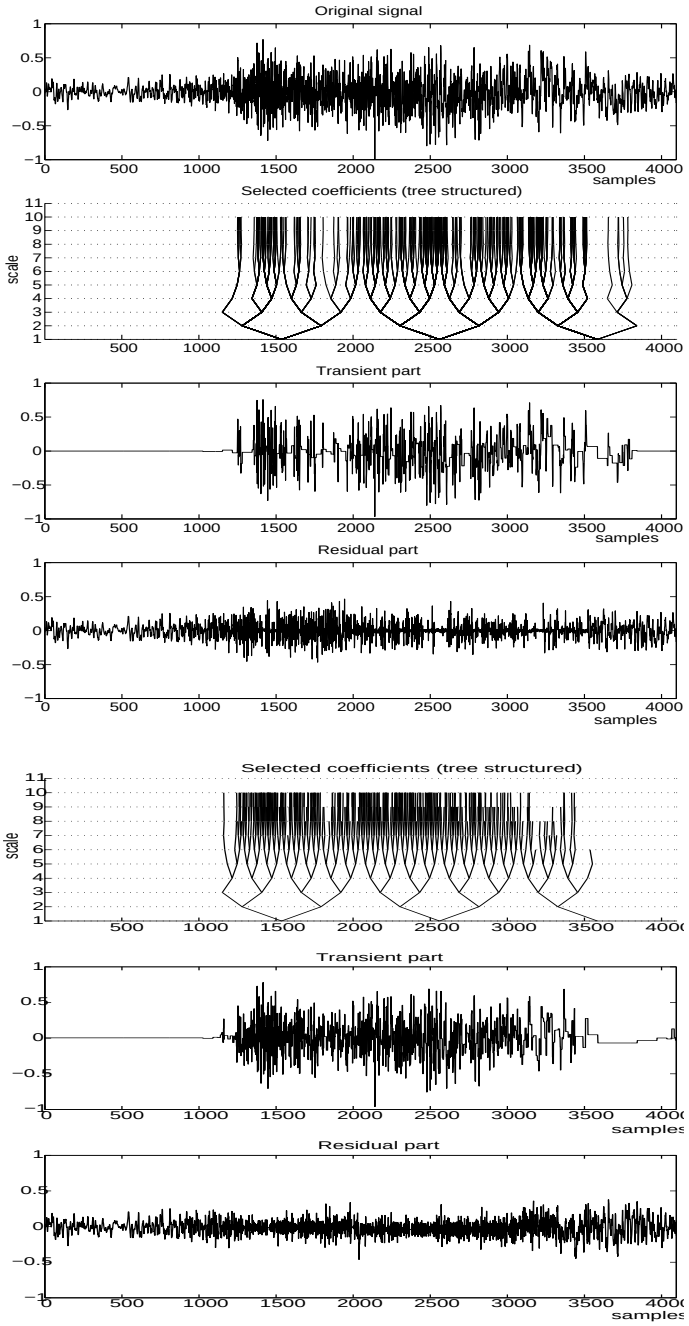

FIG. 2: Estimation of transients in audio signal: from top to bottom, input signal (with "tonal part removed"), tree of "transient-type" wavelet coefficients, estimated transient component, and residual (deterministic model); tree, estimated transient, and residual (stochastic model).

As an illustration, we show in FIG. 2 the transient estimates for about 4096 samples (center of the figure) of audio signal (a jazz recording), sampled at 44100 Hz, with the two algorithms (top: deterministic; bottom: stochastic). The chosen frame is interesting as it contains an "attack". In both cases, parameters are tuned so that about 25 percent of the coefficients are retained. As may be seen,

the attack is fairly well detected and approximated, and the residual exhibits much less "local" structures than the original signal. In other words, the residual signal is easier to model as a weakly stationary random signal with small variance, and to estimate and encode as such.

The difference between the results of the two methods is easily understood: the deterministic model imposes full branches, which results on a smaller number of branches and therefore a slightly more "lacunary" structure in the estimated transients. However, more care is needed in the interpretation as the parameters of the two models play a significant role in the type of transients to be estimated. A more detailed analysis is under progress.

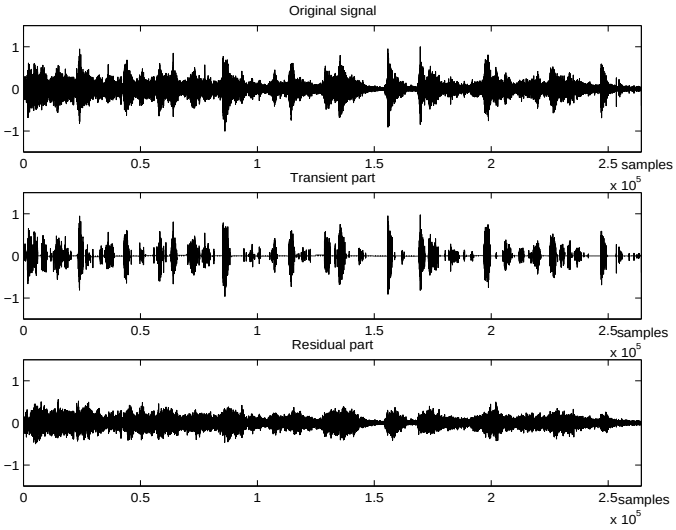

FIG. 3: HMT estimation of transients in audio signal, complete signal decomposition

The transient estimates (using the HMT algorithm) on a much larger signal sample is displayed in FIG. 3. As may be seen from the figure, the algorithm succeeds to capture the several "attacks" of the signal, but somehow fails to turn the residual into a stationary signal (in the middle part). This shows that the simple estimation procedure described here (in particular the choice of the "training frames" for the EM algorithm) has to be optimized further.

The (sound and ascii) files corresponding to the figures of this paper are publicly available at
`http://www-sop.inria.fr/sysdys/personnel/smolla/`

4.2 Perspectives

The discussion of the present paper was essentially done in a signal coding context. The advantage of encoding trees of wavelet coefficients is obvious, as it avoids run length encoding of addresses of significant coefficients.

However, the models "tonal + transient + noise" and corresponding transient estimation methods are likely to yield several additional interesting applications. Among them, let us simply quote signal modifications:

- Time-stretching without pitch modifications (or the dual problem pitch shifting without time modifications) The most efficient methods usually rely on the duplication of one elementary waveform (the sound over one period) every N periods. However, duplicating waveforms containing transients results in the smearing of attacks. Our transient detection technique would prevent such artifacts, by allowing duplication only when no transient is present.
- Attack enhancements: the resulting sound can be reconstructed using a weighted sum for the transient and tonal part, thus leading to the enhancement (or attenuation) of the transient information, which is essentially perceived as attacks.

Acknowledgements

We wish to thank the SYSDYS project (INRIA) for support and hospitality. L. Daudet is supported by the European Union Community programme *Human Potential* under contract number HPMF-CT-2000-00917. S. Molla is supported by a MENRT grant (French government).

References

[1] R. Baraniuk (1999). Optimal tree Approximation using Wavelets, Proceedings of SPIE International Conference on Wavelet Applications in Signal Processing VII, Denver, pp. 196-207.

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

[3] H. Choi and R. Baraniuk (2000). Information theoretic Interpretation of Besov spaces. SPIE International Conference on Wavelet Applications in Signal Processing VIII, San Diego.

[4] A. Cohen, I. Daubechies, W. Dahmen, and R. deVore (1999). Tree approximation and optimal encoding. Technical report, Institut für Geometrie und Praktische Mathematik, Bericht Nr. 174, Aachen.

[5] L. Daudet (2000). *Représentations structurelles de signaux audiophoniques: méthodes hybrides pour des applications à la compression*, PhD Thesis, Marseille.

[6] N.S. Jayant and P. Noll (1984), *On the digital coding of waveforms*, Prentice Hall, Englewood Cliffs, NJ, USA.

[7] S. Mallat (1998). *A wavelet tour of signal processing*, Academic Press, San Diego, CA, USA.

[8] S. Mallat, W.L. Hwang (1992). Singularity Detection and Processing with Wavelets, *IEEE. Trans. Inf. Th.* **38**, pp. 617-643.

[9] L. Rabiner and B.H. Juang (1993). *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, A. Oppenheim Ed.

[10] J.M. Shapiro (1993). Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing* **41**, pp. 3445-3462.

[11] T. Verma (2000). *A Perceptually Based Audio Signal Model With Application to Scalable Audio Compression*, PhD Thesis, Stanford University.