# REPRESENTATIONS OF AUDIO SIGNALS IN OVERCOMPLETE DICTIONARIES: WHAT IS THE LINK BETWEEN REDUNDANCY FACTOR AND CODING PROPERTIES?

*Emmanuel Ravelli, Laurent Daudet*

Laboratoire d'Acoustique Musicale
Université Pierre et Marie Curie (Paris 6)
11 rue de Lourmel, 75015 Paris, France
{ravelli|daudet}@lam.jussieu.fr

## ABSTRACT

This paper addresses the link between the size of the dictionary in overcomplete decompositions of signals and the rate-distortion properties when such decompositions are used for audio coding. We have performed several experiments with sets of nested dictionaries showing that very redundant shift-invariant and multi-scale dictionaries have a clear benefit at low bit-rates ; however for very low distortion a lot of atoms have to be encoded, in these cases orthogonal transforms such as the MDCT give better results.

## 1. INTRODUCTION

In the past few years, research has been very active in the area of sparse signal representation in overcomplete dictionaries [1, 2, 3]. These techniques have been found to give successful results for a lot of signal processing problems, including source separation, denoising, detection, classification and compression.

Coding is one of the great application of interest, in particular very low bit-rate image coding (see for ex. the work from EPFL signal processing institute [4]). However, comparable work for the compression of audio signals is quite limited. Matching-pursuit [1] based techniques are used for sinusoidal modeling of audio signals and parametric audio coding in e.g. [5, 6], but the analysis-synthesis is performed on a frame-by-frame basis and thus does not exploit the long-term structure of audio signals. Other works, including [7] consider an audio signal as a whole and represent it using an overcomplete dictionary composed by cosine and wavelet functions, exploiting the structure of the coefficients; however, the described hybrid audio coder is rudimentary, and relies only on a weakly redundant dictionary.

In this paper, we also perform a global analysis of the signal, using a matching-pursuit based algorithm, and study the performance of a coder based on such sparse representation techniques. Particularly, we investigate the link between the degree of redundancy of the dictionary and the Rate-Distortion (R-D) performance of the coder. This is done by using nested dictionaries, with different nesting strategies, and studying the obtained rate-distortion curves on various artificial and natural signals.

The remaining of the paper is as follows: in section 2, we present sparse approximation techniques and how they relate to coding; in section 3, we describe our audio coder; and finally in section 4, we present experiments and results.

## 2. AUDIO CODING AND SPARSE REPRESENTATIONS

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

$$x = \sum_{\gamma \in \Gamma} \alpha_\gamma g_\gamma \tag{1}$$

The ensemble $\mathcal{D}$ is called a dictionary and the functions $g_\gamma$ are called atoms. The representation is exact-sparse (resp. approximate-sparse) if a large number of the coefficients $\alpha_\gamma$ are zeros (resp. approximately zeros) i.e. the energy of the signal is concentrated on a small number of coefficients.

In the case when $\mathcal{D}$ has the same dimension as the signal and the functions $g_\gamma$ form an orthogonal base of $\Re^N$, the decomposition is unique and equivalent to an orthogonal transform. In state-of-the-art audio coders (e.g. AAC, [8]), an orthogonal transform based on local cosine functions is used, the Modified Discrete Cosine Transform (MDCT). The atoms corresponding to the MDCT transform of a signal of length $N = PL$ and a frame size of $2L$, are defined as:

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

with $n = 0,..,N-1$, $k = 0,..,L-1$ and $p = 0,..,P-1$. $w$ is a window which is complementary in energy i.e. verifies:

$$w^2(n) + w^2(n+L) = 1, \quad n = 0,..,L-1 \tag{3}$$

However, when the dimension of $\mathcal{D}$ is superior to the dimension of the signal, i.e. when the dictionary is overcomplete, the decomposition is not unique anymore. Hence, one can choose amongst all these decompositions one which is optimal or nearly-optimal with respect to some pre-defined criteria. Several algorithms with different complexities have been proposed in the litterature to find such decompositions (see e.g. [1, 2, 3]). We use in our case the matching pursuit algorithm [1], which is a fast sub-optimal iterative algorithm. At each iteration, Matching Pursuit chooses the atom in the dictionary most correlated with the signal, subtracts it, and iterates until some stopping condition is met.

A drawback of the MDCT for the representation of audio is that it does not carry explicitly phase information. Alternatively, one can use the Modulated Complex Lapped Transform (MCLT [9]), which is a complex extension of the MDCT that has the propriety of phase-invariance [10], at the cost of a $2\times$ increase in the number of (real) coefficients (i.e. a $2\times$ overcompleteness). In this case, both atoms and coefficients are complex ; consequently the standard matching pursuit cannot be used; instead, we use a two-dimensional matching pursuit using the projection of the signal in

the subspace composed by the complex atoms and their conjugates as described in [11].

There are two ways to further increase the redundancy of our dictionary. First, we use a generalized MDCT/MCLT transform where the resolution in frequency is increased. Second, we also use overcomplete dictionaries composed by an union of several transforms with different frame sizes. In a given experiment, in order to ensure a meaningful comparison on the influence of the redundancy factor on the R-D performance, we ensure that the dictionaries are always nested, i.e. lowest-redundant dictionaries are always included in highest-redundant ones.

## 3. OVERVIEW OF OUR AUDIO CODER

### 3.1. Analysis

We have used a fast implementation of the matching pursuit algorithm : the Matching Pursuit ToolKit (MPTK [12]). Thirteen dictionaries are tested: standard MDCT with a frame size of 2048 samples, standard MCLT with a frame size of 2048, four dictionaries composed by a generalized MCLT with a frame size of 2048 and a FFT size of respectively 4096, 6144 and 8192 frequency bins; seven dictionaries composed by a union of respectively 2,..,8 standard MDCT with frame sizes 128, 256, 512,1024, 2048, 4096, 8192,16384 samples.

### 3.2. Quantization and entropy coding

For real coefficients (MDCT), the DPCM-based quantization scheme as described in [13] is used. For complex coefficients (MCLT), an extended version of this scheme which quantizes the phase with Unrestricted Polar Quantization (UPQ, [14]) is used.

Adaptive arithmetic coding [15] is used to encode the output of the quantizers and the indexes of the coefficients in the dictionary.

## 4. EXPERIMENTAL RESULTS

We compare the Rate-Distortion (R-D) curves for different sets of dictionaries (for each figure, Rate is in Kbps and Distortion in dB). First we compare the standard MDCT with the standard MCLT; then, we study the influence of the frequency resolution for a generalized MCLT; and finally, several dictionaries using a concatenation of MDCT with different scales are tested. The signals used for the experiments are: a white noise; a synthetic signal of bell composed by a sum of damped sinusoids; and a real signal of a pop music recording (from MPEG SQAM test database).

### 4.1. MDCT vs. MCLT

The first idea is to use the MCLT which can be seen as a $2\times$ overcomplete dictionary. This redundant dictionary needs fewer atoms than the orthogonal MDCT to reach the same target SNR; and thus it needs fewer bits to code the indexes and the norm of the coefficients. However, with the MCLT, there is an additional number of bits needed to code the phase information where the MDCT needs only one bit for the sign of each coefficient. Consequently, the additional number of bits needed by the phase counterbalances the bits gained by the smaller number of atoms; and thus, depending on the number of atoms selected by the coder (and thus the target bit rate), the MCLT gives better or worse distortion-rate performance than the MDCT.

Figures 1 to 3 compare the R-D curves for the MDCT and the MCLT obtained with the three test signals. As expected, the

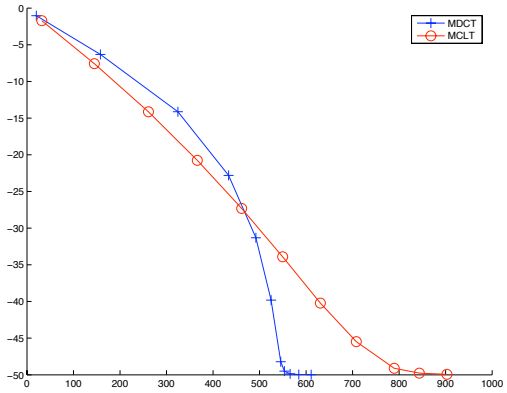


Figure 1: *R-D curve for white noise: MDCT vs MCLT (x: rate in Kbps; y: distortion in dB).*

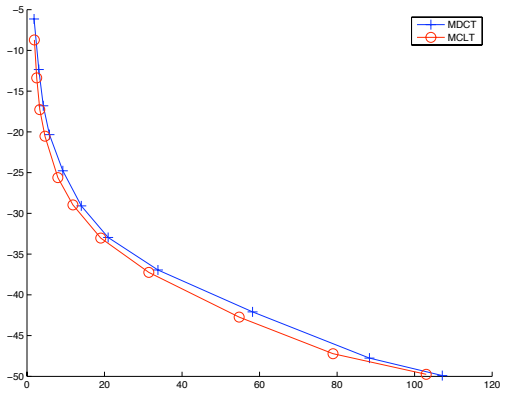


Figure 2: *R-D curve for synthetic bell signal: MDCT vs. MCLT (x: rate in Kbps; y: distortion in dB).*

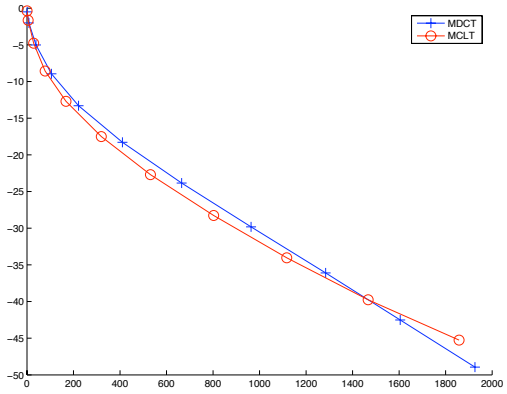


Figure 3: *R-D curve for pop music: MDCT vs. MCLT (x: rate in Kbps; y: distortion in dB).*

MDCT performs better at high rates whereas the MCLT gives better performance at low rates (this crossover would appear on fig. 2 at a highest rate than displayed; this is due to the fact that this signal is very tonal by nature).

### 4.2. Influence of frequency oversampling for the MCLT

The second idea is to increase the redundancy of the MCLT dictionary by increasing the size of the FFT i.e. increasing the precision for the frequency localization of the atoms. In the following figures, the standard MCLT uses the same size for the FFT and the frame, 2048 samples. For the generalized MCLT, the FFT is zero-padded using respectively 4096, 6144 and 8192 samples resulting a degree of redundancy of respectively 4, 6, 8.

By doing that, we hope to better estimate the sinusoidal components of the signal. As an audio signal can be modeled simply by a sum of sinusoids, such dictionaries should give better rate-distortion performance. However, the following figures show that the gain due to the better representation is lost due to the increased number of bits needed to code the indexes. And thus, for any rates, a simple MCLT performs better than the generalized MCLT.

### 4.3. Influence of the scale for the MDCT

The last idea is to use an overcomplete dictionary composed by a concatenation of several MDCT with different scales. The size of the frames are 128, 256, 512, 1024, 2048, 4096, 8192, 16384 samples. The nested dictionaries are obtained by adding one scale alternatively higher and lower, i.e.
$\mathcal{D}_1 = \{mdct(2048)\}$,
$\mathcal{D}_2 = \{mdct(1024)\} \bigcup \{mdct(2048)\}$,
$\mathcal{D}_3 = \{mdct(1024)\} \bigcup \{mdct(2048)\} \bigcup \{mdct(4096)\}$, etc.

Figs. 7 to 9 show the R-D curves obtained with the three test signals. These show that at high rate MDCT gives better rate-distortion performance whereas at low rate the most redundant dictionary performs better. However, it also shows there is no compromise between an orthogonal transform and a highly redundant dictionary at mid rate.

## 5. CONCLUSIONS

This study attempts at clarifying the role of overcompleteness in sparse representations of audio signals, in the framework of audio coding. We have shown that, at high bitrates, the orthogonal MDCT is always better in terms of rate-distortion. However, at low bitrates, introducing redundancy improves the R-D performance. Indeed, one can see parametric coders (or more generally sinusoidal models) as very redundant systems with strong signal priors. Furthermore, our findings suggest that increasing the number of scales in the dictionary is a more efficient strategy than increasing the precision of the frequency (e.g. with zero-padding).

In order to broaden these conclusions, these preliminary tests have now to be conducted on a much wider variety of sounds: although general trends are usually similar, details of the R-D behavior is very signal-dependent. Furthermore, it would be desirable to use perceptual measures of distortion instead of the simple quadratic error. Indeed, at crossover points when the MDCT wins over redundant transforms the SNR usually reaches around 30-40 dB, which is the same order of magnitude as the SNR of MDCT-based coders operating at "transparent" qualities. Through such studies, our ultimate goal is to offer a single paradigm for audio coding, encompassing transform and parametric coding.

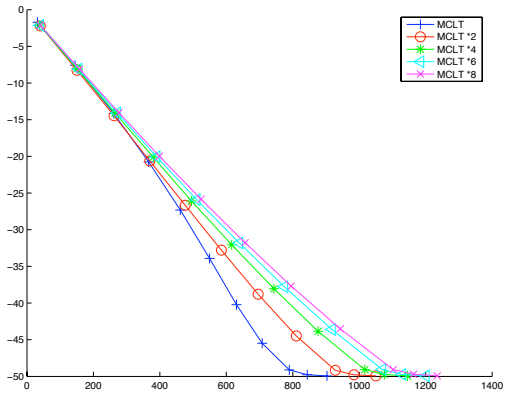


Figure 4: *R-D curve for white noise: influence of the frequency resolution for the MCLT (x: rate in Kbps; y: distortion in dB).*

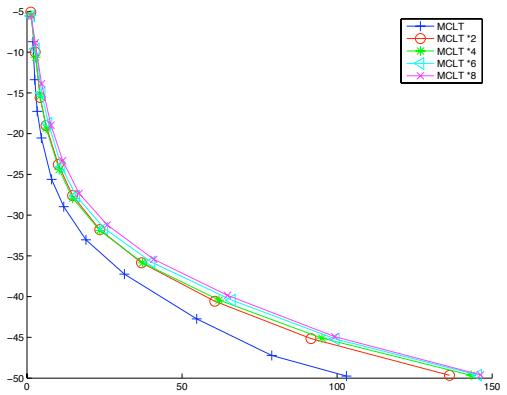


Figure 5: *R-D curve for synthetic bell signal: influence of the frequency resolution for the MCLT (x: rate in Kbps; y: distortion in dB).*

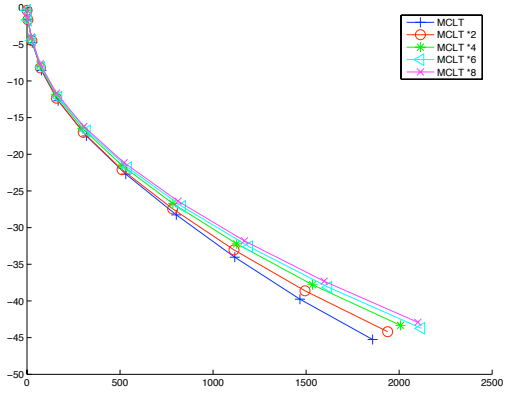


Figure 6: *R-D curve for pop music: influence of the frequency resolution for the MCLT (x: rate in Kbps; y: distortion in dB).*

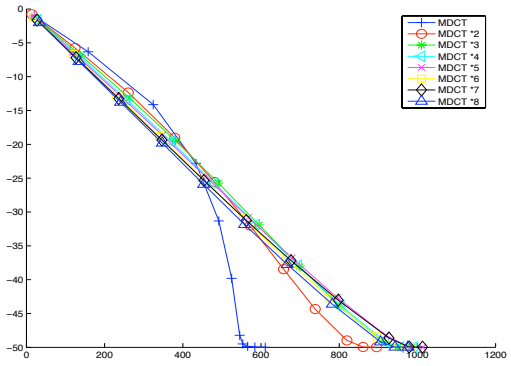


Figure 7: *R-D curve for white noise: influence of the scale for the MDCT (x: rate in Kbps; y: distortion in dB).*

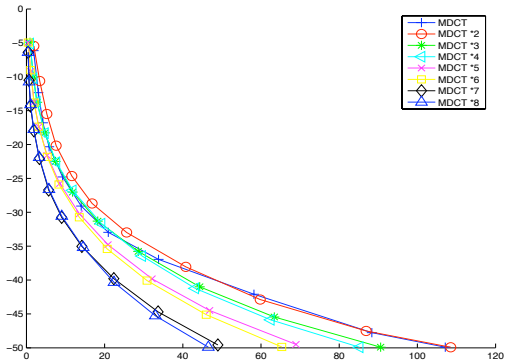


Figure 8: *R-D curve for synthetic bell signal: influence of the scale for the MDCT (x: rate in Kbps; y: distortion in dB).*

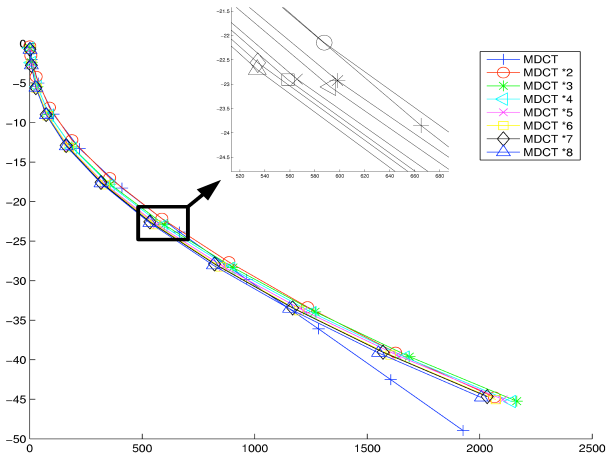


Figure 9: *R-D curve for pop music: influence of the scale for the MDCT (x: rate in Kbps; y: distortion in dB).*

## 6. ACKNOWLEDGEMENTS

The authors wish to thank the METISS team at IRISA for their support in the use and development of the MPTK project.

## 7. REFERENCES

[1] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Sig. Proc.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

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

[4] R. Ventura i Figueras, P. Vandergheynst, and P. Frossard, "Low rate and flexible image coding with redundant representations," *IEEE Trans. Image Processing*, vol. 15, no. 3, pp. 726–739, Mar. 2006.

[5] T. Verma, "A perceptually based audio signal model with application to scalable audio compression," Ph.D. dissertation, Stanford University, 2000.

[6] R. Vafin, "Towards flexible coding," Ph.D. dissertation, KTH Stockholm, 2004.

[7] L. Daudet, S. Molla, and B. Torrésani, "Towards a hybrid audio coder," in *Proc. Third Int. Conf. on Wavelet Analysis and Applications*, 2004, pp. 13–24.

[8] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "Iso/iec mpeg-2 advanced audio coding," in *101st Conv. Audio Eng. Soc.,* Los Angeles, USA, 1996, pp. 789–814.

[9] H. Malvar, "A modulated complex lapped transform and its applications to audio processing," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP'99),* Phoenix, USA, vol. 3, 1999, pp. 1421–1424.

[10] M. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Processing*, vol. 86, no. 3, 2006.

[11] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP'97),* Munich, Germany, vol. 3, 1997, pp. 2037–2040.

[12] S. Krstulovic and R. Gribonval, "GNU/GPL C++ Software, *the matching pursuit toolkit*," Retrieved June 29th, 2006, [Online] http://gforge.inria.fr/projects/mptk/.

[13] P. Frossard, P. Vandergheynst, R. Ventura i Figueras, and M. Kunt, "A posteriori quantization of progressive matching pursuit streams," *IEEE Trans. Sig. Proc.*, vol. 52, no. 2, pp. 525–535, Feb. 2004.

[14] S. G. Wilson, "Magnitude/phase quantization of independent Gaussian variates," *IEEE Trans. Communication*, vol. COM-28, no. 11, pp. 1924–1929, 1980.

[15] A. Moffat, R. M. Neal, and I. H. Witten, "Arithmetic coding revisited," *ACM Trans. Inf. Syst.*, vol. 16, no. 3, pp. 256–294, 1998.