Hidden Markov Tree Based Transient Estimation for Audio Coding

S Molla, B. Torrésani

HIDDEN MARKOV TREE BASED TRANSIENT ESTIMATION FOR AUDIO CODING

S. MOLLA and B. TORRESANI

LATP / CMI
39 rue Joliot Curie, 13453 Marseille Cedex 13, France

ABSTRACT

A new approach for transients detection and estimation in the context of hybrid audio coding is presented. The basic idea is to use an orthogonal dyadic wavelet expansion, followed by Hidden Markov Tree modeling of wavelet coefficients. Coefficients may be cast as "transient type" or "residual type", and the estimated transient is reconstructed from the transient type coefficients only. The estimation procedure involves the classical two steps of Hidden Markov Models: parameters estimation and state estimation. The implementation of those two steps in the case of wavelet coefficient trees is discussed in some details, and numerical results are given. The application to audio signal encoding is also discussed.

1. INTRODUCTION

So-called Hybrid models have received an increasing attention in recent developments on audio coding. The underlying idea is to model and encode separately different features of the signal, such as tonals, transients,... Such a scheme was developed recently in [1, 2], based upon the following idea: a tonal part is first estimated (using thresholded and weighted MCDT coefficients), encoded, and substracted from the signal. Then the transient part is estimated in a similar way using a wavelet expansion of the residual. It has been shown [2] that considerable savings are obtained if a "structured" wavelet expansion is used. Namely, rather than encoding only significant wavelet coefficients (i.e. coefficients larger than some given threshold), which forces to encode a significance map as well, model transients as sets of wavelet coefficients located at the nodes of rooted connected trees. Imposing such tree structures has two main benefits: the transient estimation is more robust, and the tree structure turns out to be much more efficient for encoding the significance map.

We discuss here in more details the probabilistic model for tree structured transient estimations proposed in [1]. Our approach is based upon the wavelet Hidden Markov Tree model (HMT) proposed by Crouse and coworkers [3], and implements a corresponding "transient + residual" model. As usual with Hidden Markov Models, the main two problems to solve are parameter estimation and state estimation, and we study several strategies for achieving these goals.

2. DESCRIPTION OF THE MODEL AND THE ESTIMATION ALGORITHM

The model we use defines transients as "chained families" of significant wavelet coefficients. More precisely, let ψ be a compactly supported wavelet, and set as usual $\psi_{jk}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right)$. Denote by $d_{jk} = \langle x, \psi_{jk}\rangle$, $j = 1, 2, \dots J$ the wavelet coefficients of a signal $x \in L^2(\mathbb{R})$. The latter are obtained via a sub-band coding algorithm (see [4] 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 $j-1$. According to the common practice, the samples x_k of the input signal are identified with small scale scaling function coefficients, and we consider wavelet expansions of the form

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

For the sake of coding efficiency, we only consider rooted connected trees. We define transients from associated trees of *relevant* coefficients [1]: *a transient structure is a connected tree of wavelet coefficients which satisfies a given relevance property.* In this work, the relevance property is defined in probabilistic terms. The model we use follows quite closely the approach described in [3] in the context of signal and image modeling and denoising. The starting point is a Hidden Markov Tree model, associating a random variable D_{jk} to each node (of coordinates (j,k), j being the level and k the position) of a fixed binary tree. The distribution of the random variables D_{jk} depends on a hidden state $X_{jk} \in \{T, R\}$ (the "transient" state T and the "residual" state R.) At each scale j, T-type coefficients follow a centered normal[1] distribution with a large variance $\sigma^2_{T,j}$. R-type coefficients follow a centered normal distribution with

[1] For the sake of simplicity, we limit ourselves to normal distributions, but the model can accomodate arbitrary distributions. Gaussian mixtures already represent quite a large family of models.

a small variance $\sigma^2_{R,j}$. The distribution of hidden states is given by a coarse-to-fine Markov chain, characterized by a 2×2 transition matrix, and the distribution of the coarsest scale state. In order to retain only connected trees of T-nodes, the transition $R \to T$ is forbidden: a "residual" type coefficient cannot have "transient" type children. Thus, the transition matrix Π takes the form

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

where π denotes the probability of transition $T \to R$:

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

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

$$f_S(y) = f(y|X = S) \ , \quad S = T, R \ .$$

According to our choice (centered Gaussian distributions), the latter are completely characterized by $\sigma_{T,j}$ and $\sigma_{R,j}$.

Given a realization of this process, the corresponding "transient" part of the so-generated signal is the wavelet synthesis from T-type coefficients:

$$x_{tr} = \sum_{j,k; X_{jk} = T} d_{jk} \psi_{jk}$$

2.1. Parameter estimation

The first step of the algorithm is the estimation of the parameters of the model: the transition matrix Π, and the standard deviations $\sigma_{T,j}$ and $\sigma_{R,j}$. The procedure is a maximum likelihood approach: find the parameters values which maximize the log-likelihood

$$\mathcal{L}(\theta) = \log \mathbb{P}\left\{D | \Theta\right\}$$

where D represents the set of wavelet coefficients under consideration, and Θ represents the collection of parameters. This estimation is done using a traditional EM (expectation-maximization, see [5] for an introduction) algorithm.

After an Upward-Downward pass of the algorithm, we are able to compute conditional probabilities $\mathbb{P}\left\{X_i = S \,|\mathcal{T}_1\right\}$ and $\mathbb{P}\left\{X_i = s_1, \, X_{\rho(i)} = s_2 | \, \mathcal{T}_1\right\}$ (for $S, s_1, s_2 = T, R$), available for the current set of parameters Θ, from which we re-estimate the parameters (denoted by a hat) as follows[2]:

$$\hat{\pi} \ = \ \frac{\frac{1}{N-1} \sum_{k=2}^{N} \mathbb{P}\left\{X_k = R, \, X_{\rho(k)} = T \,|\, \mathcal{T}_1\right\}}{\frac{1}{N-1} \sum_{k=2}^{N} \mathbb{P}\left\{X_{\rho(k)} = T \,|\, \mathcal{T}_1\right\}}$$

[2]Below, $\mu_{S,j} = 0, \ \forall S, \ \forall j$, because coefficients $\{d_k\}$ results from a wavelet decomposition.

$$\hat{\sigma}^2_{S,j} \ = \ \frac{\frac{1}{\#(scale\, j)} \sum_{k \in j} (d_k - \mu_{S,j})^2 \, \mathbb{P}\left\{X_k = S \,|\, \mathcal{T}_1\right\}}{\frac{1}{\#(scale\, j)} \sum_{k \in j} \mathbb{P}\left\{X_k = S \,|\, \mathcal{T}_1\right\}}$$

where $j = 1 \ldots J$ and $S = T, R$. Here, J denotes the depth of the tree resulting from a full dyadic wavelet decomposition over a fixed 2^J samples large frame. To lighten the notations, we denoted previously $X_{j,\ell}$ by X_i, where i is the number of the node (j, ℓ). The tree is numbered ordinally (*i.e.* 1 stands for the root $(J, 1)$, 2 and 3 are its children, and so on), $\rho(i)$ is the ancestor of i, and $c(i)$ its set of two children. Moreover, $\mathcal{T}_i$ represents the subtree of wavelet coefficients rooted at node i (coefficient d_i), and $\mathcal{T}_{i \setminus k}$ the set of nodes in $\mathcal{T}_i$ which are not in $\mathcal{T}_k$ (where $k > i$).

2.2. State estimation

The second step is the estimation of the states: this is needed in order to decide whether a given coefficient belongs to a transient structure or to the residual. This is the most difficult part, as one has to find an optimum over an extremely large configuration space: all together, the number of configurations for a frame of $N = 2^J$ coefficients is of the order of 2^N. In the classical HMC situation, the Viterbi algorithm provides an efficient way of dealing with the optimization problem. In the HMT case, the situation is more complex, because of the combinatorial explosion. Nevertheless, an adapted version of the Viterbi algorithm has recently been proposed [6], which applies to our case. This version of the EM algorithm reverses the roles of the upward and downward variables using adapted conditional variables and therefore runs the down step before the up step. Moreover, its computation is less subject to numerical limitations (in particular, a scaling factor avoids underflow problems).

We invite the reader to refer to [6] for further details about this algorithm (inspired by Devidjer's "conditional forward-backward" recursion (1985)).

2.3. A few remarks

The computation of such an algorithm in the context of transient detection in audio signals leads to consider an amount of parameters which have to be chosen adequately.

First of all, the use of a wavelet function with few vanishing moments seems adapted to our case due to the well-localization of transients in time and frequency. Typically, for a signal sampled at $44100 Hz$, analysis is done through 1024 sample long frames (*i.e.* $23.2\, ms$ width), corresponding to 10 scales depth trees, a rather sufficient depth to capture transient-like behavior. An important point is that, according to the model, the algorithm has a tendancy to detect transients in every frame, because of the "local" feature of the parameters. In other words, transients selection is only based on the current information whatever the behaviour of the signal in the neighborhood is. A more global way of

re-estimation is thus needed. We describe here two possible approaches.

On one hand, "global" parameters may be estimated by considering each local ones in a set of N frames. For instance, a mean of these local parameters over this set can be taken to compute the state estimation, or whatever multiple of this mean (this introduces an extra parameter).

The alternative, much less intuitive, is to globally and simultaneously re-estimate parameters in each frame of the N-set, modifying section 2.1 formulae by considering the whole set's conditional probabilities computed at each loop EM (known as "tying" across wavelet trees). The EM algorithm thus converges towards the set of global parameters Θ_G. The choice of the method used to reach Θ_G is discussed below. Hence, a wavelet decomposition is done on a $N \times 1024$ samples long segment of the signal, limited to 10 scales depth, which leads to N wavelet coefficient trees, and then our algorithm can be applied to these trees.

Moreover, to avoid edge effects in the decomposition, only $N-2$ center frames are used for the re-estimation, while previous and next selected coefficients in extremal frames are taken for resynthesis, and so this set of frames is slided of $N-2$ frames between two detections to permit an overlapping. More generally, we can consider an overlap set of N_0 frames, and thus a sliding over $N - N_0$ frames.

3. RESULTS ON TRANSIENT ESTIMATION

An example of tree estimation may be found in Figure 1, where a dyadic tree and a corresponding signal have been generated from a HMT model with a fixed set of parameters. The estimated tree and signal appear to be remarkably close to the original signal, except for a few glitchs, caused by "missed" branches.

We now discuss the performance of this algorithm on an audio signal [3] with various sets of parameters (essentially size of frames sets and way of re-estimate parameters).

Sound files and additional material related to this paper can be found at the following URL:
`http://www.cmi.univ-mrs.fr/~molla/GTS/index.htm`

We consider below two ways in re-restimating parameters Θ: the "global" one, in section 2.3 (tree tying), and the other we will call "local mean": it consists in taking a L_p mean to re-estimate standard deviations scale by scale,

$$\hat{\sigma}_{S,j} = \sqrt[p]{\sum_{k=1}^{N} \hat{\sigma}^p_{S,j}(k)/N}$$

where $\hat{\sigma}_{S,j}(k)$ is the estimated standard deviation of hidden sate S in the jth scale of the kth frame of the current set. Obviously, N still denote the size of a set of frames.

[3] a tonal part has been estimated and then removed from the original signal in order to obtain the signal with which we work (see [2] for details).

Note that setting p to 1 is the same as taking the arithmetic mean of the standard deviations across the frames, and raising p emphasizes the influence of large values.

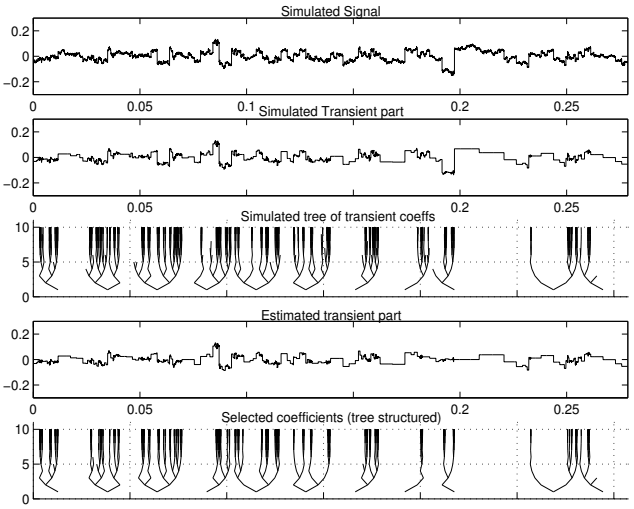

Figure 1: From top to bottom: Simulated signal and transient part for a given set of parameters, with corresponding tree. Estimated signal and trees of coefficients with a "local mean" model.

The role of the differents settings is illustrated in Figs 3-4. An audio signal of about 60000 samples long, sampled at $44.1\,kHz$ is used for these tests.

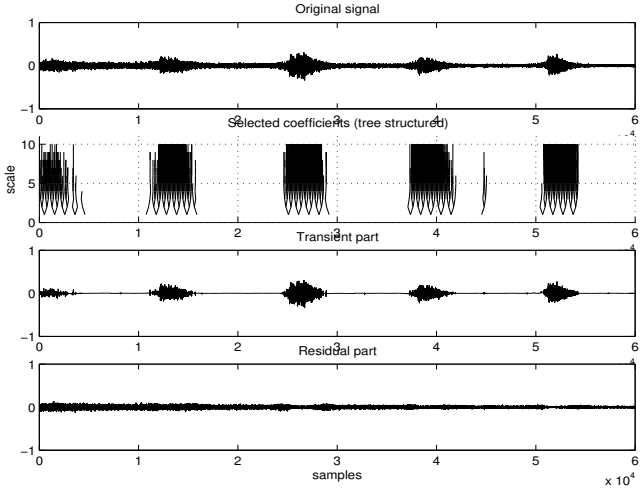

Figure 2: "Global" parameters estimation, with $N = 10$.

We can see in FIG. 3 how the choice of p can influence the behaviour of the algorithm. For $p = 1$, the transients features are pretty well selected while for $p = 2$ selected coefficients are concentrated in the "attacks", due to the importance given to strong standard deviations, and thus to strongest wavelet coefficients, and do not select all the

expected coefficients. However, for much more localized transients (such as short attacks), such a setting would yield better results.

This problem naturally leads us to consider the "global" parameter re-estimation evoked before. FIG. 2 shows that it gives excellents results, in terms of transient estimation. Therefore, the residual part after removal of the estimated transients is closer to a stationary random signal, with a rather small variance, which is one of the main goals [1, 2].

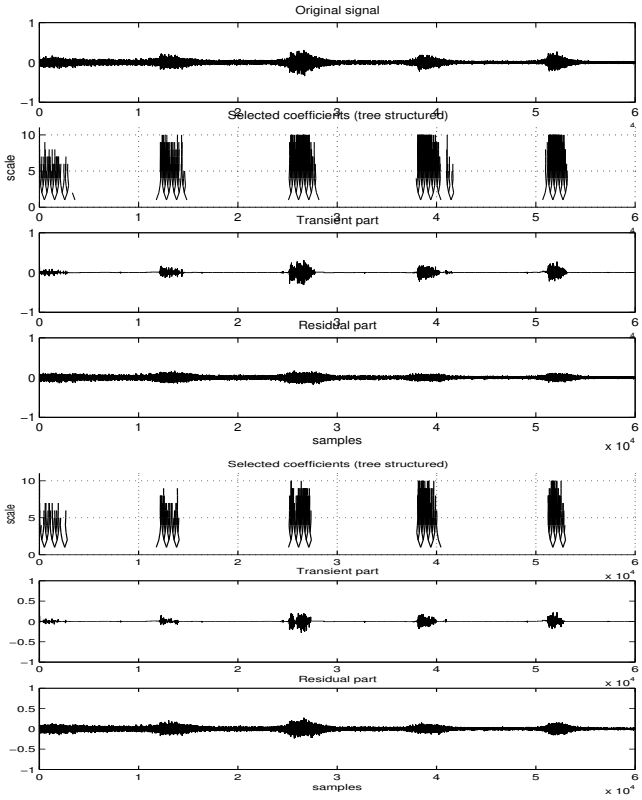

Figure 3: "Local mean" transient estimation with different values of p, with $N = 10$. From top to bottom: input signal, tree of estimated T wavelet coefficients, estimated transient component, residual (for $p = 1$); tree, estimated, residual ($p = 2$).

However, this improved precision in the transient estimation goes together with an increased number of retained wavelet coefficients, which may become a severe shortcoming in a signal coding perspective.

The optimization of the performances of this algorithm would also require optimizing the number N of frames where parameters are re-estimated. For instance, selecting the few tree branches near the 45000th sample in FIG. 2 would easily be avoided by another choice of N. Nevertheless, let us stress that the optimization of N may become an important issue, as shown in FIG. 4, where an inadequate N yields a very poor transient estimates. In general, if N is too large, some transients hidden by others within the same set are ignored. If N is too small, "ghost" transients show up.

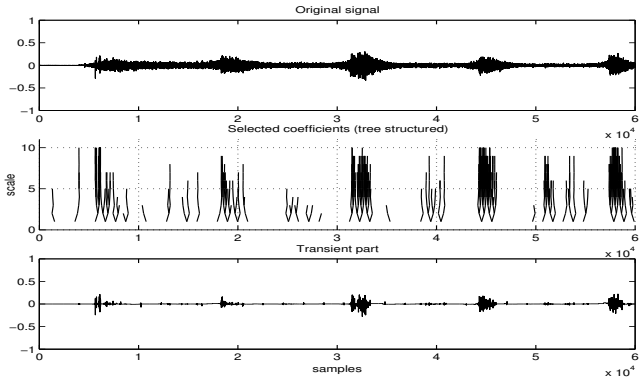

Figure 4: Transient estimation with $N = 4$, $p = 1$.

4. CONCLUSION

We presented in this paper an application to audiophonic signals of HMT models in the wavelet-domain. We showed that its adaptation to transient estimation gives relevant results if an optimization upon the parameters of the underlying model is done. This optimization will be useful to include this algorithm in a more global framework of hybrid audio signal encoding.

5. REFERENCES

[1] 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.

[2] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, 2001, Special issue on Image and Video Coding Beyond Standards.

[3] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based signal processing using hidden Markov models," *IEEE Transactions on Signal Processing*, 1998, Special Issue on Filter Banks.

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

[5] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[6] J.B. Durand and P. Gonçalves, "Statistical inference for hidden markov tree models and application to wavelet trees," Tech. Report 4248, INRIA, September 2001.