EXTENDING FINE-GRAIN SCALABLE AUDIO CODING TO VERY LOW BITRATES USING OVERCOMPLETE DICTIONARIES

Emmanuel Ravelli[1,2], *Gaël Richard*[1]

[1]TSI department
GET-ENST (Télécom Paris)
37-39, rue Dareau
75 014 Paris - France
firstname.lastname@enst.fr

Laurent Daudet[2]

[2]Institut Jean Le Rond d'Alembert, LAM team
University Pierre and Marie Curie - Paris 6
11, rue de Lourmel
75 015 Paris - France
lastname@lam.jussieu.fr

ABSTRACT

Signal representations in overcomplete dictionaries are considered here as an alternative to the traditional transform representations for fine-grain scalable audio coding. Such representations produce sparser decompositions and thus allow better coding efficiency than transform coding at very low bitrates. Moreover, the decomposition algorithms are intrinsically progressive, and flexible enough to allow an efficient transient modeling. We propose in this paper a fine-grain scalable audio coder which works on a large range of bitrates (2kbs to 128kbs). Objective measures as well as informal subjective evaluation show that this coder outperforms a comparable transform-based coder at very low bitrates.

1. INTRODUCTION

Depending on the target bitrate and/or quality, standard state-of-the-art audio coding (e.g. MPEG-4 [1]) is based on either transform coding (e.g. MPEG-4 AAC, MPEG-4 Twin-VQ) or parametric coding (e.g. MPEG-4 HILN, MPEG-4 SSC). Transform coding offers good to transparent quality for average-to-high bitrates, but its quality degrades quickly at lower bitrates; on the other hand parametric coders perform well at very low bitrates but their quality reaches a ceiling at increasing bitrates. The goal of this paper is to explore new signal representation strategies that perform better than the standard transform-based coders at very low bitrates, but that can reach arbitrary precision. Some recent hybrid algorithms combine parametric and transform coding [2, 3], but they are based on two different approaches. Using a single (and conceptually simple) paradigm on the whole range of bitrates is not only appealing from a theoretical point of view, it is also useful for seamless progressive transmission of information, a concept referred to as scalable compression.

The current algorithms for scalable compression are mostly inspired from the image coding world, for instance SPIHT [4]. These image coders are commonly used in web browsing: as the number of received bits increases, the images are progressively displayed with more and more details. In audio, a similar task is to stream the same audio content to several users with different and/or varying bandwidth possibilities. With standard fixed bitrate audio coding, the server has to store a number of bitstreams corresponding to different compression ratios, and for a given user select the one corresponding to its bandwidth with a conservative approach. In the case of fine-grain scalable coding, a unique bitstream at maximum resolution is stored; users with less bandwidth need just truncate every time frame, and still receive the file with the best quality given their bandwidth. Most existing fine grain scalable audio coders (e.g. BSAC [5], PSPIHT [6], SCALA [7]) are transform-based : the signal is decomposed in an orthogonal basis of time-frequency functions (e.g. MDCT) and the resulting coefficients are quantized and coded. These coders are known to give good results at high bitrates but introduce severe artifacts at low bitrates.

In this paper, we show that by using overcomplete waveform dictionaries, i.e. representation spaces with a dimension much higher than the signal space, it is possible to extend the range of possible bitrates of fine-grain scalable audio coding in the lower end. Moreover, the decomposition algorithms associated with these overcomplete sets are intrinsically progressive, meaning that the most salient information is extracted first : this scheme is therefore naturally scalable. The advantages of an overcomplete approach over the transform approach are the following: a much sparser decomposition i.e. the energy is concentrated on fewer coefficients and thus there are less coefficients to code; a multi-resolution decomposition which allows the modeling of simultaneous components which can have different time-frequency optimal tradeoffs; a flexible decomposition algorithm which allows modifications such as an efficient pre-echo control. However, one has to find efficient encoding strategies so that the gain of concentrating the energy in less coefficients is not offset by the cost of coding the extra set of atom's parameters. Also, a drawback of using overcomplete representations is a significant increase in the computational cost.

The remainder of this paper is as follows. In section 2, we describe the signal decomposition algorithm; in section 3, we describe the bitplane coding which is used to quantize and code the coefficients of the decomposition; in section 4, we present the results; and finally section 5 is devoted to conclusions.

2. SIGNAL DECOMPOSITION

A signal $f \in \Re^N$ is decomposed as a weighted sum of functions $g_\gamma \in \Re^N$ which form the set of functions $\mathcal{D} = \{g_\gamma, \gamma \in \Gamma\}$, called the dictionary.

$$f = \sum_{\gamma \in \Gamma} \alpha_\gamma g_\gamma \quad (1)$$

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

2.1. The transform approach: the time-varying MDCT

In the vast majority of state-of-the-art transform coders, the Modified Discrete Cosine transform (MDCT) is used [8]. In this case, $\mathcal{D}$ has the same dimension as the signal and the functions g_γ form an orthogonal basis of $\Re^N$, the decomposition is then unique. The functions $g_\gamma \in \Re^N$ corresponding to the MDCT transform with a window w of size $2L$ are of the form:

$$g_{k,p}(n) = w(n-pL)\,\cos\left[\frac{\pi}{L}\left(n-pL+\frac{L+1}{2}\right)\left(k+\frac{1}{2}\right)\right] \quad (2)$$

Usually, a block switching approach is used where two window sizes (e.g. 2048 and 256) and four window shapes (*long*, *short*, *startlong*, *stoplong*) are adaptively chosen depending on an energy or perceptual entropy criteria. Using only two windows could be considered as too rigid and other approaches (e.g. [9]) investigate the use of more than two window sizes using a time segmentation algorithm. However, these methods still remain constrained to a fixed resolution in a given time segment. This is not optimal for sound signals containing simultaneous components localized both in time and frequency. For instance, drums on top of long sustained notes would force a time segmentation algorithm to break up the long notes into smaller pieces.

2.2. The overcomplete approach: Matching Pursuit with a union of MDCT bases

We propose an approach where $\mathcal{D}$ is a union of MDCT bases with different window sizes. In this case, the dimension of $\mathcal{D}$ is greater than the dimension of the signal, and the decomposition is not unique anymore. Hence, it is possible to choose an optimal or nearly-optimal decomposition with respect to some pre-defined criteria. Several algorithms with different complexities have been proposed in the literature to find such decompositions [10, 11]; we use in our case the Matching Pursuit algorithm [10], which is a fast sub-optimal iterative algorithm. At each iteration, Matching Pursuit chooses the function in $\mathcal{D}$ most correlated with the signal, subtracts it, and iterates until some stopping condition is met (see Algorithm 1).

Algorithm 1 Standard Matching Pursuit

input: f ; $\mathcal{D} = \{g_\gamma, \gamma \in \Gamma\}$
output: α_γ
  $R^0 f = f$
  $n = 0$
  $\alpha_\gamma = 0, \forall \gamma \in \Gamma$
  repeat
    $\gamma_{\text{opt}} = \text{argmax}_{\gamma\in\Gamma} |< R^n f, g_\gamma >|$
    $c =< R^n f, g_{\gamma_{\text{opt}}} >$
    $R^{n+1} f = R^n f - c.g_{\gamma_{\text{opt}}}$
    $\alpha_{\gamma_{\text{opt}}} = \alpha_{\gamma_{\text{opt}}} + c$
    $n = n + 1$
  until a condition is met (on SNR or number of iterations)

Matching Pursuit with a union of MDCT bases invariably introduces pre-echo when decomposing signals containing transients. This problem is illustrated in Fig. 1. An extract of a glockenspiel signal is decomposed with Matching Pursuit and a union of 4 MDCT bases (window sizes 2048, 1024, 512, 256 samples). The second subplot shows the residual at iteration 10. The function

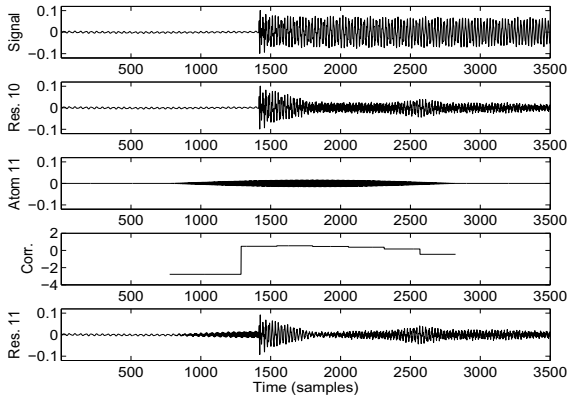

Figure 1: *Illustration of the pre-echo artifact*

which is best correlated with this residual is in the third subplot. The correlation of the un-windowed function with the original signal on subframes of size 256 (fourth subplot shows the log of the absolute value) shows that the beginning of the function is not correlated with the signal. This has as a consequence the creation of unwanted energy just before the transient, which appears in the residual at iteration 11 in the fifth subplot.

Algorithm 2 Matching Pursuit with pre-echo control

input: f ; $\mathcal{D} = \{g_\gamma, \gamma \in \Gamma\}$; $thresh$
output: α_γ
  $R^0 f = f$
  $n = 0$
  $\alpha_\gamma = 0, \forall \gamma \in \Gamma$
  M_{min} is the shortest window size
  repeat
    loop
      $\gamma_{\text{opt}} = \text{argmax}_{\gamma\in\Gamma} |< R^n, g_\gamma >|$
      $M = \max\left(g_{\gamma_{\text{opt}}}\text{window size}/8, M_{min}\right)$
      $S^i f = i-$th M-length subframe of f
      $S^i g_{\gamma_{\text{opt}}} = i-$th M-length subframe of un-windowed $g_{\gamma_{\text{opt}}}$
      $d_i = \left|< S^i f, S^i g_{\gamma_{\text{opt}}} >\right|$
      if $max_i(d_i) >= thresh * min_i(d_i)$ **then**
        $\Gamma = \Gamma \setminus \{\gamma_{\text{opt}}\}$
      else
        exit loop
      end if
    end loop
    $c =< R^n, g_{\gamma_{\text{opt}}} >$
    $R^{n+1} x = R^n x - c.g_{\gamma_{\text{opt}}}$
    $\alpha_{\gamma_{\text{opt}}} = \alpha_{\gamma_{\text{opt}}} + c$
    $n = n + 1$
  until a condition is met

Gribonval pointed out this problem in [12] with Matching Pursuit and a Gabor dictionary. He proposed a modified Matching Pursuit algorithm which removes the pre-echo artifact. However, this algorithm was designed for a complex Gabor dictionary and is not adapted for a union of real MDCT bases. Moreover, this modified Matching Pursuit significantly increases the computa-

tional cost. Alternatively, we propose a simple modification of the Matching Pursuit algorithm which reduces pre-echo artifacts. At each iteration, the function in $\mathcal{D}$ most correlated with the signal is chosen. Then the correlation of the un-windowed function with the original signal is computed on subframes of size M (as in Fig. 1). If the max of the correlations divided by the min of the correlations is superior to a threshold then the function is not selected and removed from the dictionary, otherwise the function is kept and subtracted from the residual (see Algorithm 2).

3. CODING

3.1. Coefficients grouping and interleaving

As the signal f is decomposed as a whole and the coding stage is performed frame-by-frame, the coefficients α_γ must be grouped and interleaved. First, the coefficients are grouped in frames (i.e. time segments) of length $M_{max}/2$ where M_{max} is the maximum window size. Then, in each frame, the coefficients are grouped and interleaved such that the coefficients which are close in frequency and scale are put together. For each frame, we obtain a vector of coefficients whose length is equal to the frame length multiplied by the dimension of $\mathcal{D}$. Finally, these vectors of coefficients are encoded independently.

3.2. Bitplane encoding

To achieve fine-grain scalability, the coefficients have to be quantized and coded in an embedded manner. Bitplane encoding is an efficient technique which produces an embedded bitstream. The coefficients are represented in sign-magnitude form as in Fig. 2; each column corresponds to the sign and the binary representation of the magnitude of a coefficient, and each row corresponds to a bitplane. The most significant bits of all magnitude coefficients are transmitted first, then the next most significant bits, down to the least significant bits. In this approach, the most significant coefficients are transmitted first and then successively refined. It is the opposite of the non-embedded approach where all bits of the first coefficient are transmitted first, then all bits of the next coefficients, until the last coefficient.

	c_0	c_1	c_2	c_3	c_4	c_5
	-1	5	14	-3	-12	-1
Sign	0	1	1	0	0	0
B.P. 3	0	0	1	0	1	0
B.P. 2	0	1	1	0	1	0
B.P. 1	0	0	1	1	0	0
B.P. 0	1	1	0	1	0	1

Figure 2: *Bitplanes*

The encoding of a bitplane is performed in two steps. **1) Significance map and sign encoding:** the index and sign of the newly significant coefficients (i.e. whose most significant bit belong to the bitplane) are encoded. **2) Refinement pass:** the already significant coefficients are refined, the bits in the bitplane corresponding to these coefficients are encoded. While the refinement pass is straightforward and is generally the same in all algorithms, the significance map encoding can be achieved in several different manners. In audio coding, several approaches have been experimented : arithmetic coding based [5], tree based [6], runlength encoding based [7]. We use the approach of [7], based on an adaptive runlength encoding algorithm which appears to perform well on sparse significance maps.

4. EVALUATION

Our coder is evaluated in comparison with a reference transform-based coder. The goal of this paper is to compare different signal decomposition strategies; consequently, the quantization/coding part is kept simple and no psychoacoustic model is used, this will be investigated in future works.

In our coder, $\mathcal{D}$ is a union of 8 MDCT bases whose windows are cosine-based with the following sizes: 16384, 8192, 4096, 2048, 1024, 512, 256, 128; the parameter of the pre-echo control $thresh$ is set to 100. The reference coder is very close to the one in [7], it is based on a time-varying MDCT with two window sizes (2048 and 256), a block-switching module based on an energy criteria and the same bitplane encoding algorithm.

We use the PEMO-Q [13] software as an objective measure to evaluate and compare the coders. First, the transform-based coder and our coder are compared on signals containing no strong transients. Then the algorithm with pre-echo control presented in 2.2 is evaluated on a signal containing transients.

Signals containing no strong transients. For such signals, the reference coder uses only long windows; and Matching Pursuit with or without pre-echo control gives similar results. We compare the transform-based reference coder and our coder on 4s extract of 5 signals from the SQAM database : bagpipe, horn, orchestra, trumpet, violin, all sampled at 44.1 kHz. The mean of the Perceptual Similarity Measure (PSM) and the mean of the Objective Difference Grade (ODG) obtained with PEMO-Q are given in Fig. 3. At low bitrates, our coder produces less non-zero coefficients than the transform-based coder and thus it is cheaper to obtain the same quality; while at high bitrates, a lot of non-zeros coefficients have to be encoded and it is more costly to code the indexes in our coder. Consequently the reference coder gives a slightly better PSM / ODG at high bitrates, but is outperformed significantly by our coder at low bitrates.

Glockenspiel signal and pre-echo control. A 4s extract of the glockenspiel signal from the SQAM database is coded using 3 coders: the reference transform-based coder, the matching-pursuit based coder without pre-echo control, the matching-pursuit based coder with pre-echo control. The reference coder uses short windows for the transients, i.e. the onsets of the glockenspiel signal. This reduces pre-echo but also introduces other artifacts: at low bitrates, the bit budget is spent to model the attack but the stationary part of the last note is lost as it needs a lot of bits to model a sin-like component with short windows. This is not the case for our coder, since two types of functions are superimposed in time: long windows to model the stationary part of the last note and short windows to model the attack of the current note. The Perceptual Similarity measure and the Objective Difference Grade obtained with PEMO-Q are given in Fig. 4. This results shows that our coder outperforms the reference coder on the whole range of considered bitrates; moreover, the results are improved with the pre-echo control modification of the Matching Pursuit algorithm.

Informal listening tests confirm the results obtained with the objective measure. Some audio files are available at the following address: http://www.lam.jussieu.fr/src/Membres/Ravelli/waspaa07/.

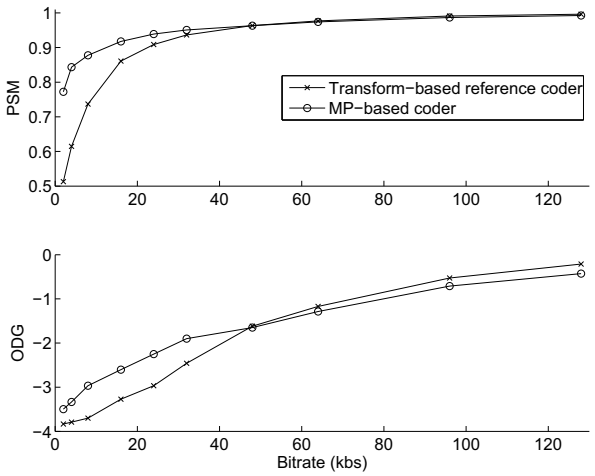

Figure 3: *Mean of Perceptual Similarity Measure and mean of Objective Difference Grade for 4 seconds extract of 5 SQAM signals*

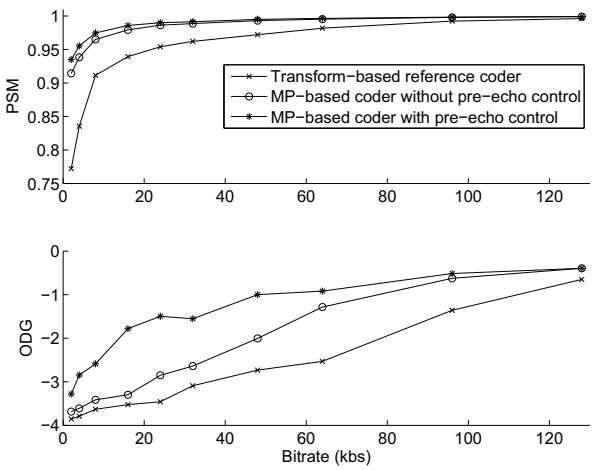

Figure 4: *Perceptual Similarity Measure and Objective Difference Grade for a 4 seconds extract of the glockenspiel signal*

5. CONCLUSION

This paper deals with signal representations in overcomplete sets and their application to audio coding. We have showed that these kinds of signal representations outperform the standard transform approach for fine-grain scalable audio coding at very low bitrates. It was shown that artifacts often encountered at low bitrates by transform-based coders such as birdies and transient deterioration are significantly reduced when using the overcomplete approach. The main limitation of our technique is its computational cost, as most efficient Matching Pursuit implementations require typically one hour for a 30 s musical piece on a current desktop computer, decomposing up to transparency. Note that the bitplane coding part is in comparison extremely fast (much faster than real time), as are decoding and the inverse transform.

Since the goal of this paper is to experiment with different transform strategies, we have adhered to simple (although efficient) quantization and coding schemes. As such, the coder presented in this paper is not yet competitive to state-of-the art parametric / transform coders as it does not incorporate any psychoacoustic masking model. This will be the goal of further research. However, it shows the potential of signal representations in overcomplete sets and its potential for scalable audio coding.

6. REFERENCES

[1] I. O. for Standardization, "ISO/IEC 14496-3, information technology - coding of audio-visual objects - part 3: Audio."

[2] M. Wolters, K. Kjrling, D. Homm, and H. Purnhagen, "A closer look into MPEG-4 high efficiency AAC," in *Proc. 115th AES Convention*, Oct. 2003.

[3] N. van Schijndel and S. van de Par, "Rate-distortion optimized hybrid sound coding," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 235–238.

[4] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Systems Video Tech.*, vol. 6, no. 3, pp. 243–250, Jun. 1996.

[5] S.-H. Park, Y.-B. Kim, and Y.-S. Seo, "Multi-layer bit-sliced bit rate scalable audio coding," in *Proc. of the 103rd Convention of the Audio Engineering Society*, 1997, preprint 4520.

[6] M. Raad, A. Mertins, and I. Burnett, "Scalable to lossless audio compression based on perceptual set partitioning in hierarchical trees (PSPIHT)," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc.*, vol. 5, May 2003, pp. 624–627.

[7] C. Dunn, "Scalable bitplane runlength coding," in *Proc. 120th Convention of the Audio Engineering Society*, 2006, preprint 6749.

[8] S. Shlien, "The modulated lapped transform, its time-varying forms, and its applications to audio coding standards," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 4, pp. 359–366, July 1997.

[9] O. A. Niamut and R. Heusdens, "RD optimal time segmentations for the time-varying MDCT," in *Proc. 12th European Signal Processing Conference*, 2004, pp. 1649–1652.

[10] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[11] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[12] R. Gribonval, E. Bacry, S. Mallat, P. Depalle, and X. Rodet, "Analysis of sound signals with high resolution matching pursuit," in *Proc. of the International Symposium on Time-Frequency and Time-Scale Analysis*, 1996, pp. 125–128.

[13] R. Huber and B. Kollmeier, "PEMO-Q & new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.