"Sparsification" of Audio Signals using the MDCT/IntMDCT and a Psychoacoustic Model – Application to Informed Audio Source Separation

Jonathan Pinel[1], Laurent Girin[1]

[1]*Grenoble Laboratory of Images, Speech, Signal and Automation (GIPSA-lab) CNRS UMR 5216 , Grenoble Institute of Technology, Grenoble, France*

Correspondence should be addressed to Laurent Girin (laurent.girin@gipsa-lab.grenoble-inp.fr)

ABSTRACT

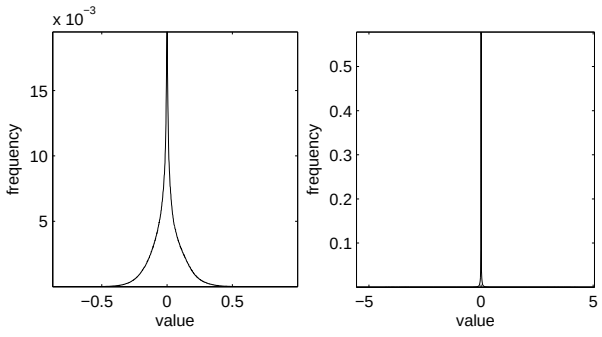
Sparse representations have proved a very useful tool in a variety of domain, *e.g.* speech/music source separation. As strictly sparse representations (in the sense of ℓ^0) are often impossible to achieve, other ways of studying signals sparsity have been proposed. In this paper, we revisit the irrelevance filtering analysis-synthesis approach proposed in (Balazs et al., IEEE Trans. ASLP, 18(1), 2010), where the TF coefficients that are below some masking threshold are set to zero. Instead of using the Gabor transform and a specific psychoacoustic model, we use tools directly inspired from perceptual audio coding, for instance MPEG-AAC. We show that significantly better "sparsification performances" are obtained on music signals, at lower computational cost. We then apply the sparsification process to the informed source separation (ISS) problem and show that it enables to significantly decrease the computational cost at the ISS decoder.

1. INTRODUCTION

Sparse representations of signals are representations where most of the signal coefficients are zero (or close to zero, in the weak sense). They are interesting in signal processing algorithms because the "useful" information to be processed is concentrated in a small proportion of the representation space. Additionally, in some cases, the coefficients corresponding to different signals can be disjoint. Sparsity is thus used in a large variety of audio applications, such as enhancement [1], source separation [2, 3, 4], or music transcription [5]. Sparsity is also the core principle of compressed sensing.

Many works have been dedicated to studying sparse representations [6, 7, 8]. In the audio case, signals may sometimes be sparse in the time domain (*e.g.*, with silent portions and/or Laplacian or Gaussian distribution) but it is well-known that they are far more sparse in the Time-Frequency (TF) domain (see Fig.1 for an example). When unicity of decomposition is not at stake, it has been shown that even more sparse representations can be found using overcomplete bases (e.g., "$8 * \text{MDCT}$" [9]).

The intuitive way of measuring the sparsity of a given



(a) time domain (b) time-frequency domain

Fig. 1: Samples distribution in the time (1a) and time-frequency (1b) domains for a rock music signal (each distribution is represented on its total support).

signal representation $\mathbf{x}$ is to use the ℓ^0 norm:[1]

$$\|\mathbf{x}\|_0 = \#\{j, x_j \neq 0\}. \quad (1)$$

However, in the strict sense of ℓ^0, "real-life" audio signals are generally not sparse at all: even if some coefficients are close to zero they are generally non zero. To overcome this problem, different measures have been proposed to study signals sparsity, using tools such as ℓ^p norms/quasinorms, tanh function, kurtosis or Gini index [10]. However as shown in [11], one must be careful when using those measures as they can lead to quite wrong assessments in certain circumstances.

In [12], the authors use a different approach: instead of looking for a representation that is strictly sparse (in the sense of ℓ^0), they use a representation where the audio signals are sparse in the weak sense (the TF domain using the Gabor transform[2]) and they set the lower coefficients to zero so that the representation becomes strictly sparse. Then, the "sparsified" time-domain signal in obtained using the inverse TF transform. To ensure preservation of audio quality, their algorithm uses a PsychoAcoustic Model (PAM) based on simultaneous frequency masking[3] to adjust the selection threshold of what they call the "irrelevance filter".

Although different in implementation and destination, this algorithm is related to Perceptual Audio Coding (PAC) [14], such as MPEG AAC [15]. PAC uses a PAM to estimate the maximum noise power that can be introduced by quantization at each frequency, and then adjusts the allocation of binary resource accordingly: the less relevant (*i.e.* more masked) a coefficient is, the coarser its quantization can be (and some low but non-zero coefficients are actually coded to zero).

Despite the more or less clear link between the irrelevance filter algorithm and PAC, or maybe because of this more or less clear link, the authors of [12] have used specific TF transform and PAM, with specific settings. In the present paper, we revisit their algorithm using tools that are much closer to the PAC approach and more efficient: instead of the Gabor transform, we use either a Modified Discrete Cosine Transform (MDCT) or its integer version (IntMDCT), similar to the one used in AAC (MPEG-4), and we use a PAM also directly inspired from AAC. Using those PAC tools, we show that a better "sparsification" of audio signals can be obtained when compared to the results of [12], with a simpler and more efficient signal decomposition/reconstruction framework.

Finally, we show that the proposed sparsification process can be exploited efficiently in the audio informed source separation (ISS) system proposed in [4]. In this system, the source signals are assumed to be available and the (linear instantaneous) mixing process is assumed to be controlled at the so-called ISS encoder. An a priori information about those sources and the mixing process is extracted and embedded within the mixture signal using a high-capacity data hiding technique [16]. At the ISS decoder, where only the mix signal is available, the side-information is extracted and used to help the separation process to deliver high-quality separated sources. We show that the sparsification process can be used as a pre-processing step for such ISS system to significantly reduce the computational cost of the separation.

This paper is organized as follows: Section 2 presents the irrelevance filter algorithm in a general manner. Section 3 provides the details of our implementation using the MDCT and AAC-inspired PAM. Section 4 and 5 present experiments and results, and conclusions are drawn in Section 6.

2. THE BASIC "IRRELEVANCE FILTER" ALGORITHM

The block diagram of the irrelevance filter/sparsification algorithm is presented in Fig. 2. A time-frequency representation $\mathbf{X}$ of the signal $\mathbf{x}$ is calculated (block ①) (using the Gabor transform in [12] and the (Int)MDCT in the present study). A masking threshold $\mathbf{M_x}$ is derived using a psychoacoustic model (block ②). The signal power spectral density (PSD) $\mathbf{S_x}$ is calculated (as the square of TF coefficients; block ③) and is compared with the masking threshold $\mathbf{M_x}$ (block ④). This yields a binary mask $\mathbf{m}$ that is used to set to zero the coefficients of $\mathbf{X}$ that are below the masking threshold (block ⑤):

$$\forall (t,f),\ m(t,f) = \begin{cases} 0 \text{ if } S_x(t,f) < M(t,f), \\ 1 \text{ else.} \end{cases} \quad (2)$$

[1] Actually ℓ^0 is not a norm but the limit of $(\ell^p)^p$ when $p \to 0$ with ℓ^p the usual norms/quasinorms. However this is not important here.

[2] Basically, the Gabor transform is a discrete Short-Term Fourier Transform (STFT) with specific conditions for analysis-synthesis signal reconstruction.

[3] When two pure tones or narrow-band noises close in frequency and with significantly different power are produced simultaneously, the human hear may perceive only the loudest [13]).

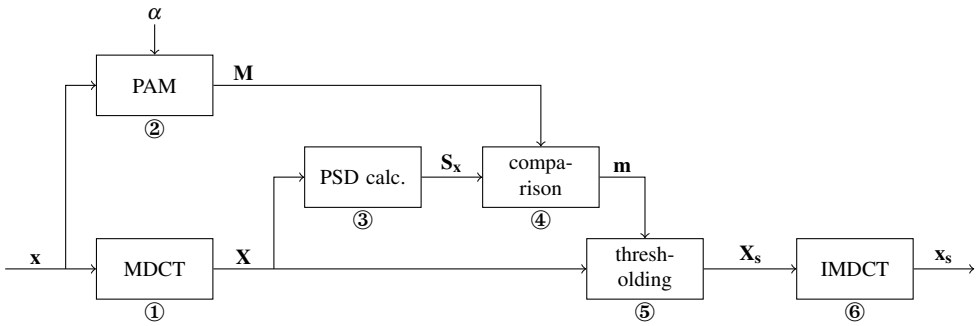


Fig. 2: Block diagram of the system.

$$\forall(t,f),\ X_s(t,f) = X(t,f)\cdot m(t,f). \tag{3}$$

This binary masking process leads to the sparsified representation $\mathbf{X_s}$ in the TF domain. The signal is finally transformed back into the time-domain sparsified signal $\mathbf{x_s}$ using the inverse TF transform (block ⑥).

Following the same line as in [12], a parameter α (in dB) controls an arbitrary translation of the masking threshold $\mathbf{M_x}$. As mentioned by the authors of [12], "this shift gives a conservative way to deal with uncertainty effects resulting from removing time-frequency components and with inaccuracies in the masking model." The goal is to experimentally find the α value that provides the sparsest representation without impairing the audio quality. Different values of α will be tested and discussed in Section 4.

3. IRRELEVANCE FILTER REVISITED

3.1. MDCT and IntMDCT

As mentioned before, the Gabor transform was used as the TF transform in [12]. In the present study, we propose to use the Modified Discrete Cosine Transform (MDCT) or its integer version (IntMDCT). We first explain the foundations of this proposition, and then we provide a very brief technical presentation of the MDCT/IntMDCT (technical details can be found in many papers, *e.g.*, the foundation paper [17].)

The MDCT is a real-valued time-frequency transform commonly used in audio signal processing, since it presents several interesting properties: i) the MDCT is critically sampled (it has the same overall number of coefficients in the time domain and the time-frequency domain), while being a lapped transform (with 50% overlap), and ii) under simple conditions, the MDCT ensures perfect reconstruction of the signal (when the transformed coefficients are not modified); more generally, even if the coefficients are modified, it has very good robustness against block effects during signal reconstruction by inverse MDCT (IMDCT). This property is intrinsically exploited in AAC compression to minimize the effects of MDCT coefficients quantization.

The combination of i) and ii) makes the MDCT a very interesting transform in the present study, since it has the potential for a very efficient decomposition/reconstruction of the signal, while ensuring the audio quality of the reconstructed sparsified signal. Also, and perhaps more importantly, the MDCT is characterized by the time-domain aliasing cancellation (TDAC) property: if the MDCT coefficients of a given signal are modified and the time-domain signal is reconstructed, the time-domain signal is modified as compared to the original signal, but when reapplying the transform to the modified signal the same modified coefficients are exactly recovered, despite the overlap process of the reconstruction. In other words the cross-frame influence of the overlap cancels out when returning to the transformed domain. Consequently, going anew to the TF domain from a sparsified reconstructed signal leads to a strictly sparse representation, *i.e.* the coefficients that have been zeroed by the irrelevance filter remain strictly equal to zero.

This property is not easily ensured by the Gabor transform: if a window with good frequency resolution is needed (as is the case here for applying the PAM and accurate irrelevance filter), a high level of redundancy

with specific windows is required to ensure perfect reconstruction (which is actually not exactly perfect but very close to it). In [12] the hop size is set to 1/8 of the analysis window size and therefore, it requires a quite larger amount of computations compared to the critically-sampled MDCT.

Finally, the preservation of zeroed coefficient values when chaining IMDCT and MDCT is verified if the time-domain signal is not modified. In practice, we deal with 16-bit PCM signals, and the mix signal is generally converted to this format (for storage and transmission; we assume that we work with uncompressed signals here). It can be shown that the 16-bit quantization introduces an additive white Gaussian noise (AWGN) on the MDCT coefficients, thus the zeroed coefficients can actually be corrupted by this noise. The use of the integer version of the MDCT (IntMDCT) can solve this problem. The IntMDCT is close to the MDCT (and it is also TDAC) but it is an integer-to-integer transform, so that PCM signals are transformed into integer MDCT coefficients, and conversely, integer coefficients are transformed back to integer PCM values. Therefore, the sparsification process directly leads to strictly sparse PCM signals, and we use this transform when the signals are to be stored in the .wav format.

Technically, the MDCT coefficients of a given frame t of N samples (N being even) of the host signal $\mathbf{x}$ is given for each $f \in \left[0, \frac{N}{2}-1\right]$ by:

$$X(t,f) = \frac{2}{\sqrt{N}} \sum_{n=0}^{N-1} x(t,n) w_a(n) \cos\left(\frac{2\pi}{N} n' f'\right), \quad (4)$$

where w_a is the analysis window, $x(t,n) = x\left(n + t\frac{N}{2}\right)$, $n' = n + \frac{N}{4} + \frac{1}{2}$, and $f' = f + \frac{1}{2}$. The inverse transformation (IMDCT) of the same frame is given for each $n \in [0, N-1]$ by:

$$\widetilde{x}(t,n) = \frac{2}{\sqrt{N}} w_s(n) \sum_{f=0}^{\frac{N}{2}-1} X(t,f) \cos\left(\frac{2\pi}{N} n' f'\right), \quad (5)$$

with w_s the synthesis window. To ensure perfect reconstruction of the signal, w_a and w_s must satisfy the Princen-Bradley conditions [17]. When $w_a = w_s = w$ (which is often the case), those conditions can be written as:

$$\forall n \in \left[0, \frac{N}{2}-1\right] \begin{cases} w^2(n) + w^2\left(n + \frac{N}{2}\right) = 1 \\ w(n) = w(N-1-n) \end{cases} \quad (6)$$

In this paper we use the Kaiser-Bessel Derived window, commonly used with the MDCT (for example in MPEG-AAC). The window length N is set to 2048, an usual choice for 44.1kHz audio signals.

As for the IntMDCT, as said earlier this transform is an integer approximation of the classical MDCT. The technique to achieve this approximation is called the *lifting scheme* and is described in several papers, e.g., [18]. The basic principle is to decompose the MDCT matrix (and the windowing process) into a product of matrices composed of 2×2 matrices of the form:

$$L_a = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}, \quad (7)$$

with $a \in \mathbb{R}$. This matrix L_a (which inverse is L_{-a}) represents the linear application:

$$\mathscr{L}_a : \begin{cases} \mathbb{R}^2 \longrightarrow \mathbb{R}^2 \\ (x,y) \longrightarrow (x, y + ax) \end{cases} \quad (8)$$

The principle of the lifting scheme is to replace all these applications by their integer approximation:

$$\text{int}\mathscr{L}_a : \begin{cases} \mathbb{Z}^2 \longrightarrow \mathbb{Z}^2 \\ (x,y) \longrightarrow (x, y + [ax]) \end{cases} \quad (9)$$

where $[.]$ denotes the integer rounding operation. The inverse of $\text{int}\mathscr{L}_a$ is $\text{int}\mathscr{L}_{-a}$ but as the application is an automorphism of $\mathbb{Z}^2$, it is exactly invertible, even with finite-accuracy computations.

3.2. PAM

The PAM used in our implementation is directly inspired from the PAM of the MPEG-AAC standard, and similar to the one that we used in [16]. The calculations are made in the time-frequency domain, however the transform used for the computations of the PAM is not the MDCT but the FFT. The main computations consist in a convolution of the FFT power spectrum of the host signal with a spreading function that models elementary frequency masking phenomenons, to obtain a first masking curve. This curve is then adjusted according to the

tonality of the signal.[4] After that, some pre-echo control is applied, and finally the threshold is translated by the "conservative factor" α (in dB), resulting in the final masking threshold $\mathbf{M_x}$. As mentioned earlier, different values of α will be tested in Section 4.

4. EXPERIMENTS ON SPARSIFICATION

4.1. Settings

As in [12], the algorithm has been evaluated in a listening experiment which goal is to find the value of the threshold translation parameter α. In those experiments, the MDCT is used; note that the two transforms provide quite close coefficient values, so that the sparsification results obtained with those two transforms are similar. The tested values of α are selected to -3, -4.5 and -6 dB after preliminary listening experiments. 10 musical excerpts (of different musical styles) of 5 seconds duration were used and 8 normal hearing subjects completed the experiments. The test was a classical ABX test: for each configuration (one excerpt and one alpha, unknown to the subject), the listener was presented a reference (the original signal) and two other signals in a random order. One of the signal was the reference and the other the sparsified signal. The subject had to chose which one was the original.

4.2. Results

For $\alpha = -6$ dB, the percentage of correct answer (*i.e.* identification of reference signal) is 53.75%; for $\alpha = -4.5$ dB, it is 55%; and for $\alpha = -3$ dB, it is 78.75%. Therefore, these results show that on average, the sparsification is inaudible for $\alpha = -6$ dB and -4.5 dB, or at the very least, it is inaudible for most of the excerpt/subject combinations. However, a difference can generally clearly be made for $\alpha = -3$ dB.[5]

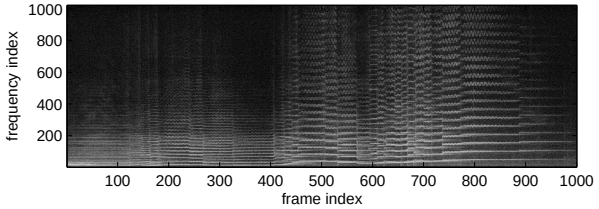
Table 1 shows the average proportion of suppressed coefficients and suppressed energy, for an extended set of α values. It can be seen from this table that $\alpha = -6$ dB (resp. $\alpha = -4.5$ dB) corresponds to a suppression of about 74% of the TF coefficients (resp. 78%), representing only less than 3% (resp. 4%) of the signal energy.

[4]The main reason why the PAM of the AAC works with the FFT and not the MDCT is because the phase information given by the FFT can be used to estimate the tonality of the signal in a better way than with the MDCT.

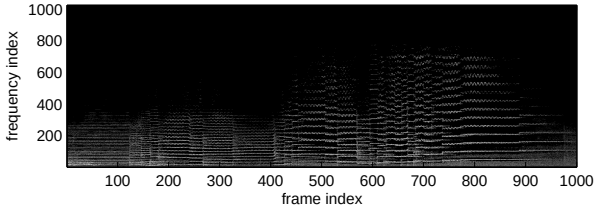
[5]The PAM is not the same as the one used in [12] so the range of α values and the inaudibility limit may slightly differ from the one reported in [12].

α (dB)	-3	-4.5	-6	-7.5	-9
Suppressed coefficients (%)	82.3	78.6	74.4	69.4	65.4
Suppressed energy (%)	4.8	3.4	2.4	1.7	1.2

Table 1: Proportion of Suppressed coefficients and suppressed energy for several values of α.



(a) Original signal



(b) Sparsified signal

Fig. 3: PSD (in dB) of a classical music excerpt before and after sparsification. The higher the amplitude the brighter the representation.

When analyzing those results, several remarks must be made:

- 74% of suppressed coefficients may seem a huge proportion, but it must be mentioned that a quite large proportion of those suppressed coefficients were generally already very low before sparsification (as they only represent a few percents of the total energy of the original signal). This is the case for most high-frequency coefficients for example, as illustrated by Fig. 3. This clearly illustrates the natural sparsity of audio signals, in the weak sense.
- Most of the test subjects were untrained listeners, hence even if the test seems to show that for $\alpha = -6$ dB, the sparsification is inaudible, this result is an average result valid for naive ears. The sparsification at $\alpha = -6$ dB may sometimes be audible for

"good" listeners after a few training period and several listenings. For example one of the author that is familiar with the test signals and sparsification effects can hear differences for α up to -7.5 dB for some excerpts. In particular, some specific effects on specific instruments can be detected after careful listening, *e.g.* typical musical noise on cymbals.

- However, even the "best" listeners (including the authors) could not find any difference for $\alpha = -9$ dB and lower. In this case an average proportion of about 65% of the coefficients are suppressed (representing 1.2% of the energy), which remains quite impressive.
- In [12], a proportion of 32% of suppressed Gabor coefficients, representing less than 2% of signal energy, is reported (however, quite surprisingly, 16-kHz music signals were considered in [12], leading to a lower amount of high-frequency low-energy coefficients). Therefore, our PAC-inspired modification of the irrelevance filter algorithm using MDCT and PAM seems to lead to a significant improvement in signal sparsification, whereas the analysis/synthesis process is much more efficient.

5. APPLICATION TO THE INFORMED SOURCE SEPARATION PROBLEM

5.1. Informed Source Separation of linear instantaneous mixtures

In previous papers [4, 19], we presented an Informed Source Separation (ISS) system that is able to separate $I > 2$ sources from stereo linear instantaneous mixtures. This system is designed with a specific coder-decoder configuration. At the coder, source signals are assumed to be available and the mixing process is assumed to be controlled. In each time-frequency (TF) bin, the mixture signal is assumed to be composed of at most two predominant sources, and the coder looks for the two sources that provide the best separation results by local 2×2 matrix inversion applied to MDCT coefficients (the other separated sources are set to zero). The side-information transmitted to the decoder is the index of those two predominant sources for each TF bin (and the mixing matrix parameters). In [4, 19] the side-information is embedded in the mix signal using a high-capacity data-hiding technique similar to the one presented in In previous papers [16]. At the decoder, where the original source signals are unknown, the extraction of this side-information enables to invert the (embedded) mixture in each TF bin to recover the source signals. This system is summarized in the diagram of Fig. 4. With such *informed* approach, it has been shown in [4, 19] that 5 instruments and singing voice signals can be efficiently separated from 2-channel stereo musical mixtures, with a quality that significantly overcomes the quality obtained by a semi-blind reference method and that enables separate manipulation of the source signals during stereo music restitution (*i.e.* remixing).

5.2. Sparsification as a pre-process for ISS

In the ISS system of [4, 19], the separation is made by local matrix inversion applied to the MDCT coefficients of the mixture signal, since the linear instantaneous mixture in the time-domain results in an identical linear instantaneous mixture on the source MDCT coefficients. Therefore, if the sparsification process of Section 3 is first applied to the signals involved in the separation process, this can lead to a reduced number of non-zero MDCT coefficients and a significant simplification of this separation process. Here, we propose to apply the sparsification process independently to each source signal at the coder level before the mixing process. This sparsification is assumed to have no consequence on the quality of each source signal taken separately, and it is very likely to have no (or not much) consequence on the quality of the resulting mix. For each TF bin, we can thus mention the following cases:

Case 1: The MDCT coefficient of each source has been zeroed by the sparsification process; this results in a zero-coefficient for each channel of the mixture; in such case, the separation process by matrix inversion is unnecessary, all separated sources can be directly set to zero.

Case 2: The MDCT coefficient of at least one source has not been zeroed; in that case, the matrix inversion is needed, but we can distinguish the following subcases:

Case 2.1: If the MDCT coefficient of only one or two sources are non-zero, then the separation leads to the exact reconstruction of the MDCT coefficient of every (sparsified) sources (as opposed to the original system in [4, 19], where

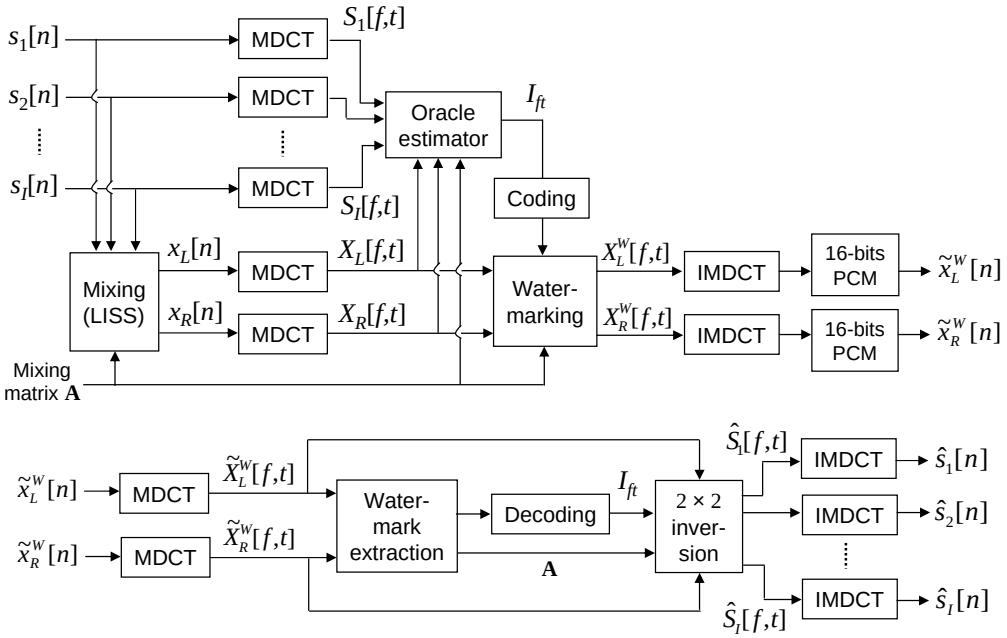


Fig. 4: Block diagram of the ISS system.

the weakest but non-zero sources can corrupt the inversion process).

Case 2.2: If the MDCT coefficient of more than two sources are non-zero, then the inversion of the two predominant sources can be disturbed by the other non-zeroed sources, as in [4, 19], but also as in [4, 19], those non-zero sources are assumed to be small compared to the two predominant sources and their influence on the separation process is assumed to be reasonable.

Note that after sparsification of the sources, the mixing process is carried out in the TF domain. Because the mixture coefficients are generally not integer (usually between 0 and 1), the resulting mixed TF coefficients are first rounded and then the mix signal is transformed back in the time-domain. This ensures the preservation of the values zeroed by the sparsification process (case 1 above), as mentioned in Section 3.1. The rounding error on non-zero coefficients is assumed to be negligible compared to their dynamics.

In the next subsection, we lead some experiments that measure the occurrence of each of those cases on realistic musical mixtures, and thus provide a first estimation of the computational gain that can be obtained at the ISS decoder

5.3. Experiments on sparsified ISS

The experiments were performed using 4 real music tracks of different styles (pop-rock, new-wave, funk, electro-jazz; duration 3, 4, 5 and 6 min) of 5 sources each (among guitar, bass, drums, lead vocals, saxophone, synthesizer, percussions). Table 2 shows the effect of the pre-mix source sparsification on the overlapping of the sources in the mix. It can be seen that on the average:

- Approx. 32% of the mix coefficients are zero (**Case 1**), leading directly to 32% of computational cost saving for the 2×2 matrix inversion of the separation process. Since approx. 65% of the coefficients are zeroed by the sparsification for each source taken separately, it can be deduced that approx. half of these sparsified coefficients are zeroed simultaneously for the 5 sources of the mix signal.
- Approx. 43% of the mix coefficients contain only 1

Number of non-zero sources	3	4	5
Average % of energy in the 2 predominant sources	97.92	96.53	95.70

Table 3: Sources overlapping in the mix after sparsification of each source.

or 2 sources (**Case 2.1**), meaning that the separation in those TF bin will be perfect (exact reconstruction of the sparsified sources).

To further complement those results, Table 3 shows that even in the case of more than 2 non-zero sources, the energy of the 2 most energetic sources represents on average more than 95% of the energy of all the sources (confirming the similar results presented in [19]). This confirms the two-predominant sources assumption on which the ISS method is based. In fact, as also shown in Table 2, the perfect reconstruction cases (**Case 1** and **Case 2.1**) represent only about 10% of the total energy of the mix signal. Therefore, the sparsification as a preprocess for informed source separation yields only few improvements in separation quality. Its main interest in the present case lies in the very significant computational gain (32% on the average).

6. CONCLUSIONS AND PERSPECTIVES

The present study shows that it is possible to largely sparsify musical signals in the TF domain (in the strict sense of ℓ^0, and in a conservative manner, *i.e.* with several back-and-forth transformations from time domain to TF domain), with more than 65% of zeroed coefficients without impairing audio quality. This is made possible by the use of PAC tools (MDCT and PAM), here borrowed from MPEG, which seem to be more efficient than the Gabor transform and specific PAM used in the previous inspiring study [12].

The sparsification as a pre-process has been applied within the informed source separation system presented in [4, 19]. The sparsification of instrument/voice signals before making the mix leads to approx. 1/3rd computational cost saving at the ISS decoder where the separation is processed (for 5 sources). In addition, this process contributes to reduce the overlapping of the source signals in the TF domain, and thus enables sparsity-based source separation algorithms to yield better results than when applied on normal mix.

7. ACKNOWLEDGMENTS

This work is supported by the French National Research Agency (ANR) as part of the DReaM project (ANR 09 CORD 006).

8. REFERENCES

[1] M. Moussallam, P. Leveau, and S.M. Aziz Sbai. Sound enhancement using sparse approximation with speclets. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 221 –224, 2010.

[2] H. Mohimani, M. Babaie-Zadeh, and C. Jutten. A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm. *Signal Processing, IEEE Transactions on*, 57(1):289 –301, 2009.

[3] R. Gribonval. Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 3, page III, 2002.

[4] M. Parvaix and L. Girin. Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 245 – 248, 2010.

[5] S.A. Abdallah and M.D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *Neural Networks, IEEE Transactions on*, 17(1):179 – 196, 2006.

[6] L. Daudet. Sparse and structured decompositions of signals with the molecular matching pursuit. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1808 –1816, 2006.

[7] A. Nesbit and M.D. Plumbley. Oracle estimation of adaptive cosine packet transforms for underdetermined audio source separation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008.

Number of non-zero sources	0	1	2	3	4	5
% of coefficients	32.64	25.10	18.50	12.88	7.75	3.13
% of energy	0	3.30	7.45	18.07	34.73	36.45

Table 2: Sources overlapping in the mix after sparsification of each source.

[8] Y.F.V. Tan and C. Févotte. A study of the effect of source sparsity for various transforms on blind audio source separation performance. In *Int. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS'05)*, Rennes, France, 2005.

[9] M.D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies. Sparse representations in audio and music: From coding to source separation. *Proceedings of the IEEE*, 98(6):995 –1005, 2010.

[10] N. Hurley and S. Rickard. Comparing measures of sparsity. In *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*, pages 55 –60, 2008.

[11] J. Karvanen and Cichocki A. Measuring sparseness of noisy signals. In *International Symposium on Independent Component Analysis (ICA) and Blind Source Separation (BSS)*, pages 125–130, Nara, Japan, 2003.

[12] P. Balazs, B. Laback, G. Eckel, and W.A. Deutsch. Frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(1):34 –49, 2010.

[13] E. Zwicker and U. Zwicker. *Psychoacoustics Facts and Models*. Springer-Verlag, 1990.

[14] T. Painter and A. Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451 –515, April 2000.

[15] K. Brandenburg and M. Bosi. Overview of MPEG audio: Current and future standards for low bit-rate audio coding. *J. Audio Eng. Soc.*, 45(1):4–21, 1997.

[16] J. Pinel, L. Girin, and C. Baras. A high-capacity watermarking technique for audio signals based on mdct-domain quantization. In *International Congress on Acoustics*, Sydney, Australia, 2010.

[17] J. Princen and A. Bradley. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(5):1153 – 1161, October 1986.

[18] R. Geiger, J. Herre, J. Koller, and K. Brandenburg. Intmdct - a link between perceptual and lossless audio coding. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, page II, 2002.

[19] M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2011, Pending Publication.