Determining local transientness of audio signals

Stéphane Molla and Bruno Torrésani

***Abstract*— We describe a new method for estimating the degree of "transientness" and "tonality" of a class of compound signals involving simultaneously transient and harmonic features. The key assumption is that these two layers admit sparse expansions in wavelet and local cosine bases. The estimation is performed using particular form of entropy (or theoretical dimension) functions. We provide theoretical estimates on the behavior of the proposed indices, as well as numerical simulations. Audio signals provide a natural field of application.**

***Index Terms*— audiophonic signal, transient, tonal, wavelet basis, local Fourier basis, sparsity.**
EDICS: 1.TFSR, 2.AUEA

I. Introduction

Many generic signal classes feature significantly different "components", such as transients, (locally) sinusoidal or harmonic "partials", or stochastic-like components in sounds, or edges, textures,... in images. Detecting the presence of such components is one of the classical signal processing problems. Another interesting problem is to estimate whether a given portion of signal is for example more transient than harmonic or periodic, or in other words to estimate "transientness" or "tonality" indices. This finds immediate applications in several contexts, including the hybrid signal coders [4], [8], [15] which use different methods for encoding transient or tonal regions (and were the main motivation of this work), more general purpose hybrid models [1], or similar recent ideas in image coding [10], [13]. We propose here simple criteria, based on transform coding ideas, for estimating such indices. The main idea is to use orthonormal bases in signal spaces which are significantly different from each other in the following sense: a given component has a sparse expansion in a given basis, while the others have dense expansions. Information theoretic criteria (we elaborate on the case of a variant of Shannon's entropy) therefore yield estimates for the indices.

We focus here on the case of transient and locally sinusoidal (or harmonic) layers in audio signals, using wavelet and local cosine bases. However, the approach we develop may be adapted to different signal layers (chirps for example), or in higher dimensions. We provide theoretical estimates for the behavior of transientness and tonality indices, and illustrate our results by numerical simulations and tests on real sounds.

II. A model for sparse audio signals

We focus on the particular application to audio signals, and limit ourselves to transient and tonal features. Our starting point is the assumption that transient signals have a sparse expansion in a wavelet basis (provided the wavelets have small enough support), and that tonals have sparse expansion in local cosine basis (with smooth enough window function). We are naturally led to consider a generic redundant "dictionary" made out of two such orthonormal bases, denoted by ψ_λ and w_δ respectively (we refer to [2], [9], [16] for detailed tutorials), and signal expansions of the form

$$x = \sum_{\lambda\in\Lambda} \alpha_\lambda \psi_\lambda + \sum_{\delta\in\Delta} \beta_\delta w_\delta + r \ , \qquad (1)$$

where Λ and Δ are (small, and this will be the main *sparsity* assumption) subsets of the index sets, termed *significance maps*. The nonzero coefficients α_λ are independent $\mathcal{N}(0,\sigma_\lambda)$ random variables, and the nonzero coefficients β_δ are independent $\mathcal{N}(0,\tilde\sigma_\delta)$ random variables: r is a residual signal, which is not sparse with respect to the two considered bases (we shall talk of *spread residual*), and is to be neglected or described differently.

Given a signal assumed for simplicity to be of the form (1), with unknown values of $|\Delta|$ and $|\Lambda|$, we are interested in finding estimates for the latter, or at least for the "transientness" and "tonality" indices

$$I_{ton} = \frac{|\Delta|}{|\Delta|+|\Lambda|} \ ; \qquad I_{tr} = \frac{|\Lambda|}{|\Delta|+|\Lambda|} \ . \qquad (2)$$

We propose a procedure close to the notions of *theoretical dimension* or α-entropies, advocated by M.V. Wickerhauser [17] and coworkers, which may in some situations be shown to be closely connected to the notion of Shannon entropy [14].

For the sake of simplicity, we shall work in this section in a finite dimensional context.

Definition 1: Given an orthonormal basis $\mathcal{B} = \{e_n, n \in S\}$ of a given N-dimensional signal space $\mathcal{E}$, define the logarithmic dimension of $x \in \mathcal{E}$ in the basis $\mathcal{B}$ by

$$\mathcal{D}_\mathcal{B}(x) = \frac{1}{N} \sum_{n\in S} \log_2 \left(|\langle x, e_n\rangle|^2 \right) \qquad (3)$$

It follows from a simple calculation that in the framework of the signal models under consideration,

Lemma 1: Given an orthonormal basis $\mathcal{B} = \{e_n, n \in S\}$, assuming that the coefficients $\langle x, e_n\rangle$ are $\mathcal{N}(0,\sigma_n)$ random variables, one has

$$\mathbb{E}\left\{\mathcal{D}_\mathcal{B}(x)\right\} = C + \frac{1}{N} \sum_{n\in S} \log_2(\sigma_n^2) \qquad (4)$$

where $C = 1 + \gamma/\ln(2)$ is a universal constant ($\gamma \approx .5772156649$ being Euler's constant.)

Returning to the model (1), and assuming that the coefficients $\alpha_\lambda, \lambda \in \Lambda$ and $\beta_\delta, \delta \in \Delta$ are respectively $\mathcal{N}(0,\sigma_\lambda)$ and $\mathcal{N}(0,\tilde\sigma_\delta)$ independent random variables, the coefficients

$$a_\lambda = \langle x, \psi_\lambda\rangle \ ; \quad b_\delta = \langle x, w_\delta\rangle \ ,$$

Both authors are with LATP, CMI, 39 rue Joliot-Curie, 13453 Marseille cedex 13, France; email: molla@cmi.univ-mrs.fr; torresan@cmi.univ-mrs.fr.
Work supported in part by the European Union's Human Potential Programme, under contract HPRN-CT-2002-00285 (HASSIP)

are centered normal random variables, whose variance depends on whether $\lambda \in \Lambda$ (or $\delta \in \Delta$) or not. For example,

$$\mathrm{var}\{a_\lambda\} = \begin{cases} \sigma_\lambda^2 + \sum_{\delta\in\Delta} \tilde\sigma_\delta^2 |\langle x, w_\delta\rangle|^2 & \text{if } \lambda \in \Lambda \\ \sum_{\delta\in\Delta} \tilde\sigma_\delta^2 |\langle x, w_\delta\rangle|^2 & \text{if } \lambda \notin \Lambda \end{cases} \quad (5)$$

we obtain, for the $\Psi = \{\psi_\lambda\}$ basis

$$\mathbb{E}\{\mathcal{D}_\Psi(x)\} = C + \frac{1}{N}\log_2\Big[\prod_{\lambda\in\Lambda}\Big(\sigma_\lambda^2 + \sum_{\delta\in\Delta}\tilde\sigma_\delta^2|\langle\psi_\lambda, w_\delta\rangle|^2\Big) \times \prod_{\lambda'\notin\Lambda}\Big(\sum_{\delta\in\Delta}\tilde\sigma_\delta^2|\langle\psi_{\lambda'}, w_\delta\rangle|^2\Big)\Big] , \quad (6)$$

and a similar expression for the logarithmic dimension $\mathcal{D}_W(x)$ with respect to the $W = \{w_\delta\}$ basis.

In the simpler case where $\sigma_\lambda = \sigma$, $\forall\lambda \in \Lambda$ and $\tilde\sigma_\delta = \tilde\sigma$, $\forall\delta \in \Delta$, we introduce the Parseval weights

$$p_\lambda(\Delta) = \sum_{\delta\in\Delta} |\langle w_\delta, \psi_\lambda\rangle|^2 , \quad \tilde p_\delta(\Lambda) = \sum_{\lambda\in\Lambda} |\langle w_\delta, \psi_\lambda\rangle|^2 , \quad (7)$$

which satisfy the following property, which is an immediate consequence of Parseval's formula: for all f, $\sum_\lambda |\langle f, \psi_\lambda\rangle|^2 = \|f\|^2$.

Lemma 2: With the above notations, the Parseval weights satisfy

$$0 \le p_\lambda(\Delta) \le 1 , \quad 0 \le \tilde p_\delta(\Lambda) \le 1 , \quad .$$

Introducing the *relative redundancies* of the bases Ψ and W with respect to the significance maps

$$\epsilon(\Delta) = \sup_{\lambda\in\Lambda} p_\lambda(\Delta) , \quad \tilde\epsilon(\Lambda) = \sup_{\delta\in\Delta} p_\delta(\Lambda) , \quad (8)$$

one then obtains simple estimates for the logarithmic dimension

Theorem 1: With the above notations, assuming that the significant coefficients $\{\alpha_\lambda, \lambda \in \Lambda\}$ and $\{\beta_\delta, \delta \in \Delta\}$ are independent identically distributed $\mathcal{N}(0,\sigma)$ and $\mathcal{N}(0,\tilde\sigma)$ normal variables respectively, and assuming $r = 0$, the following bounds hold

$$\mathbb{E}\{\mathcal{D}_\Psi(x)\} \ge C + \frac{|\Lambda|}{N}\log_2(\sigma^2) + \log_2\left(\prod_{\lambda'\notin\Lambda}\left(\tilde\sigma^2 p_{\lambda'}(\Delta)\right)^{1/N}\right) \quad (9)$$

$$\mathbb{E}\{\mathcal{D}_\Psi(x)\} \le C + \frac{|\Lambda|}{N}\log_2(\sigma^2 + \epsilon(\Delta)\tilde\sigma^2) + \log_2\left(\prod_{\lambda'\notin\Lambda}\left(\tilde\sigma^2 p_{\lambda'}(\Delta)\right)^{1/N}\right) , \quad (10)$$

with $C = 1 + \gamma/\ln(2)$. Exchanging the roles of Δ and Λ, a similar bound holds for the other logarithmic dimension $\mathcal{D}_W(x)$.

Proof: the proposition follows directly from the fact that in such a situation, equation (6) reduces to

$$\mathbb{E}\{\mathcal{D}_\Psi(x)\} = C + \log_2\Big(\prod_{\lambda\in\Lambda}\left(\sigma^2 + \tilde\sigma^2 p_\lambda(\Delta)\right)^{1/N} \times \prod_{\lambda'\notin\Lambda}\left(\tilde\sigma^2 p_{\lambda'}(\Delta)\right)^{1/N}\Big) , \quad (11)$$

from Lemma 2 and the definition of $\epsilon(\Delta)$. ♠

This result is quite appealing in several respects

i. The bounds in Equations (9) and (10) differ by $|\Lambda|\log_2(1 + \epsilon(\Delta)\tilde\sigma^2/\sigma^2)/N$. Let us assume for a while that this term may be neglected (more on that below). Then the behavior of $\mathbb{E}\{\mathcal{D}_\Psi(x)\}$ is essentially controlled by

$$\log_2\left(\prod_{\lambda'\notin\Lambda}\left(\tilde\sigma^2 p_{\lambda'}(\Delta)\right)^{1/N}\right)$$

The behavior of this term is not easy to understand, but a first idea may be obtained by replacing $p_{\lambda'}(\Delta)$ by its "ensemble average"

$$\frac{1}{N}\sum_{\lambda=1}^{N} p_\lambda(\Delta) = \frac{1}{N}\sum_{\delta\in\Delta}\|w_\delta\|^2 = \frac{|\Delta|}{N} ,$$

which yields the approximate expression:

$$\mathbb{E}\{\mathcal{D}_\Psi(x)\} \approx C + \frac{|\Lambda|}{N}\log_2(\sigma^2) + \left(1 - \frac{|\Lambda|}{N}\right)\log_2\left(\tilde\sigma^2\frac{|\Delta|}{N}\right) . \quad (12)$$

Therefore, if the "Ψ-component" of the signal is sparse enough, i.e. if $|\Lambda|/N$ is sufficiently small (compared with 1), $\mathbb{E}\{\mathcal{D}_\Psi(x)\}$ may be expected to behave as $\log_2\left(\tilde\sigma^2\frac{|\Delta|}{N}\right)$, which suggests to use

$$\hat N_\Psi(x) = 2^{\mathcal{D}_\Psi(x)} \approx 2^C\,\tilde\sigma^2\frac{|\Delta|}{N} \quad (13)$$

as an estimate (up to a multiplicative constant) for the "size" of the W component of the signal. Similarly, it is tempting to use

$$\hat N_W(x) = 2^{\mathcal{D}_W(x)} \approx 2^C\,\sigma^2\frac{|\Lambda|}{N} \quad (14)$$

as an estimate (up to a multiplicative constant) for the "size" of the Ψ component of the signal.

ii. As mentioned above, the difference between the lower and upper bounds depends on two parameters: the sparsity $|\Lambda|/N$ of the Ψ-component, and the relative redundancy parameters $\epsilon(\Delta)$. The latter actually describe the intrinsic differences between the two considered bases. When the bases are significantly different, the relative redundancy may be expected to be small (notice that in any case, it is smaller than 1).

iii. The relative redundancy parameters ϵ and $\tilde\epsilon$ which pop up in our model differs from the one which is generally considered in the literature, namely the *coherence* of the dictionary $W \cup \Psi$ (see e.g. [5], [6], [7])

$$M[W \cup \Psi] = \sup_{\substack{b,b'\in W\cup\Psi \\ b\neq b'}} |\langle b, b'\rangle| .$$

The latter is intrinsic to the dictionary, while the Parseval weights and corresponding ϵ and $\tilde\epsilon$ provide a finer information, as they also account for the signal models, via their dependence in the significance maps Λ and Δ.

iv. Precise estimates for the behavior of the ϵ and $\tilde{\epsilon}$ parameters are fairly difficult to obtain. What would be needed is a model for the significance maps Δ and Λ, in the spirit of the structured models described in the two previous sections. Returning to the wavelet and MDCT case, it is quite natural to expect that models implementing time persistence in Δ and scale persistence in Λ (as in [12], where more numerical simulations are given) would yield smaller values for the relative redundancies than models featuring uniformly distributed significance maps.

Another interesting point is the sensitivity of such tools with respect to departures to the model, or noise. We show that results similar to the above ones still hold true in the presence of white noise, i.e. assuming that the residual r in (1) is a centered Gaussian white noise. In such a situation, denoting by s^2 the variance of the noise r, equation (6) becomes

$$\mathbb{E}\left\{\mathcal{D}_\Psi(x)\right\} = C + \frac{1}{N}\log_2\left[\prod_{\lambda\in\Lambda}\left(\sigma_\lambda^2 + \sum_{\delta\in\Delta}\tilde{\sigma}_\delta^2|\langle\psi_\lambda, w_\delta\rangle|^2 + s^2\right)\right. \\ \left.\times \prod_{\lambda'\notin\Lambda}\left(\sum_{\delta\in\Delta}\tilde{\sigma}_\delta^2|\langle\psi_{\lambda'}, w_\delta\rangle|^2 + s^2\right)\right] , \tag{15}$$

and a similar expression for the logarithmic dimension $\mathcal{D}_W(x)$ with respect to the $W = \{w_\delta\}$ basis. Hence, the approximate expression (12) becomes

$$\mathbb{E}\left\{\mathcal{D}_\Psi(x)\right\} \approx C + \frac{|\Lambda|}{N}\log_2(\sigma^2 + s^2) \\ + \left(1 - \frac{|\Lambda|}{N}\right)\log_2\left(\tilde{\sigma}^2\frac{|\Delta|}{N} + s^2\right) , \tag{16}$$

and the discussion above still holds (after suitable adaptation) as long as the signal's energy $\tilde{\sigma}^2|\Delta|$ exceeds the noise's energy $s^2 N$.

III. Numerical results

These estimates are confirmed by numerical simulations, run on the sparse hybrid models given in (1). We generated several realizations of the signal model (with $r = 0$ first), with fixed number M of MDCT atoms, and variable numbers L of wavelet atoms, and computed the estimated rates

$$\hat{I}_{ton} = \frac{N_\Psi}{N_\Psi + N_W} , \quad \hat{I}_{tr} = \frac{N_W}{N_\Psi + N_W} , \tag{17}$$

to be compared with the ground truth (2), i.e. $I_{ton} = M/(M+L)$ and $I_{tr} = L/(M+L) = 1 - I_{ton}$.

As may be seen from the numerical simulations presented in Figure 1 (which corresponds to averages over 10 realizations of the model), the estimated curves reproduce quite well the theoretical ones. Some discrepancies may be observed at the right hand side of the curves, where the sparsity assumptions are not valid any more, and the correction terms in (12) comes into play. Observe that the curves cross precisely at the correct location $M = L$.

The influence of the noise may be seen on Figure 2: a white noise, whose energy equals 30% of the signal's energy, has been added. The effect is what can be anticipated from (16), namely the presence of an additional noise term moves the experimental curves away from the theoretical ones.

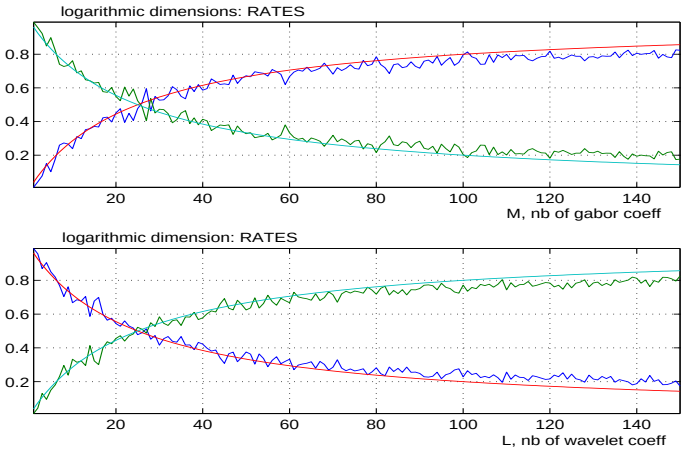

Fig. 1. Transientness and tonality estimates for the model (averaged over 10 realizations). *Top:* $L = 25$, and $M \in \{1, \dots 150\}$; increasing curves: I_{ton} and $\hat{I}_{ton}$; decreasing curves: I_{tr} and $\hat{I}_{tr}$. *Bottom:* $M = 25$, and $L \in \{1, \dots 150\}$; increasing curves: I_{tr} and $\hat{I}_{tr}$; decreasing curves: I_{ton} and $\hat{I}_{ton}$.

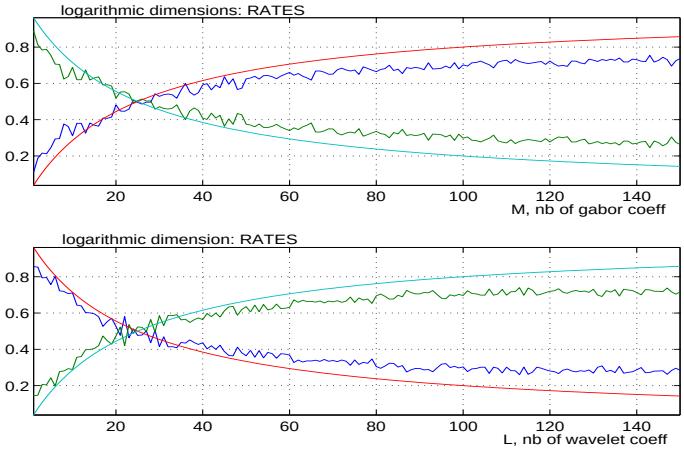

Fig. 2. Influence of white noise: Transientness and tonality estimates for the model (averaged over 10 realizations) with additional white noise. Same legends as before.

Besides the numerical simulations above, the transientness and tonality indices have been tested on real audio signals, yielding very sensible results.[1] A first example, based upon a simple castagnette signal (6 sec long, sampled at 44,100 kHz) is shown in Figure 3. A value for the transientness index and the tonality index was computed for all time frames (23msec. long). Since $I_{ton} = 1 - I_{tr}$, only the transientness index is displayed for the sake of clarity. This signal is quite simple, as it essentially exhibits attacks followed by harmonic tones, and is thus a "perfect" test for the proposed approach. As may be seen from the bottom plot of Figure 3, all attacks are correctly captured, and the corresponding index is quite high.

[1] Additional material, including sound files, may be found at the web site `http://www.cmi.univ-mrs.fr/~torresan/papers/balance`.

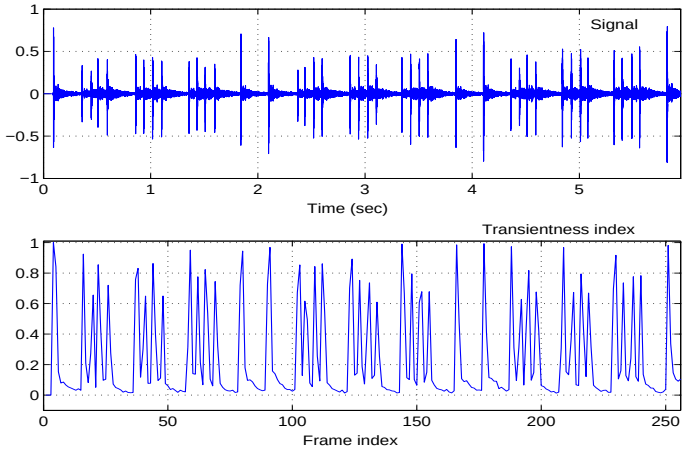

Fig. 3. Transientness index for the test "castagnette" signal. Signal (top) and transientness index (bottom).

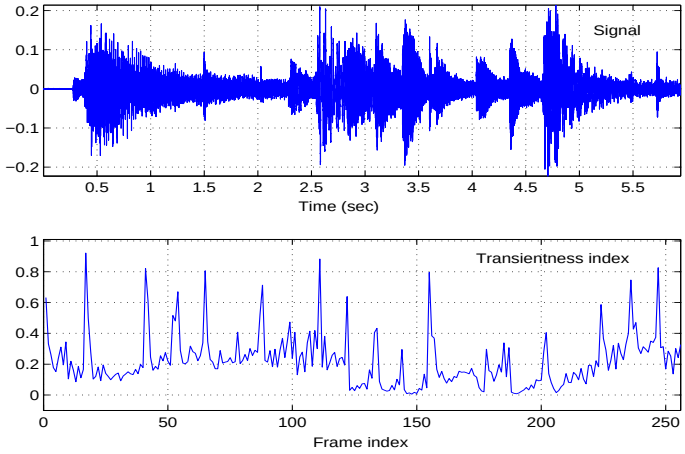

Fig. 4. Transientness index for the test "jazz" signal. Signal (top) and transientness index (bottom).

In between attacks, the transientness index is very low, which is also natural since the signal is essentially harmonic, thus sparsely represented by local cosine basis.

The second sound example displayed here is a more complex audio signal, extracted from a jazz recording (about 6 sec. long, sampled at 44,100 kHz) which features "mixed" tonals and transients. The numerical results are displayed in Figure 4. Notice again that the "obvious" attacks of the signal have been captured by the method. A closer examination of the signal (using a "spectrogram type" representation, not shown here) shows that in the middle part of the signal (more precisely, between seconds 3 and 5), the harmonic content is stronger, which explains the lower average value of I_{tr} there. This illustrates the fact that I_{tr} really provides an estimate of the *proportion* of transients relative to tonals, rather than an absolute indicator of the presence of transient, such as the ones used in transient detection [8] for example.[2]

More numerical results, in the framework of the hybrid audio coding scheme developed in [4], will be given in a forthcoming publication [12].

[2]Let us again refer to the web site `http://www.cmi.univ-mrs.fr/~torresan/papers/balance` for supplementary details.

IV. Conclusions

We have shown that sparsity of wavelet and MDCT signal representations may be exploited in order to balance the amount of tonal and transient components present in the signal. This approach proves to be extremely effective in the context of hybrid audio signal coding [4], [12], and possesses a wider range of applications, including image coding [10].

The theoretical analysis we have outlined here is based on strong a priori assumptions on the signal (essentially, a hybrid model such as (1), with sparse significance maps Λ and Δ, and equal (or comparable) energies for the two layers. When this is not the case, the approach may easily be refined to account for departures from such a situation.

Finally, let us simply mention that the approach may be extended to more than two layers, provided that the considered orthonormal bases are sufficiently different (in terms of their "Parseval weights", see above) to allow the separation. Again, this may prove useful in the context of image coding, where new types of waveforms (e.g. curvelets) may be introduced.

References

[1] 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.

[2] R. Carmona, W.L. Hwang, and B. Torr´esani. *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.

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

[4] L. Daudet and B. Torr´esani. Hybrid representations for audiophonic signal encoding. *Signal Processing*, 82(11):1595–1617, 2002. Special issue on Image and Video Coding Beyond Standards.

[5] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decompositions. *IEEE Trans. Inf. Th.*, 47(7):2845–2862, 2001.

[6] M. Elad and A.M. Bruckstein. A generalized uncertainty principle and sparse representations. *IEEE Trans. Inf. Th.*, 48(9):2558–2567, 2001.

[7] R. Gribonval and M. Nielsen. Sparse representations in union of bases. Technical Report 1499, Institut National de Recherches en Informatique et Automatique, IRISA Rennes, 2003.

[8] S. Levine and J. O. Smith, A switched parametric and transform audio coder, in *Proc. of the International Conference on Acoustics, Speech and Signal Processing, Phoenix*, 1999.

[9] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.

[10] F.G. Meyer, A.Z. Averbush and R.R. Coifman. Multilayered Image Representation: Application to Image Compression. *IEEE Transactions on Image Processing*, 11:1072–1080, 2002.

[11] S. Molla and B. Torr´esani. Hidden markov trees of wavelet coeffi - cients for transient detection in audiophonic signals. In A. Benassi, editor, *Proceedings of the conference Self-Similarity and Applications, Clermont-Ferrand (May 2002)*, 2003. to appear.

[12] S. Molla and B. Torr´esani. An Hybrid Audio Scheme using Hidden Markov Models of Waveforms Preprint (September 2003), submitted.

[13] J. Romberg, M. Wakin and R. Baraniuk, Approximation and Compression of Piecewise Smooth Images Using a Wavelet/Wedgelet Geometric Model. *IEEE International Conference on Image Processing*, sept 2003.

[14] A. Trgo and M. V. Wickerhauser. A relation between Shannon–Weaver entropy and "theoretical dimension" for classes of smooth functions. Preprint, Washington University, Saint Louis, Missouri, 1995.

[15] T. Verma and T. Meng, Sinusoidal Modeling Using Frame-Based Perceptually Weighted Matching Pursuits, in *Int. Conf. Acoustics, Speech and Signal Processing*, Pheonix, USA, 1999.

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

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