SPARSE AND STRUCTURED DECOMPOSITIONS OF AUDIO SIGNALS IN OVERCOMPLETE SPACES

Laurent Daudet

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

ABSTRACT

We investigate the notion of "sparse decompositions" of audio signals in overcomplete spaces, ie when the number of basis functions is greater than the number of signal samples. We show that, with a low degree of overcompleteness (typically 2 or 3 times), it is possible to get good approximation of the signal that are sparse, provided that some "structural" information is taken into account, ie the localization of significant coefficients that appears to form clusters. This is illustrated with decompositions on a union of local cosines (MDCT) and discrete wavelets (DWT), that are shown to perform well on percussive signals, a class of signals that is difficult to sparsely represent on pure (local) Fourier bases. Finally, the obtained clusters of individuals atoms are shown to carry higher levels of information, such as a parametrization of partials or attacks, and this is potentially useful in an information retrieval context.

1. INTRODUCTION

Sparse decompositions of audio signals are extremely useful in many signal processing applications : compression, noise reduction, source separation, detection, etc. The goal is to decompose the signal onto a small number of basis functions, called "atoms" (typically time-frequency atoms, such as Gabor atoms or local cosines; or time-scale atoms, such as wavelets). The fundamental problem : the bigger your set of atoms (i.e. the more redundant), the more likely you will have a good match between your signal and the atoms, but the larger your set of possible solutions. The difficulty is to find the "optimal" (whatever that means) solution amongst them, usually a problem of high algorithmic complexity.

The work presented here tries to tackle this problem with practical, however suboptimal, methods. Optimality is here defined in terms of compression, i.e. bitrate vs. quality of approximation. Our space of representation is limited to the union of a basis of local cosines (MDCT) and a basis of dyadic discrete wavelets (DWT). This choice of basis is relevant for audio signals, especially for percussive sounds where the well-defined attacks are difficult to capture with purely Fourier-based approaches ; and where assumptions of harmonicity of the partials are not always verified. Based on the observation that significant coefficients are not randomly distributed across the time-frequency (-scale) plane but rather tend to form clusters, we do not select individual large atoms as significant, but groups of neighboring large atoms, called "molecules" or sound "micro-objects". In the MDCT domain, these form spectral lines; in the DWT domain, these form subtrees located around transient parts of the signal. Besides a significant reduction in the algorithmic complexity, "micro-objects" are more meaningful from an analysis point of view than isolated coefficients, and are cheap to encode. In short, this work presents a tentative framework for the unification of two developments that have emerged independently in recent years : sparse representations in overcomplete spaces [1], and structured representations (such as SPIHT [2] in image coding).

This paper is organized as follows : after a short introduction to sparse representations (part 2), we will detail how this can be implemented within the framework of sparse representations (part 3). Finally, the conclusion (part 4) will discuss potential applications for information processing or musical purposes.

2. SPARSE OVERCOMPLETE DECOMPOSITIONS

2.1. What does "sparse" mean ?

There are many definitions of sparsity for representations of signals. Here, we work in the context of representations (projections) of the signal on a set of pre-defined functions or "atoms" $\{b_n\}$, and the signal is simply "represented" by the set of mixing coefficients $\{\alpha_n\}_{n=0\ldots N-1}$.

$$x(t) = \sum_{n=0}^{N-1} \alpha_n b_n(t) \tag{1}$$

A given representation is said to be sparse if the number of non-zero coefficients is small compared to the dimension N of the space (the total number of samples in the signal). Mathematically speaking, this corresponds to a so-called L_0 measure of sparsity, that counts the amount of non-zero coefficients in a given set.

This definition can be generalized, and we can define a number of L_p sparsity measures, with $p > 0$, that represent how the "energy" is concentrated on a small number of coefficients :

$$L_p(\{\alpha_n\}) = \sum_{n=0}^{N-1} |\alpha_n|^p \tag{2}$$

Amongst these, the L_1 measure is a popular choice since some algorithms can be implemented with linear programming techniques. Note that L_2 is just a the standard measure of signal energy, and is invariant if the space of representation is an orthonormal basis, or a union thereof. Note that recent results [3] show that, if a given signal admits a "very sparse" representation under a given norm, then this representation is also the sparsest with respect to all L_p sparsity measures, $0 \leq p \leq 1$.

Also commonly used are sparsity measures based in the normalized information measure:

$$Li_p(\{\alpha_n\}) = (\frac{|\alpha_n|}{A})^p \sum_{n=0}^{N-1} \log(\frac{|\alpha_n|}{A})^p \quad (3)$$

where A is the L_2-norm of the α's.

2.2. Sparse approximations

In this study, we are going to relax the above constraints: we do not look for strictly sparse representations of signals, but for *approximations* of the signals that are sparse. More specifically, we assume that our signal is the sum of a signal that admits a sparse representation and a small noise component :

$$x(t) = \sum_{n=0}^{N-1} \alpha_n b_n(t) + \gamma(t) \quad (4)$$

where γ is our approximation error that is assumed to be small. In the context of audio compression, the tradeoff between quality of approximation and sparsity can be formalized by way of rate-distorsion curves.

2.3. Approximations in overcomplete spaces

In general, decompositions on othonormal spaces do not provide sparse representations: it is indeed very unlikely that everywhere the signal locally resembles the basis functions. Actually, the basis functions have by themselves little to do with the signal -neither do they sound like the signal - , and it it the union of (usually a large number of) them that allow a good approximation. Figure 1 shows a few of this indivudual atoms, taken from two popular choices of orthonormal basis : the Modified Discrete Cosine Transform (MDCT), and the Dyadic Wavelet Transform (DWT).

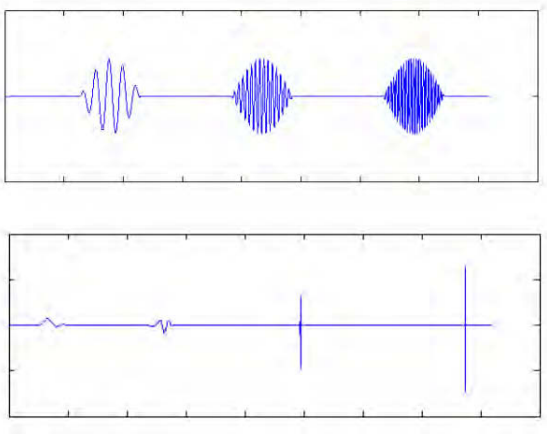

Figure 1: *Individual atoms : top: three time-frequency (MDCT) atoms; bottom : four time-scale (DWT) atoms.*

Therefore, it is usually preferable to use overcomplete spaces of approximation, ie. spaces with a number of basis functions that is greater than the dimension of the space. One may for instance choose Gabor frames [4], which can be seen as a set of discretized windowed Fourier atoms.

The problem with overcompleteness is that one loses the nice orthogonality principle that grants us the uniqueness of the decomposition : indeed, for a given signal there is an infinity of possible decompositions. The problem here is to find, amongst these decompositions, the one that is the most sparse, or that admits a sparse approximation, according to one of the above definition of sparsity.

2.4. Previous approaches

The above problem is in general not tractable, since its algorithmic complexity is huge (they belong to the NP-complete class of problems). However, there are approaches that give practical solutions, at a cost of suboptimality. Amongst these, let us mention the approach of Matching Pursuits [5] and Basis Pursuits [6]. The main drawback of these methods that their algorithmic complexity is still too high to be used on high-dimension signals such as audio signals.

3. STRUCTURED DECOMPOSITIONS

The method proposed here relies on the observation that large (ie significant) coefficients are not randomly located, but form structures, or clusters, in the parameter plane. Here we restrict ourselves to spaces with a small degree of overcompleteness, typically 2 times or 3 times ; and this allows for good visualization of the clusters. For the simplicity of the decompositions, we will choose the union of 2 (or 3) orthonormal bases. Preliminary results indicate that, if the bases are sufficiently different from each other, there is little to gain in choosing higher degrees of overcompleteness.

Our choice of basis will be the union of a basis of Modified Discrete Cosine Transform (MDCT) and a basis of Dyadic (discrete) Wavelet Transform (DWT). The MDCT is a popular choice in the audio coding community, since it is similar to a windowed Fourier transform while keeping the orthogonality for real signals. It is well adapted to the representation of locally tonal signals. The DWT, with short wavelets, is well known for its capacity to analyze transient portions of the signal, such as the attacks of percussive sounds.

With these notations our problem may be restated as follows. For a given signal x find the best overcomplete MDCT / DWT approximation of x :

$$x(t) = \sum_{n\in A} \alpha_n a_n(t) + \sum_{m\in B} \beta_m b_m(t) + \gamma(t) \quad (5)$$

where the a_n (resp. b_m) are the MDCT (resp. DWT) basis functions, and A and B the set of significant coefficients. Here, "best" means that the set of significant coefficients $\{\alpha_n\}_{n\in A} \bigcup \{\beta_m\}_{m\in B}$ is sparse according to our sparsity measure.

3.1. Tonal structures

In the MDCT time-frequency plane, large coefficients form tonal structures that appear along the spectral lines, as in Figure 2. On a practical point of view, tonal structures are detected as places where the MDCT *pseudo-spectrum* (a smoothed near-shift-invariant version of the MDCT spectrum $|\beta_m|$, described in [7]) is strongly

correlated across time. A tonal structure is then described as a set of MDCT coefficients with a width of 3 frequency bins, extending over a number of adjacent windows (see Figure 2).

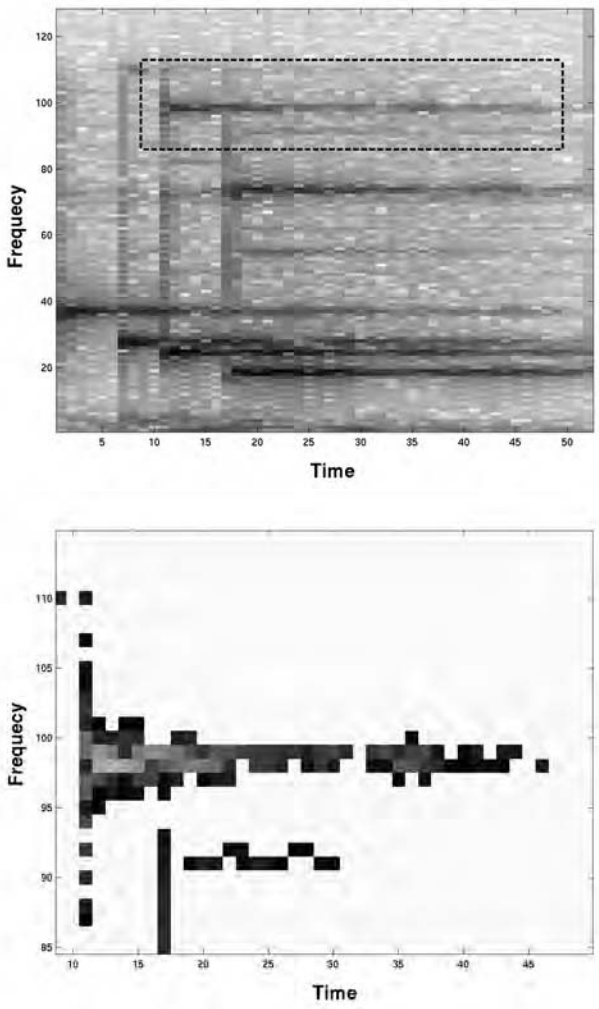

Figure 2: *Top: MDCT spectrogram of a typical percussive signal. Bottom: detail of the structure of significant tonal coefficients for the partial framed on the top figure.*

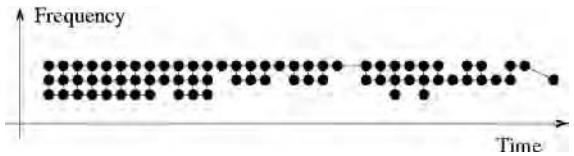

Figure 3: *"Tonal molecule" corresponding to the coefficients in the bottom plot of Figure 2, centered around the frequency bin 98. Note that the selected molecule has connected separate groups of significant coefficients (the gaps in time frames 32 and 45 corresponds to interference processes in the decomposition and do not bear a physical meaning).*

Figure 4 represents the MDCT spectrogram of a glockenspiel recording, and Figure 5 shows the detected tonal structures on the same file. Although some partials are not detected, the obtained pattern can be seen as a signature of the original sound, and indeed it is very close to it from a perceptive point of view, except at the onset of the notes.

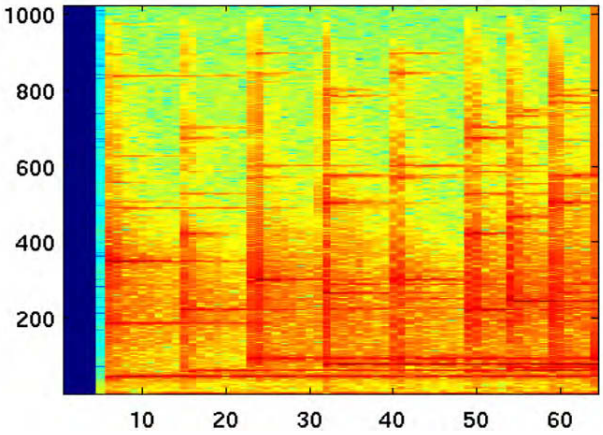

Figure 4: *Time-frequency MDCT spectrogram of a glockenspiel.*

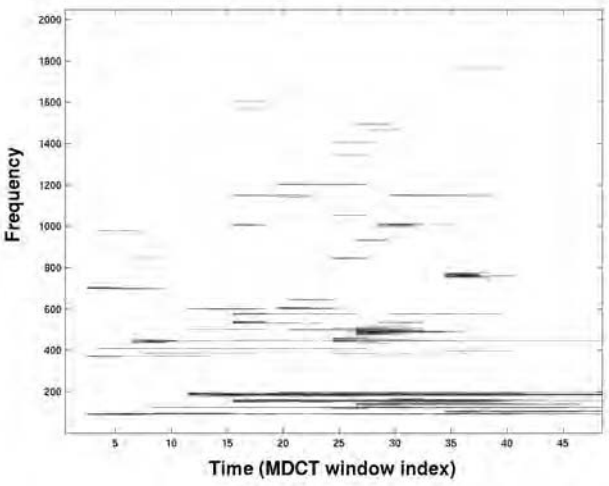

Figure 5: *Time-frequency MDCT spectrogram of the tonal molecules selected by the algorithm, same soundfile as Figure 4.*

3.2. Transient structures

Here, we work on the residual of the previous tonal extraction, that appears to contain mainly the transient sequences at the onsets of the notes. These will be represented as structures in the Dyadic Wavelet Transform (DWT) domain, which is organized in a dyadic tree structure in the time-scale plane. We use compact wavelet with

short support, such as Haar or Daubechies-4, since they provide good time localization properties [8].

Here, transient structures will be defined as "trees" in the time-scale dyadic plane where large coefficients are correlated *across scales* [9], as shown in Figure 6. When a tree is selected, we prune out the small coefficients with a top-down search, that ensures that the remaining trees are connected and fully connected to the largest scale, ie. the root of the dyadic tree. For very percussive signals, we observe that the selected structures correspond to the sharp attack transients of the sounds.

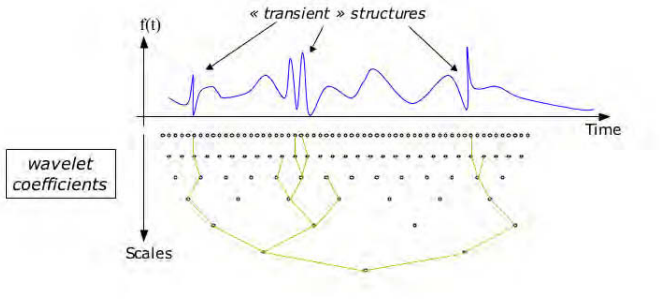

Figure 6: *In the dyadic time-scale plane, large wavelet coefficients cluster around branches of sparse trees.*

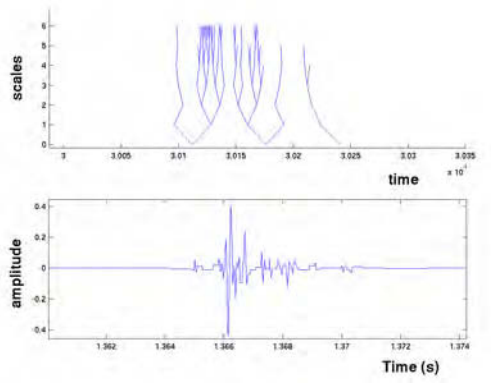

Figure 7: *Top : Pruned tree of selected wavelet atoms. Bottom : corresponding waveform.*

4. CONCLUSION AND PERSPECTIVES

This article has shown the capabilities of a new decomposition algorithm for sparse decompositions of signals in overcomplete spaces. The strength of this algorithm lies in the fact the structural information, through the localization of significant coefficients, is taken into account. This ensures that the obtained decompositions are intrinsically sparse. The algorithm is computationally very efficient, as compared to the original Matching Pursuit a large number of coefficients (typically between 5 and 50) is selected at every iteration. Furthermore, at every iteration we only need to update half of the scalar products between the signal and the basis, since we work on a union of two orthonormal basis. Higher degrees of overcompleteness give little improvement in the sparsity, with a large penalty in computational efficiency. In our model, the only improvement was noted when using in addition a basis of Modified Discrete Sines (MDST), that provides shift-invariance to our tonal part.

We have tested this algorithm on a number of signals. Preliminary results indicate that the best results are obtained on the sounds that are difficult to represent on classical Fourier-based spaces, namely the percussive signals. The main characteristics of these signals is that they are not harmonic (previous models extending Matching Pursuits have assumed harmonicity [10]) , and they are strongly non-stationary at their attacks (hence requiring a large number of Fourier coefficients). Although this model is very general, the sparsity of the decompositions degrades quickly when the local frequencies of the tonal partials is not constant, e.g. in the case of a frequency chirp or vibrato. More complex tonal decompositions can be implemented to account for these classes of signals.

Furthermore, one of the main drawbacks of the algorithm at present is that one needs *a priori* estimate of the relative tonal / transient importance. Future improvements will offer a simultaneous rather than sequential estimation of these two components, through a modified version of the Matching Pursuit [11]. However, this is expected to induce a significant increase in the computational requirements.

It is also interesting to note that the reconstructed signals from the obtained structures are meaningful from an auditive point of view (whereas individual atoms were not). Tonal structures sound like individual partials, and transient structures sound like the attacks of each note. This may be seen as intermediate level of representations, between the low-level time-frequency (/scale) planes, and the (relatively) high-level MIDI-like representation as individual notes (see Figure 8). The difficulty of automatic transcription of audio files makes these intermediate levels (called "micro-objects") an attractive option. This could also be useful for audio indexation purposes, used in information retrieval systems (for instance these objects bear information about the structure of the timbre, that is lost in the MIDI information).

Finally, one may wonder whether musical applications of this techniques are sensible. The author believes that this may be the case, since structures can be manipulated (sound transformations / effects) or completely created as reorganizations of molecules from different (possibly natural) sounds (this can be seen as an extension to the widely-used granular synthesis).

5. ACKNOWLEDGEMENTS

The author would like to thank Profs. B. Torrésani, Université de Provence, and M. Sandler, Queen Mary, University of London, for their contribution to this work.

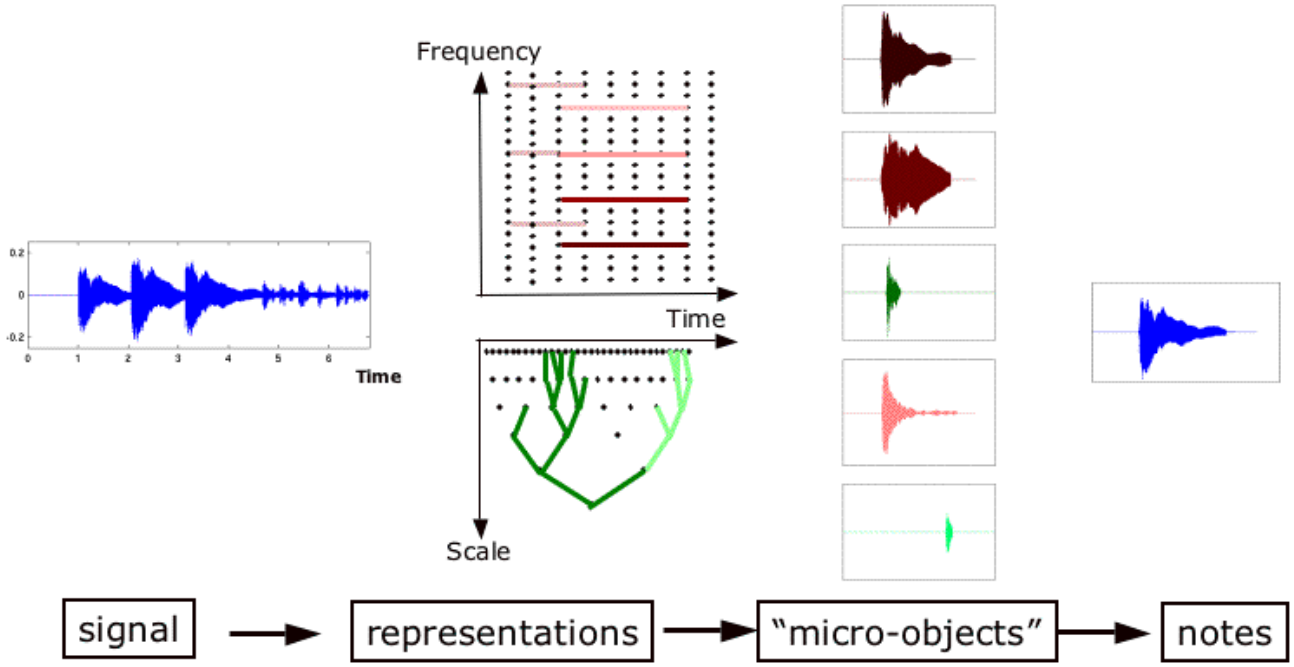

Figure 8: *Structured elements ("micro-objects") are seen as intermediate levels of representation between the time-frequency (or time-scale) spaces, and the individual notes.*

6. REFERENCES

[1] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," Tech. Rep. ICES 03-04, The University of Texas at Austin, February 2003.

[2] A. Said and W. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 243–250, 1996.

[3] X. Huo and D. Donoho, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Information Theory*, vol. 47, no. 7, 2001.

[4] P. Wolfe, M. Dörfler, and S. Godsill, "Multi-Gabor dictionaries for audio time-frequency analysis," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 43–46.

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

[6] H. Krim, S. Mallat, D. Donoho, and A. S. Willsky, "Best basis algorithm for signal enhancement," in *Proc. of the 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1995, pp. 1561–1564.

[7] Daudet L. and Sandler M., "MDCT analysis of sinusoids: explicit results and applications to coding artifacts reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 302–312, 2004.

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

[9] L. Daudet, S. Molla, and B. Torrésani, "Transient detection and encoding using wavelet coefficient trees," in *Proc. 18th Symposium GRETSI'01 on Signal and Image Processing, Toulouse*, 2001.

[10] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, Jan. 2003.

[11] L. Daudet, "Structured decompositions of signals with the Molecular Matching Pursuit," in preparation.