MATCHING PURSUIT IN ADAPTIVE DICTIONARIES FOR SCALABLE AUDIO CODING

Emmanuel Ravelli[1,2]*, Gaël Richard*[2]*, and Laurent Daudet*[1]

[1]UPMC - Univ. Paris 06
Institut Jean le Rond d'Alembert-LAM
11 rue de Lourmel, 75015 Paris, France

[2]TELECOM ParisTech
Institut TELECOM
37-39 rue Dareau, 75014 Paris, France

Contact: ravelli@lam.jussieu.fr; daudet@lam.jussieu.fr

ABSTRACT

In previous work [10], we have proposed a new signal representation for audio coding, where the signal is decomposed in a union of MDCT bases using matching pursuit. The resulting coder gave better performance than a transform coder at low bitrates but slightly worse at high bitrates. In this paper, we propose an adaptive matching pursuit algorithm that in the first iterations decomposes the signal into the redundant union of MDCT bases, and then, when the residual energy decay becomes too low, switches to an orthogonal basis (one of the MDCT bases). We investigate simple strategies to determine in which iteration switching is near-optimal in terms of rate-distortion. We present in this paper a prototype audio coder based on this algorithm, that reaches the performance of the previous approach at low bitrates and the one of transform coding at high bit rates.

1. INTRODUCTION

When transparency or near-transparency is required, state-of-the-art audio coders are mostly transform-based and generally use the Modified Discrete Cosine Transform (MDCT). One example of such a coder is MPEG-4 Advanced Audio Coding (AAC) [6] which is able to encode general audio at 64 kbps per channel with a near-transparent quality. However, MDCT-based coders are known to introduce severe artefacts at lower bitrates and they are now outperformed by other coders. These new coders are either purely parametric (e.g. SSC [2]) or hybrid (e.g. HE-AAC [3], AMR-WB+ [7]) and allow better performance than transform coding at 24 kbps per channel or lower. However, parametric and hybrid coders only model a subspace of the input signal and consequently cannot reach transparent quality, even at high bitrates.

In [10], we have proposed a new signal representation that allows better performance than transform coding at low bitrates while allowing transparent quality at high bitrates. This approach is related to transform coding : it could be seen as a generalization of the transform approach since it is based on a simultaneous use of a union of MDCT bases. This allows us to use efficient scalable encoding techniques used in transform coding (e.g. [4]), while producing a sparser decomposition than the transform approach. In comparison, for the same target SNR, there are less significant coefficients to encode than in the transform case, but encoding the parameters of the significant coefficients (the significance map) is more costly. In [10], we have showed that the tradeoff between the number of significant coefficients, and the coding cost of these coefficients, significantly favors our approach at low bitrates. However, at high bitrates, it is necessary to encode a high number of coefficients in both approaches, and the cost of encoding large significance map becomes prohibitive: in this case our approach is outperformed by the standard transform approach.

In this paper, we propose a new decomposition algorithm that, under mild assumptions, provides the "best of both worlds": the same performance as the previous approach [10] at low bitrates and the same performance as transform coding at high bitrates. The signal is first approximated in the overcomplete set of time-frequency atoms used in [10] (union of 8 MDCT bases), and then the residual of this approximation is decomposed using an orthogonal transform (one of the MDCT bases). The signal decomposition is performed on the whole signal using a modified Matching Pursuit (MP) algorithm, with an adaptive dictionary that is changed locally i.e. the union of MDCT is reduced to one MDCT on a frame-by-frame basis. The decomposition is then encoded using similar bitplane encoding methods as used in previous work [10]. The main issue is the design of an efficient strategy to decide on the appropriate MP iteration, in a given frame, to switch from the overcomplete to the complete dictionary.

A similar idea is found in image coding [9], where an overcomplete set of 2D atoms is used to model the edges of an image and the residual of this approximation is coded using a wavelet transform. However, this approach is based on two different dictionaries and two different coding methods. The image coder combines these two different approaches in a rate-distortion way. What we propose is different as we use a unique dictionary and a unique coding method, both are adapted online using the novel switching procedure proposed in this paper.

Another similar approach is found in audio coding [11], where SSC is used to approximate a signal and the residual is coded using a MDCT-based coder. A rate-distortion optimization is used to allocate the available bit budget among the two coders. However, the same remarks apply here. This approach is based on two different paradigms, which is different from our approach as we propose a single paradigm for the signal representation and the coding.

The remainder of this paper is as follows. In section 2, we introduce the signal model and notations. In section 3, we recall the decomposition algorithm used in previous work and describe the new approach. In section 4, we describe how the decomposition is encoded. In section 5, we derive an optimal parameter value for the adaptive decomposition algorithm. In section 6, we present the results, and finally we conclude in section 7.

2. SIGNAL MODEL

We proposed in [10] a signal model based on a union of 8 MDCT bases, where the window length ranges from 128 to 16384 samples (i.e. from 2.9 to 370 ms) in powers of 2. We use the same model here : the smallest windows are needed to model very sharp attacks while larger windows are useful for modeling long stationary components. The signal $f \in \mathbb{R}^N$ is then decomposed as a weighted sum of functions $g_\gamma \in \mathbb{R}^N$ plus a residual of negligible energy r

$$f = \sum_{\gamma \in \Gamma} \alpha_\gamma g_\gamma + r \tag{1}$$

where α_γ are the weighting coefficients. The set of functions $\mathscr{D} = \{g_\gamma, \gamma \in \Gamma\}$ is called the dictionary and is a union of M MDCT bases (called blocks). The functions g, called atoms are defined as:

$$g_{m,p,k}(n) = w_m(u)\sqrt{\frac{2}{L_m}}\cos\left[\frac{\pi}{L_m}\left(u + \frac{1+L_m}{2}\right)\left(k+\frac{1}{2}\right)\right] \tag{2}$$

where

$$u = n - pL_m - T_m. \tag{3}$$

and m is the block index, p is the frame index, k is the frequency index, L_m is the half of the analysis window length of block m (defined as power of two $L_m = L_0 2^m$), P_m is the number of frames of block m, T_m is a time offset introduced to "align" the windows of different lengths ($T_m = \frac{L_m}{2}$, see Fig. 1) and $w_m(u)$ is the sine window defined on $u = 0, .., 2L_m - 1$.

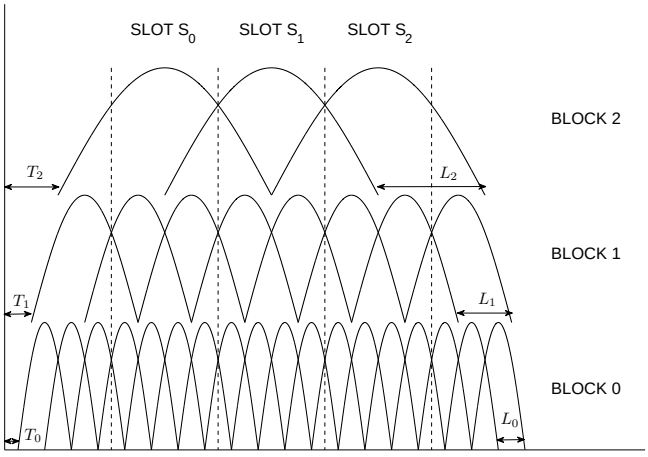

Figure 1: *Analysis windows for $M = 3$ different MDCT window sizes. Dashed vertical lines indicate timeslots.*

3. DECOMPOSITION ALGORITHM

In the case of the orthogonal transform ($M = 1$), $\mathscr{D}$ forms a basis of $\mathbb{R}^N$ and the atoms $\{g_\gamma\}$ are linearly independent. The decomposition of f over $\mathscr{D}$ is then unique, and is simply obtained by projecting the signal on the atoms.

The dictionary is overcomplete when $M > 1$: the dimension of $\mathscr{D}$ is superior to the dimension of the signal, and the decomposition of f in $\mathscr{D}$ is not unique anymore. We are looking for a sparse solution, where the signal is represented by a small number of atoms. Finding an optimally sparse solution is a NP-hard problem if the dictionary is unrestricted [1]. Instead, it is possible to find a sub-optimal solution using e.g. Matching Pursuit (MP) [8]. MP is an iterative algorithm which selects the optimal atom at each iteration (see Algorithm 1 in [10]). We present in the following a modified version of this algorithm that allows us to reduce the size of the dictionary adaptively.

Standard MP is performed globally on the whole signal. To better adapt the model to the local statistics of the signal, we consider in the following temporal segments called "timeslots" whose length is equal to the half of the maximum analysis window length (see Fig. 1). These slots group coefficients in subsets $\mathscr{S}_q$ defined as

$$\mathscr{S}_q = \left\{ \alpha_{m,p,k} \mid \text{floor}\left(\frac{p - P'_m + 1}{P'_m}\right) = q \right\} \tag{4}$$

with $P'_m = 2^{M-m-1}$ is the number of frames of block m in each timeslot. Since the first and last frames of each block are discarded in this scheme, it is necessary to fill the signal f with zeros at both sides before the decomposition to avoid any problem at the edges. In the timeslot $\mathscr{S}_q$, the time support of the largest scale atom includes all smaller scale atoms. Consequently, we define the time support U_q of the timeslot $\mathscr{S}_q$ as the time support of the largest scale atom.

At every MP iteration, one atom is picked up, that can belong to any timeslot. We thus define $n_q(i)$ the MP iterations where the selected atom belongs to the timeslot $\mathscr{S}_q$. The atom selected at iteration $n_q(i)$ decreases the energy of the residual on the time support U_q. We thus define the SNR of the timeslot $\mathscr{S}_q$ as

$$SNR_q(i) = 10\log_{10}\left(\frac{||R^{n_q(i)}_{U_q}||^2}{||R^0_{U_q}||^2}\right) \tag{5}$$

and the SNR decay of the timeslot S_q as

$$DEC_q(i) = 10\log_{10}\left(\frac{||R^{n_q(i)+1}_{U_q}||^2}{||R^{n_q(i)}_{U_q}||^2}\right) \tag{6}$$

with $R^n_{U_q}$ is the part of the residual at iteration n corresponding to the time support U_q. Fig. 2 plots the decay curve $DEC_q(i)$ as a function of $SNR_q(i)$, obtained with the standard MP for a timeslot of an audio signal and two dictionaries: one single MDCT with window length 2048 samples and the union of 8 MDCT bases presented in Sec. 2. In the first iterations, each atom of the overcomplete dictionary decreases significantly the SNR of the timeslot, but after some iterations, the SNR decay becomes small and almost equal to the one of the orthogonal dictionary at same SNR. This is the same phenomenon as the one described in [8]: the atoms extracted in the first iterations are the coherent structures of the signal and after some iterations the residue R^n converges to a process called the dictionary noise. When coding such a decomposition, there is a gain in the first iterations but after some point, it is less costly to encode an atom from an orthogonal dictionary. Consequently, from a coding point of view, it is better to decompose the first iterations in the union of MDCT bases, and then to reduce the dictionary when the decay becomes too low. The proposed modified MP is detailed in Alg. 1.

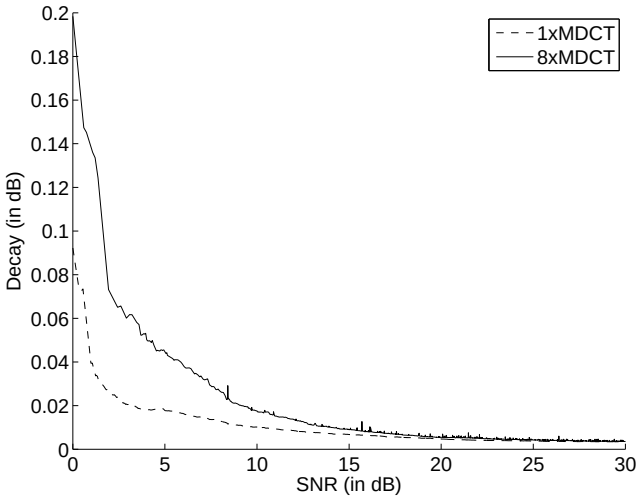

Figure 2: *SNR decay in function of SNR for a timeslot*

Algorithm 1 Adaptive MP

Require: f ; $\mathscr{D} = \{g_\gamma, \gamma \in \Gamma\}$
Ensure: α_γ
  $n = 0$
  $R^0 = f$
  $\alpha_\gamma = 0, \forall \gamma \in \Gamma$
  repeat
    $\gamma_{\text{opt}} = \text{argmax}_{\gamma\in\Gamma} \left| < r, g_\gamma > \right|$
    $c = < r, g_{\gamma_{\text{opt}}} >$
    $R^{n+1} = R^n - c.g_{\gamma_{\text{opt}}}$
    q is the timeslot index of $g_{\gamma_{\text{opt}}}$
    if $DEC_q < SW$ **then**
      remove all atoms in timeslot q except those from the orth. dict.
    end if
    $\alpha_{\gamma_{\text{opt}}} = \alpha_{\gamma_{\text{opt}}} + c$
  until only atoms from the orth. dict. remain in the dict.
  Project the residual R^n on the orth. dict.

4. CODING

As in [10], the coefficients are first interleaved in each timeslot to produce a vector of coefficients per timeslot. Then, we use a slight modified version of the coding algorithm used in [10]: each vector of interleaved coefficients is encoded using an adaptive bitplane encoding algorithm that reduces the size of the significance map after the switch.

We first recall the interleaving process used in [10]. To simplify notations in the following, we introduce a new frame index $p' = \text{mod}\,(p - P'_m + 1, P'_m)$ such that the frame index starts at 0 in each timeslot. The mapping process between the coefficients of a timeslot $\alpha_{m,p',k}$ and the corresponding vector values v_i is then formalized as follows. First we define a recursive function r that performs a permutation of the frames:

$$r(p', M-1) = p' \tag{7}$$

and for $m < M-1$

$$r(p', m) = \begin{cases} r(\frac{p'}{2}, m+1) & \text{if } p' \text{ is even} \\ r(\frac{p'-1}{2}, m+1) + P'_{m+1} & \text{if } p' \text{ is odd} \end{cases} \tag{8}$$

Then, values are mapped according to:

$$v_i = \alpha_{m,r(p',m),k} \tag{9}$$

with

$$i = (kP'_m + p')M + m \tag{10}$$

Then, each vector of interleaved coefficients is encoded using a bitplane encoding algorithm. In [10], the same algorithm as in [4] was used. The basic principle of bitplane encoding is to send successively each bitplane starting from the most significant bitplane. This is done using a scheme in two passes: the significance pass and the refinement pass. The significant pass transmits the subset of the j-th bitplane corresponding to the j-th most significant bits of the non already significant coefficients. The significance pass also transmits the sign of the new significant coefficients. The refinement pass transmits the subset of the j-th bitplane corresponding to the j-th most significant bits of the already significant coefficients. All existing bitplane encoding algorithms differ essentially in the way they perform the significance pass. The approach used in [4] and [10] is based on adaptive Golomb codes: the significance pass does not transmit directly the bits in the current significance map but instead transmits the number of zeros between ones using adaptive Golomb codes.

We use here a slightly modified version of this algorithm. The first bitplanes have the same size as the length of the vector v_i, i.e $M = 8$ times the number of signal samples in a timeslot. After the dictionary switch from overcomplete to orthogonal, all remaining bitplanes contain only significant coefficients from the reduced dictionary. Consequently, the size of the significance map is reduced after the switch. One bit per bitplane is added in order that the decoder knows when the significance map size is reduced ("0" = full overcomplete dictionary, "1" orthogonal transform).

5. OPTIMAL SWITCHING PARAMETER

The proposed adaptive MP depends only on one parameter *SW*, which is the energy decay per coefficient under which it is better to switch from overcomplete to orthogonal. Of course, *SW* can be estimated on-line, by computing at each iteration the most favorable rate-distortion configuration. However, since the coefficients are not encoded separately but in bitplanes, this technique leads to a significant complexity increase (at each iteration, one has to "look ahead" a large number of steps to decide on the best strategy). Instead, we would like to compute a fixed value for *SW* that is approximately optimal (in terms of rate-distortion), without any additional computation in the MP loop.

In the following, we consider one particular slot $\mathscr{S}_q$, and we suppose that the switch is done for that timeslot at the iteration $n_q(i_{switch})$. We make the assumption that the SNR decay DEC_q in that timeslot is constant for the few iterations following the switch (this is approximately verified if the switch does not occur in the first iterations). Then, we consider two cases: if at iteration $n_q(i_{switch})$, we keep in the overcomplete dictionary (no switch), the decay value would be DEC_q^O; if at iteration $n_q(i_{switch})$, the dictionary is reduced to an orthogonal dictionary (switch), the decay value would be lower and equal to DEC_q^T. We now make the empirical observation that there exists a simple relation between DEC_q^O and DEC_q^T. We decompose a 20 seconds signal composed of several audio types contents (monophonic, polyphonic) with

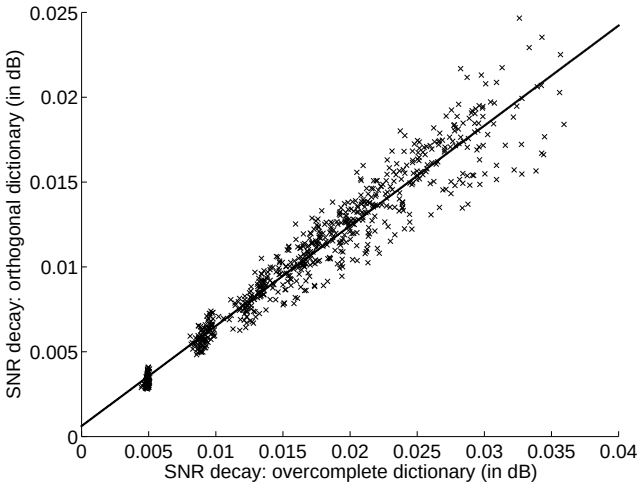

Figure 3: *Decay of the orthogonal dictionary as a function of the overcomplete dictionary; the decay is computed on the* 100 *iterations following the switch.*

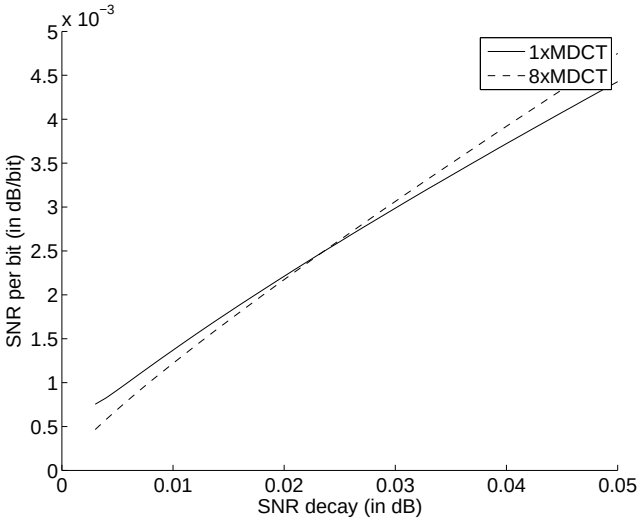

Figure 4: *Distortion per bits as a function of the SNR decay for two dictionaries.*

the adaptive MP and several values of the parameter *SW*. We also decompose the same signal in the overcomplete dictionary with the standard MP (no switch) as a reference to compute DEC_q^O. In each timeslot, and for each parameter value, we compute the mean of the decay on the 100 iterations $n_q(i)$ following the switch at iteration $n_q(i_{switch})$ for each case and we plot DEC_q^T in function of DEC_q^O. Fig. 3 shows the obtained values and a linear approximation. We thus approximate the relation between DEC_q^T and DEC_q^O as: $DEC_t = \gamma DEC_o$ with $\gamma = 0.6$.

Now, as we have supposed that the SNR decay DEC_q in a timeslot is constant for the few iterations following the switch, and if we neglect the influence of the neighboring timeslots $\mathscr{S}_{q-1}$ and $\mathscr{S}_{q+1}$, the energy and the absolute value of the coefficients have approximately the same decay: DEC_q^O if we remain in the overcomplete dictionary, and DEC_q^T if we switch to the orthogonal dictionary. As a bitplane corresponds to a division per two of the absolute value of the coefficients, we are now able to compute the approximate number of coefficients per bitplane:

$$Nb^T = \frac{20}{DEC_q^T \log_2 10} \tag{11}$$

for the orthogonal dictionary and

$$Nb^O = \frac{20}{DEC_q^O \log_2 10} \tag{12}$$

for the overcomplete dictionary. As the major contribution to the bitrate is due to coding the significance maps, in following computations we neglect the sign and refinements bits. Assuming that the coding cost for the significance map can be estimated by entropy, we can then compute the average rate per significant coefficient as

$$R^T = \frac{1}{p}\left(-p\log_2(p) - (1-p)\log_2(1-p)\right) \tag{13}$$

with $p = \frac{Nb^T}{S_{len}}$ for the orthogonal dictionary and

$$R^O = \frac{1}{\tilde{p}}\left(-\tilde{p}\log_2(\tilde{p}) - (1-\tilde{p})\log_2(1-\tilde{p})\right) \tag{14}$$

with $\tilde{p} = \frac{Nb^O}{MS_{len}}$ for the overcomplete dictionary. In short, each coefficient in the bitplane after the switch iteration decreases the SNR by DEC_q^O dB at a cost of R^O bits if we remain the overcomplete dictionary, and decreases the SNR by DEC_q^T dB at a cost of R^T bits if we switch to the orthogonal dictionary. Consequently, we decide to switch to the orthogonal dictionary if $DEC_q^T/R^T > DEC_q^O/R^O$. We are now able to compute the optimal value *SW* numerically. Fig. 4 plots DEC_q^T/R^T and DEC_q^O/R^O in function of DEC_q^O for $\gamma = 0.6$. We finally find numerically the optimal parameter value which is approximately $SW = 0.025$ (numerical simulations have shown that the exact value is not really critical).

6. RESULTS

We have tested our algorithm on the same signals as used in [10], these are 4s-long signals sampled at 44.1 kHz: bagpipe, glockenspiel, harpsichord, horn, orchestra, violin. We compare three coders based on three different signal representations. First, the two coders compared in previous work [10], these are a transform coder based on a single MDCT ($M = 1$) with an analysis window length of 2048 samples, and the overcomplete approach coder with the standard MP in a union of 8 MDCT bases. These two coders are compared with the novel coder proposed in the paper, based on the modified MP with an adaptive dictionary that switches from 8xMDCT to 1xMDCT (with length 2048 samples). We remark that contrary to [10], the evaluation measure is based here on SNR as the decomposition algorithm is based on SNR too. More relevant objective measure for audio coding such as PEMO-Q [5] or even listening tests are planned for future work. The results are shown on Fig. 5.

It clearly shows that the performance of the new coder is the same as the previous approach at low bitrates and the

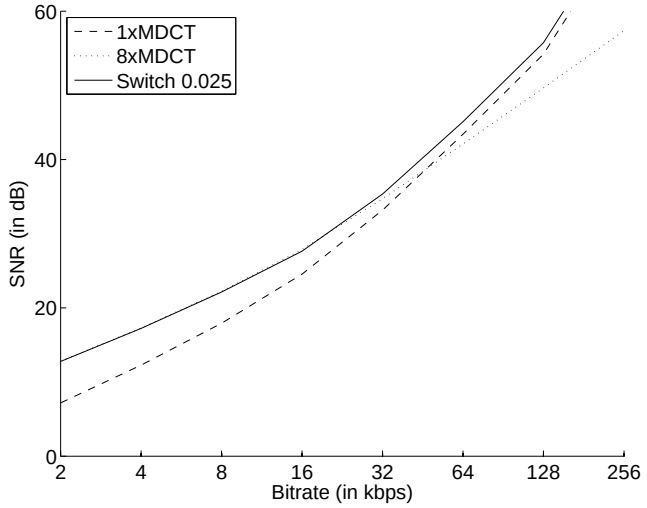

Figure 5: *Mean SNR for six signals in function of the bitrate for 3 coders: transform coder with 1xMDCT, standard MP with 8xMDCT, adaptive MP from 8xMDCT to 1xMDCT with a switch parameter value of SW = 0.025.*

same as transform coding at high bitrates. Preliminary tests also seem to indicate that the approximate switching scheme presented above is very close to the optimum. Fig. 1 compare the computation time needed to code the six files with the three approaches: the single MDCT coder, the proposed adaptive MP coder and the standard MP coder (with a precision of 60 dB). Though it is still much slower than a single MDCT, the new approach is faster than the previous approach.

	MDCT	Adaptive MP	MP (60dB)
Bagpipe	0.03	3.51	129.03
Glockenspiel	0.03	2.87	56.49
Horn	0.03	7.20	21.65
Orchestra	0.03	3.85	133.28
Trumpet	0.03	4.78	61.77
Violin	0.03	4.62	123.35

Table 1: *Normalized computation times on a Core 2 Duo 2.0GHz laptop (in seconds/seconds)*

7. CONCLUSION

The scalable audio coder previously proposed in [10] gave better performance than transform-based coding at low bitrates but slightly worse performance at high bitrates. The signal representation was based on a standard Matching Pursuit decomposition in a redundant union of MDCT bases. We have showed that the energy decay in this overcomplete dictionary is high on the first iterations and becomes almost equal to the decay of an orthogonal dictionary after some iterations. As it is less costly to encode atoms in an orthogonal dictionary, it is better from a coding point of view to decompose the residual in an orthogonal dictionary when the energy decay becomes too low. We thus have proposed a modified MP which switches from an overcomplete dictionary to an orthogonal dictionary when the energy decay is below a threshold. We have then derived an optimal switching parameter value. Finally, we have shown experimentally that the resulting coder reaches the performance of the previous approach at low bitrates and the performance of a transform coder at high bitrates.

This study also raises some questions: first, it is not clear whether there is a fundamental reason for such a simple (approximate) relationship between DEC_q^O and DEC_q^T. Second, this leads to wonder if there are more signal-independent techniques to perform the switch near the optimum. Finally, further studies will have to study whether the rate-distortion optimization as performed here, with distortion as mean quadratic error, is also optimal from a perceptual point of view.

REFERENCES

[1] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *J. Constr. Approx.*, 13:57–98, 1997.

[2] A.C. den Brinker, E.G.P. Schuijers, and A.W.J. Oomen. Parametric coding for high-quality audio. In *Proc. of the 112th AES Convention*, 2002. Paper 5554.

[3] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz. Spectral band replication, a novel approach in audio coding. In *Proc. of the 112th AES Convention*, 2002. Paper 5553.

[4] C. Dunn. Scalable bitplane runlength coding. In *Proc. 120th Audio Eng. Soc. Conv.*, 2006. paper 6749.

[5] R. Hubert and B. Kollmeier. PEMO-Q–A new method for objective audio quality assessment using a model of auditory perception. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 14(6):1902–1911, Nov. 2006.

[6] ISO/IEC, JTC1/SC29/WG11 MPEG. Information technology - coding of audio-visual objects - part 3: Audio, IS14496-3 2001.

[7] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb. AMR-WB+: a new audio coding standard for 3rd generation mobile audio services. In *Proc. Int. Conf. on Acoustics, Speech, and Sig. Proc.*, volume 2, pages 1109–1112, March 2005.

[8] S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12):3397–3415, Dec. 1993.

[9] L. Peotta, L. Granai, and P. Vandergheynst. Image compression using an edge adapted redundant dictionary and wavelets. *Signal Processing*, 86(3):444–456, 2006.

[10] E. Ravelli, G. Richard, and L. Daudet. Extending fine-grain scalable audio coding to very low bitrates using overcomplete dictionaries. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 195–198, 2007.

[11] N.H. van Schijndel and S. van de Par. Rate-distortion optimized hybrid sound coding. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 235–238, 2005.