A Watermarking-Based Method for Informed Source Separation of Audio Signals with a Single Sensor

Mathieu Parvaix, *Student Member, IEEE,* Laurent Girin, and Jean-Marc Brossier

***Abstract*—In this paper, the issue of audio source separation from a single channel is addressed, *i.e.* the estimation of several source signals from a single observation of their mixture. This challenging problem is tackled with a specific two levels coder-decoder configuration. At the coder, source signals are assumed to be available before the mix is processed. Each source signal is characterized by a set of parameters that provide additional information useful for separation. We propose an original method using a watermarking technique to imperceptibly embed this information about the source signals into the mix signal. At the decoder, the watermark is extracted from the mix signal to enable an end-user who has no access to the original sources to separate these signals from their mixture. Hence, we call this separation process Informed Source Separation (ISS). Thereby, several instruments or voice signals can be segregated from a single piece of music to enable post-mixing processing such as volume control, echo addition, spatialization, or timbre transformation. Good performances are obtained for the separation of up to four source signals, from mixtures of speech or music signals. Promising results open up new perspectives in both under-determined source separation and audio watermarking domains.**

***Index Terms*—under-determined source separation, watermarking, audio processing, speech processing.**

I. INTRODUCTION

DURING the past twenty years, source separation has become one of the most challenging problems in signal processing. It can be stated as follows: source separation aims at estimating I unknown source signals, $s_i[n], i \in [1, I]$, from J observations of their mixtures $x_j[n], j \in [1, J]$. Blind source separation (BSS) [1] [2], where very little knowledge about the sources and the mixture configuration is available, has been intensively studied since the early 90's. The generality of the blind approach makes BSS techniques relevant for many application areas such as medical engineering [3] [4], communications [5], or antenna processing [6]. The (over)determined case, *i.e.* when $J \geq I$, is generally processed using the assumption of mutual independence of sources.

The *under-determined* (or degenerate) case, where fewer observations than sources are available, is of particular interest in audio processing since, in the typical mono or stereo configuration, many instruments and voices are to be separated from only two channels. To achieve source separation when $J < I$, many relevant techniques are based on the use of sparse representations of signals [7] [8] [9] [10] [11]. The basic idea is that in each localized region of the time-frequency (TF) plane, one single source is supposed to be more active than the other sources. Therefore, most methods are based on the transformation of the mixture signals in the TF domain where the separation is processed. Short-time Fourier transform were used in [11] before estimating the mixing matrix in the sparse transformed domain using a potential-function-based method. The separation method for non-stationnary signals introduced in [12] is based on TF distribution with the assumption that source signals are disjoint in the TF domain. Specific points of the TF plane corresponding to a single source are isolated and used to estimate the TF distribution of this source, from which sources waveforms are reconstructed. The DUET (Degenerate Unmixing Estimation Technique) BSS method was proposed in [7] to separate a large number of speech sources using a 2-channel mixture. The disjointness of source signals in the TF plane enables to determine the mixing parameters, and then to perform a relevant masking of the mixture's decomposition in the TF domain to estimate the source signals. This technique is assumed to be adapted to speech signals, particularly sparse in the TF domain. Another sparsity-based approach is introduced in [8] which addresses the problem of BSS of convolutive speech mixtures with a two-stage process. Hierarchical clustering is first used to estimate the mixing system, before sources are estimated under L1-norm minimization constraints. In [9], a Bayesian method was used to model the source signals using a sparse linear combination of Modified Discrete Cosine Transform (MDCT) atoms and a Markov hierarchical modeling to recreate the harmonic structure of sound signals. Note that, the under-determined case is also processed using the Computational Audio Scene Analysis (CASA) approach (see, e.g., [13] for an overview).

In the present study, we focus on audio source separation of linear instantaneous mixtures in a very specific configuration: in addition to the (single channel) observation of the (linear) mixture signal at the separation level (so-called here the decoder), source signals are assumed to be available at the mixing level (so-called here the encoder). This is quite an original and, at first sight, surprising configuration in the source separation framework, but not unnatural, since in some key audio applications, mixing and demixing can be processed separately by cooperative users. For instance, we address the audio-CD configuration: in a recording studio, the different tracks (corresponding to the different instruments and singing voice(s)) are recorded separately, and are then mixed in a very controlled way. At home, an end-user only has the mix signal available on a CD (actually, two stereo channels are available but the exploitation of stereo is not considered in the present paper). The objective is there to enable the separation of the

different elements of the audio scene, so that they can be manipulated separately: for example, the volume, the color and the spatialization of an instrument can be modified. This is an old dream of music lovers, that can be referred to as *active listening*.

In this paper, we propose a first approach to address this original problem at both coder and decoder levels. This method is based on the original combination of source separation with another major domain of signal processing, namely *watermarking*. Audio watermarking consists of embedding extra information within a signal in an inaudible manner. It is mainly dedicated to Digital Right Management (DRM). For instance, data identifying the designer or the owner of a digital media can be embedded by watermarking [14]. Extracting the watermark enables to prove the authenticity of the media. Since malicious users may try to remove or modify the watermark to appropriate the media, the embedding process is specifically designed to be robust against attacks such as compression, scaling or filtering [15] [16]. More recently, another type of watermarking emerged, using a signal as a channel for data transmission [17]. Properties of the host signal, known at the transmitter, can be used to increase the performances of the watermark detection at the decoder: the watermark is fitted to the host signal to optimize the data transmission rate while remaining inaudible. New applications of watermarking for data transmission have appeared, such as audio signal indexation [18], improvement of transmission systems [19] or document identification for broadcast monitoring [20]. Constraints on robustness and embedding capacity of the watermark are here different from DRM watermarking.

In the present study, we are closer to watermaking for data transmission than to watermarking for media protection. We take advantage of the knowledge of source signals at the encoder to extract a set of descriptors from these signals (see Fig. 1). These descriptors consist in parameters that describe the structure of the source signals, or their contribution to the mixture. This information is embedded into the mix signal using a high-capacity watermarking process. This technique exploits the defaults of the human hearing system to insert the information into TF coefficients of the mix signal. At the decoder, the source descriptors are extracted from the mix signal, and then used for the separation process. Since the method exploits the knowledge of unmixed signals, the corresponding framework can be labeled as "informed source separation" (ISS), as opposed to BSS.

This paper is organized as follows. Section II is a general overview of the method. In Section III a deepened description of the technical implementation is given, for each functional block of the system. Results of speech and music signals separation are given in Section 3. Finally, some perspectives are presented in Section 4.

II. General overview of the system

As already mentioned in the introduction and seen on Fig. 1, the general principle of the proposed ISS method is based on two main structures, a coder and a decoder. A more detailed functional schema is given on Fig. 2. In this section, a general overview of the entire system is given before presenting each functional block in details in the next section.

At the coder, the source signals $s_i[n], i \in [1, I]$ are assumed to be available, in addition to the mix signal $x[n]$. Therefore, characteristic parameters can be extracted from each source signal $s_i[n]$ and coded (block 4). These parameters are referred to as *descriptors* (as introduced in Section I), and they will be accurately defined in Section III-D. These descriptors are then embedded into the mix signal (block 5), using a "high-capacity" quantization-based watermarking technique (Section III-C). At the decoder, only the (watermarked) mix signal $x_W[n]$ is available. The descriptors are extracted from $x_W[n]$ using quantization again and decoding (blocks 11 and 12), before being used to separate each source signal from the mix signal (block 13). This is the heart of the process that will be described in details in Section III. Since the source signals generally strongly overlap in the time domain (TD), the separation process has to take place in a much sparser domain, *i.e.* a domain where each source is described by a sparse representation, that enables much less inter-source overlapping. This is the purpose of the time-frequency (TF) decomposition at the inputs of the coder (blocks 1 and 1') and at the input of the decoder (block 8). The dual operation, TF-to-TD synthesis, is done at the output of the coder (block 6) to provide the time samples of the watermarked mix signal. These samples are converted to the audio-CD format (16-bits uniform quantization at block 7, see Section III-C). The synthesis is also done at the outputs of the decoder (blocks 14) to reconstruct the estimated source signals from separated TF coefficients. The TF decomposition/resynthesis processes are described in Section III-A. Because of the need for a high embedding capacity and for a compact representation of descriptors, the TF coefficients are gathered into groups of neighboring coefficients: this is the molecular grouping of blocks 3, 3' and 10 (described in Section III-B and exploited in Section III-D). Finally, the role of the additional quantization blocks 2 and 9 will be detailed in the following. It can be seen from this general overview that the source signals characterization, the watermarking process, and the separation process are all carried out in the transformed domain.

III. The Informed Source Separation

In this section, we describe in details each functional block of the proposed ISS system. Note that when the role of a block is similar at the coder and at the decoder, it will only be described once for concision. The articulation between blocks has been given in the previous section.

A. Sparse decomposition

The signals to be segregated are issued from several voices and/or instruments playing a piece of music, and as a consequence, the different source signals that compose the additive mixture generally strongly overlap in the time domain. Also, audio sources can be non-stationary and have a large spectral variability, so that their respective contributions in the mix strongly depend on both time and

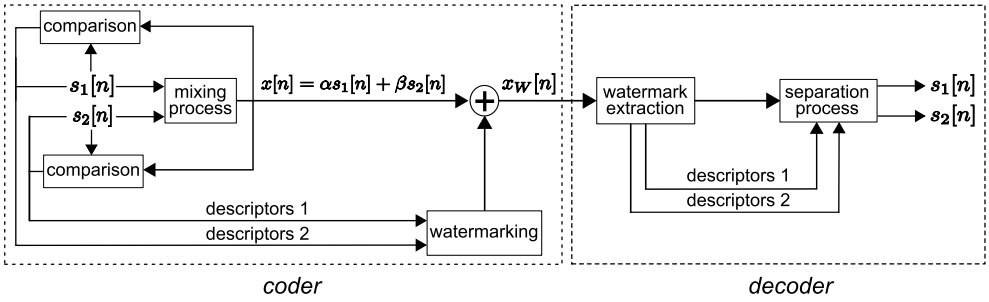

Fig. 1: Basic principle of the proposed watermarking-based source separation method.

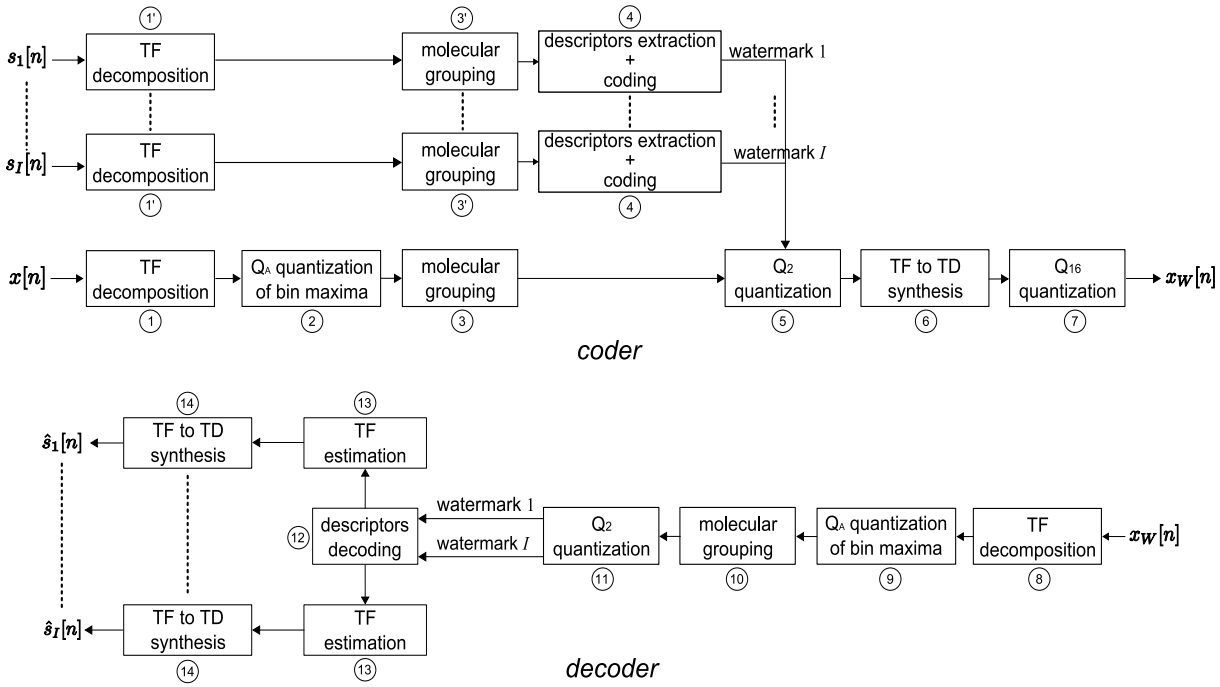

Fig. 2: Detailed Coder/Decoder structure.

frequency. A relevant transform domain is thus necessary to facilitate the separation process. As mentioned before, sparse decompositions such as TF decompositions appear to be particularly adapted to the case of under-determined audio source separation. Remind that we address here an extreme under-determined configuration with a single observation of the mixture available. For all these reasons, a TF decomposition is first applied to all signals. At the encoder, this decomposition is separately applied to each source signal (blocks 1' of Fig. 2) and to the mix signal (block 1) in order to extract the descriptors of each source and to embed them into the mixture. At the decoder, the decomposition is applied to the mix signal (block 8) to extract the descriptors and process the separation.

The Modified Discrete Cosine Transform (MDCT) [21] has been selected as the TF decomposition, for its ability to concentrate the energy of audio signals into limited regions of the TF plane, *i.e.* into few coefficients. This property, already exploited in audio coding, is expected to be useful for source separation. Since the MDCT is a linear transform, the source separation problem initially stated in the time domain is directly transposed in the transformed domain (*i.e.* the mixture remains linear). The chance sources have disjoint supports in the transformed domain is increased by the sparsity of this representation.

The MDCT is a discrete cosine transform applied on short overlapping signal frames. For a block of W consecutive samples of a signal $x\,[n]$ (beginning at $t \times W/2$, t being an integer denoting the time frame index), the MDCT transform results in $W/2$ frequency samples $m_t^x\,[f]$:

$$m_t^x\,[f] = \sum_{n=0}^{W-1} x\,[n + tW/2]\, w_a\,[n] \cos\left(\frac{2\pi}{W}\,(n+n_0)\left(f+\frac{1}{2}\right)\right) \quad (1)$$

where $f \in [0, W/2-1]$, $n_0 = (W/2+1)/2$, and w_a is the time analysis window of length W. For a time signal of N samples, with 50%-overlap between two consecutive frames, the MDCT decomposition results in a matrix $\mathcal{M}_x = \{m_t^x\,[f]\}, f \in [0, W/2-1], t \in [0, 2N/W]$, $2N/W+1$ being the number of time frames (see Fig. 3). We chose for w_a a Kaiser-Bessel derived (KBD) window of $W = 512$ samples

corresponding to 32ms of signal for a sampling frequency $f_e = 16$kHz and about 12ms if $f_e = 44.1$kHz. W is chosen to follow the dynamics of speech/audio signals and to have a frequency resolution suitable for the separation. The MDCT is an invertible transform, and a block of W time samples can be recovered from the $W/2$ coefficients $m_t^x[f]$ by inverse MDCT (IMDCT):

$$\tilde{x}[n+tW/2] = w_s[n]\frac{4}{W}\sum_{f=0}^{W/2-1} m_t^x[f]\cos\left(\frac{2\pi}{W}(n+n_0)\left(f+\frac{1}{2}\right)\right) \quad (2)$$

where w_s is the synthesis window chosen here to be identical to w_a. The complete signal is reconstructed by the overlap-add of successive frames. This process is applied at block 6 of Fig. 2 to generate the watermarked mixture signal from watermarked MDCT coefficients, and it is applied at block 14 for source signals reconstruction.

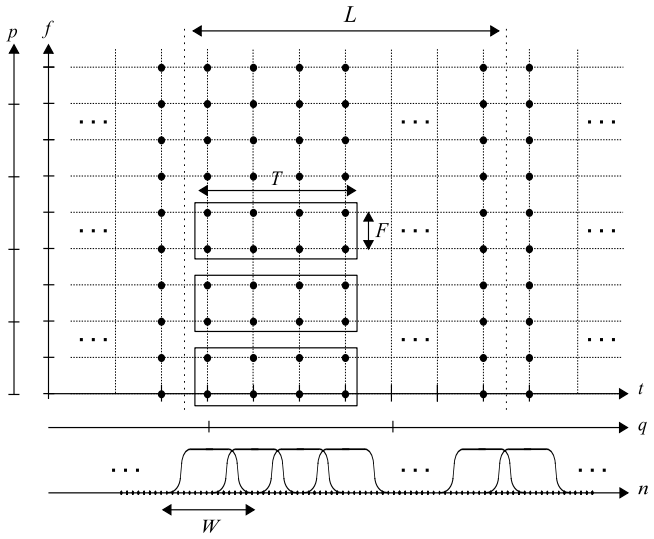

Fig. 3: Schematic representation of the time-frequency decomposition and molecular grouping.

If the MDCT is relevant to obtain a sparse representation of signals, it is non-realistic to assume that this decomposition will be sufficient to separate sources completely. Many regions of the TF plane remain where several sources overlap. Different methods have been proposed to address this problem within the multichannel case (using MDCT or other decomposition techniques), e.g., [22] [23] [24] [25]. As previously mentioned, the method developed in the present paper specifically addresses this problem by using extra information on source signals that can be inserted into the mixture.

B. Molecular grouping

The embedding of descriptors information into the mixture signal is made in the TF domain. As further detailed in Section III-D, the descriptors of source signals used for separation generally require a large embedding capacity from the host signal to be encoded with enough precision. Because of that, both the descriptors and embedding process cannot be defined at the level of a single MDCT coefficient. In other words, the contribution of the different sources to a given MDCT coefficient cannot be embedded into this MDCT coefficient alone. Therefore, we propose to gather MDCT coefficients into what is here referred to as *molecules*, after [26], [27] and [28]. A molecule of a signal x is a group of MDCT coefficients $m_t^x[f]$ located in the same neighborhood in the TF plane. Following the notations of Section III-A, a molecule M_{pq}^x located at the so-called *molecular* frequency and time bins (p,q) in the TF plane is defined by:

$$M_{pq}^x = \{m_t^x[f]\}_{\substack{f\in P=[(p-1)F,pF-1]\\ t\in Q=[(q-1)T,qT-1]}} \quad (3)$$

where T and F are the size of M_{pq}^x in time bins t and frequency bins f, respectively (see Fig. 3). Hence, a molecule is the elementary pattern from which the source signals descriptors used in the separation process are defined and computed. Molecular grouping is thus applied to each source signal (blocks 3'). A molecule is also chosen to be the elementary support of the watermark process. Molecular grouping is thus also applied to the mixture signal (block 3 of the encoder and block 10 of the decoder). In other words, the description, watermarking, and separation processes are all carried out on MDCT coefficients at the molecule level. The larger a mix signal molecule is, the larger is its watermarking capacity, and thus, the more information on source descriptors can be inserted into this molecule. Conversely, too large molecules cannot provide a good resolution for the description and thus the separation of overlapped source signals. Therefore a trade-off between watermarking capacity, descriptors accuracy, and quality of source separation needs to be found when setting the values of T and F.

C. Watermarking process

Even though watermarking is used in an original way in ISS, it has to verify the following basic principles: being imperceptible to human ear, being exactly retrievable at the decoder, and being resistant to attacks. In the specific context of ISS, for instance for the audio-CD application, we consider that the system does not have to cope with intentional malicious attacks. Since the purpose of the inserted watermark is to help to separate source signals, a user has no interest in damaging the watermark, contrary to the DRM case. The conversion to the audio-CD format is the only "attack" to be considered. This conversion corresponds to a 16-bits linear quantization of the time samples of the watermarked mixture signal (block 7 of Fig. 2). This leads to a slight modification of the corresponding MDCT coefficients and thus a possible damaging of the watermark[1]. Even though it must be taken into account, this attack has very limited effects in comparison to classical attacks against DRM watermarks. That is why we focus on a (low robustness) high embedding rate method, to encode the source signal descriptors. We selected a quantization-based watermarking method inspired by [29], in which the inserted message is carried by a modification of quantization levels of

[1]For the same reason, the proposed method is not suitable for compressed signals.

the host signal. The performances of this technique approach the theoretical bounds established in [30].

In the present study, the quantization principle is applied to the MDCT coefficients of the mixture signal. We take advantage of the fact that MDCT coefficients can be quantized with a quite coarse resolution without noticeable deterioration of the perceptual quality of the resulting time signal. This principle is exploited in compression algorithms such as MPEG [31]. In the proposed method, the watermark is embedded into MDCT coefficients using two uniform scalar quantizers, denoted Q_1 and Q_2. Actually, only Q_2 is used for practical quantization of MDCT coefficients in the system (as shown in Fig. 2, block 5). Q_1 is used as a virtual reference to simplify the presentation. We remind that a scalar quantizer is defined by two parameters: its resolution R, directly related to the number 2^R of quantization levels, and its scale factor A which is the maximum amplitude of the signal to encode. The quantization step between two successive levels is $\Delta = 2A/2^R$, and the value of the k^{th} level is $-A+k\times\Delta, k \in \left[0, 2^R-1\right]$. Q_1 and Q_2 have respective resolutions $R_1 < R_2$, and a common scale factor A (actually, the grids Q_1 and Q_2 are intertwined, as illustrated on Fig. 4). As the behavior of MDCT coefficients appears to be particularly dependent on their spectral location, the quantizers are defined for every frequency bin f. They are also updated every L time frames of signal (typically every 1.5 seconds, hence we take L=256), to take into account the non-stationarity of audio signals. Therefore, in the following, all quantizer parameters are functions of f and l which indexes the current L-block of time frames.

Watermarking a MDCT coefficient consists in the following process: the coefficient is first virtually quantized using the reference grid $Q_1(l,f)$. The watermarking corresponds to the modification of the reference level to an arbitrary "up-level" on the grid $Q_2(l,f)$ within the close neighborhood of the reference level ("sub-sampling" from grid $Q_2(l,f)$ to $Q_1(l,f)$ must lead to retrieve the reference level). This up-level is completely determined by the watermark code, as illustrated on Fig. 4. Hence, the embedding capacity of each MDCT coefficient at frequency f and within the l^{th} block is given by:

$$C(l,f) = R_2(l,f) - R_1(l,f). \tag{4}$$

This capacity does not depend on t (within the current L-block). The capacity of a $T \times F$ molecule M^x_{pq} is thus:

$$C_M(l,p) = T \times \sum_{f\in P} C(l,f),\ \ P = \left[(p-1)F, pF-1\right]. \tag{5}$$

Eq. (4) clearly shows that, to maximize $C(l,f)$, $R_1(l,f)$ has to be minimized and $R_2(l,f)$ has to be maximized. However, the inaudibility of the watermark induces a lower bound for $R_1(l,f)$, and the robustness to audio-CD format conversion induces an upper bound for $R_2(l,f)$. $R_1(l,f)$ has indeed to be high enough so that the Q_1-quantization of MDCT coefficients has no consequence on the audio quality of the mixture. To ensure this quality, the scale factor $A(l,f)$ is determined for each frequency bin f and each time block l from the maximum absolute value of the MDCT coefficients on the block: $m^x_{max}(l,f) = max_{t\in[(l-1)L, lL-1]}(|m^x_t\,[f]\,|)$. This ensures optimal dynamics of the signal inside the quantization grid. Given this adaptive scale factor, the reference resolution $R_1(l,f)$ is set to a fixed value in the present study, for simplicity purpose. To set this resolution, extensive listening tests were processed on a large set of various speech and music signals. 8-bits quantization of MDCT coefficients proved to have no audible consequences on the quality of resynthesized time signals, and provided signal-to-quantization noise ratios of about 40dB for speech and music signals. Therefore, we chose $R_1(l,f) = R_1 = 8$ bits for every frequency bin f (and every block l). Note that this resolution is quite comfortable regarding the standards used in transform-based compression algorithms, like MPEG for example. This suggests the possibility for a significant improvement of the embedding capacity in a future work. An adaptive resolution $R_1(l,f)$ may be determined for each signal frame using a psychoacoustic model since less precision is needed for high frequencies than for low frequencies.

As for the Q_2-quantization, it has to be resistant to the audio-CD format conversion. Indeed we remind that the time waveform of the watermarked signal is quantized using a 16-bits uniform quantizer at the final block of the coder. For the watermark to be correctly retrieved, the Q_2-quantization of the MDCT coefficients of the watermarked signal (block 11 of the decoder in Fig. 2) must provide the same result as the Q_2-quantization of the MDCT coefficients at the coder (block 5 in Fig. 2, i.e. before the 16-bits time-domain quantization). This results in an upper limit for $R_2(l,f)$. This conversion can indeed be considered as the addition of a uniform additive noise in the time domain. In the MDCT domain, this noise is observed to be approximately white Gaussian, in compliance with the central limit theorem. Its standard deviation σ_{16} is independent of f (and t), and it has been determined on a large database of music and speech signals. We assume that the amplitude of the maximum deviation of MDCT coefficients caused by the time-domain quantization remains lower than $4\sigma_{16}$. For the watermark to be preserved after audio-CD format conversion, the quantization step $\Delta_2(l,f) = 2A(l,f)/2^{R_2(l,f)}$ has to verify:

$$4\sigma_{16} < \frac{\Delta_2(l,f)}{2} \tag{6}$$

Hence the following condition on $R_2(l,f)$:

$$2^{R_2(l,f)} < \frac{A(l,f)}{4\ \sigma_{16}} \tag{7}$$

As shown in the result section, this condition generally enables significant embedding capacity, since the power of the 16-bits quantization noise is low.

Contrary to audio coding algorithms, in the proposed source separation method the quantizer parameters are not transmitted to the decoder, although they are necessary for the extraction of the source descriptors. As shown below, the quantizers parameters can be retrieved from the watermarked mix signal, despite of the modification of the MDCT coefficients by the watermark. This enables to allocate all the embedding capacity

to the source signals descriptors. Actually, only the scale factor $A(l,f)$ has to be retrieved since R_1 is fixed and $R_2(l,f)$ is specified by (7). To ensure that the decoding of the watermark is correct, for each frequency bin f and each time block l, $A(l,f)$ has to be the same at the coder and at the decoder. However, the value of $m^x_{max}(l,f)$, which determines $A(l,f)$, is modified by the watermarking process, as any other MDCT coefficient. A solution to this problem is to use a scale factor quantizer $Q_A(f)$, at both the coder (block 2 of Fig. 2) and the decoder (block 9) levels. This quantizer has a lower resolution than $Q_2(l,f)$ so that the Q_A-quantization of $m^x_{max}(l,f)$ is robust to the watermark. The result of this quantization is used as the scalar factor. Thus, a 6-bits quantizer $Q_A(f)$ is chosen, which is, i) known at both the encoder and the decoder, ii) defined for each frequency bin f, but independent of time, and iii) insensitive to the watermarking process. $Q_A(f)$ has been determined from a large set of music and speech signals (about 1 000 speech files from TIMIT [32] and about 10 minutes of music signals of various musical styles extracted from audio-CDs).

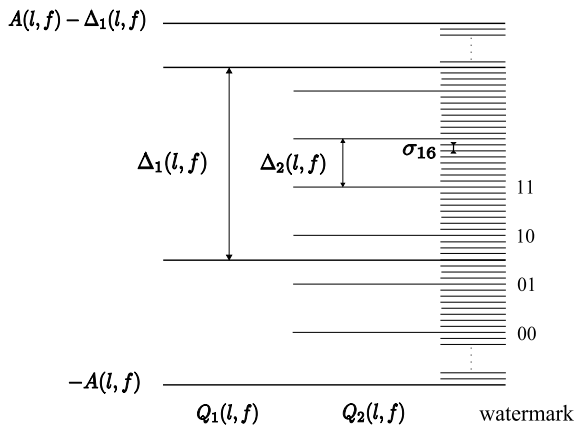

Fig. 4: Q_1- and Q_2-quantizers. $A(l,f)$ is provided by the quantization of $m^x_{max}(l,f)$ using $Q_A(f)$.

D. Source signals descriptors and their use for separation

In the method described throughout this paper, the watermark embedded into the mix signal contains descriptors of the source signals which are used for the separation process. The descriptors are defined at the molecule level. Two different sets of descriptors are used to describe a molecule of source signal depending on the embedding capacity of the corresponding molecule of the host mix signal (*i.e.* the molecule with same coordinates in the TF plane). In the case of a poor capacity, the descriptor of a given source is a single parameter that characterizes the local contribution of this source to the mixture, in terms of energy. In the case of a larger embedding capacity, the descriptors are a set of parameters directly encoding the source signal independently of the mixture. In the following we specify the descriptors definitions in both cases.

In the case of a poor capacity, the descriptor of the source signal s_i is defined as the energy ratio between a molecule of s_i and the corresponding molecule of the mix x:

$$E_{s_i/x}(p,q) = \frac{\sum\limits_{(f,t)\in\{P\times Q\}} |m^{s_i}_t[f]|^2}{\sum\limits_{(f,t)\in\{P\times Q\}} |m^x_t[f]|^2} \quad (8)$$

This ratio is quantized to $\check{E}_{s_i/x}(p,q)$ using a scalar quantizer and the resulting index is embedded as watermark information into the mix molecule M^x_{pq} (block 5 of Fig. 2) using $Q_2(l,f)$. At the decoder, the descriptor is extracted and decoded at blocks 11 and 12. The molecule $M^{s_i}_{pq}$ of the source signal s_i is then reconstructed (block 13) by the corresponding mix molecule M^x_{pq} weighted by (8) according to:

$$\hat{M}^{s_i}_{pq} = M^x_{pq} \times \sqrt{\check{E}_{s_i/x}(p,q)} \quad (9)$$

This is done for each molecule of each source signal. Therefore, the separation is based here on molecular energy segregation[2]. Note that the resolutions of descriptor quantizers are determined using the bit allocation tables presented in the next subsection. The scale factors are quantized using a 6 bits uniform quantizer known at both coder and decoder. The result is transmitted via watermarking because this information characterizes individual source signals, and it cannot be retrieved from the mix MDCT coefficients at the decoder. The cost of this additional embedding is assumed to be very low in comparison to the cost of descriptors, since those parameters are encoded only once in a given L-block of signal.

In the case several sources overlap, the *shape*[3] of the mix molecule can be quite different from the shape of the source molecule it is supposed to reconstruct. For this reason, if $C_M(l,p)$ is large enough, additional information describing the shape of source molecules is embedded. Coding the shape of a molecule would ideally correspond to code the amplitude of each of its MDCT coefficients. However, the required capacity is much larger than $C_M(l,p)$. Hence, we use *vector quantization* techniques (actually matrix quantization techniques) [33]: molecule prototypes are used as shape descriptors to strongly reduce the coding cost of these descriptors (compared to individual MDCT coefficient coding). Codebooks of (matrix) prototype shapes have been designed for each molecular frequency bin p using the Linde-Buzo-Gray algorithm [34] applied on a large training database of signal molecules (see Section IV-A). Note that these codebooks are adapted to a given instrument, or a given type of voice, using a corresponding database for training. This is expected to provide an improvement of separation results as in, e.g., [35]. For each kind of source signal (*i.e.* each voice and each instrument) a set of codebooks $\mathcal{D}_i = \{D^r_i(p)\}, i \in [1,I], r \in [6,10], p \in [1, W/2F]$ of different sizes 2^r is computed for each molecular frequency bin p. The shape descriptor of a source molecule is the closest prototype shape in the corresponding codebook, according to the Euclidean distance. The codebooks are assumed to be known at the

[2]A similar constant-power processing on regions of MDCT coefficients has been considered in the Gaussian models-based source separation method of [24], but on the frequency axis only, hence sub-bands processing, instead of TF molecules.

[3]The shape of a molecule is defined as the normalized set of amplitudes of the MDCT coefficients composing this molecule.

decoder, and the index of the prototype in the codebook is embedded as shape information (blocks 4 and 5). To increase the coding efficiency in term of quality/bitrate ratio, molecules are normalized before coding (hence, all codebooks contain normalized prototype molecules). A molecule $M_{pq}^{s_i}$ with mean (across the MDCT coefficients of the molecule) $\mu_{pq}^{s_i}$ and standard deviation (idem) $\sigma_{pq}^{s_i}$ is normalized by:

$$N_{pq}^{s_i} = \frac{M_{pq}^{s_i} - \mu_{pq}^{s_i}}{\sigma_{pq}^{s_i}} \tag{10}$$

Additional descriptors $\mu_{pq}^{s_i}$ and $\sigma_{pq}^{s_i}$ are encoded separately, as in e.g. [36], to provide $\check{\mu}_{pq}^{s_i}$ and $\check{\sigma}_{pq}^{s_i}$. This is done using scalar quantizers similar to those used to encode the energy descriptor (8) in the case of a poor capacity. As was done for the energy descriptor, the parameters of the additional mean/standard deviation quantizers are transmitted via watermarking once for each L-block. At the decoder, if $\check{N}_{l_q}$ denotes the molecule of the shape codebook $D_i^r(p)$ closest to $N_{pq}^{s_i}$, $M_{pq}^{s_i}$ is estimated by (blocks 12 and 13):

$$\hat{M}_{pq}^{s_i} = \check{\sigma}_{pq}^{s_i} \times \check{N}_{l_q} + \check{\mu}_{pq}^{s_i}. \tag{11}$$

Note that in this case, the decoding and estimation of a source molecule is a single process.

Finally, the source signals are reconstructed from the estimated source molecules by applying inverse MDCT (block 14, see Section III-A).

E. Bit allocation

Distributing the embedding capacity (5) among the different descriptors is a complex optimization problem. It depends on the number of sources to separate, their nature (an instrument and a speech signal may not need the same amount of data to be accurately described), the nature of the descriptors, and their required coding accuracy. Moreover, limits to the size of codebooks must be taken into account, since too small a codebook cannot efficiently represent the shape of a molecule, while too large a codebook would considerably increase the coding/decoding computational cost. In the present study, a series of bit allocation tables were determined empirically for different separation configurations, and validated by listening tests. For each configuration, the number of sources to separate from the mixture is fixed for the entire mixture signal. This number, as well as the corresponding bit allocation tables, are assumed to be known at both the encoder and the decoder.

Table I is an example of bit allocation tables designed for the separation of two, three, or four signals with molecules of size 2×4. Note that the allocation is given per molecule and per source, and each source signal is allocated the same resource. This table illustrates the switch between different coding configurations depending on the embedding capacity. When the embedding capacity is within the range 3–7 bits (for 2 signals) or 3–6 bits (for 3 or 4 signals), the ratio (8) is quantized (below 3 bits, the coding accuracy is too poor and it is better to directly use the mix molecule for source signal reconstruction). For a high capacity, within the range 14–26 bits (for 2 signals), 14–20 bits (for 3 signals), or 12–16 bits (for 4 signals), the mean-gain-shape (MGS) descriptors of (11) are quantized. The upper bound results from the limitation of both the shape codebooks size and the total embedding capacity. Between 7 or 8 and 11 or 13 bits, it was found relevant to split each 2×4 molecule into two 1×4 molecules and to allocate half of the embedding capacity to encode an energy ratio descriptor for each of the two sub-molecules.

Note finally that the streaming of embedded data among the MDCT coefficients of a molecule is a trivial task and is not detailed here.

TABLE I: Bit allocation tables (per source) for the separation of 2 to 4 signals with molecules of size 2×4. M, G and S, stand for the mean, the gain (standard deviation), and the shape in (11), respectively. When (8) is used, it is also denoted by G. In the case of molecule split, G is the energy ratio for the sub-molecule with lower frequency, and G' is the energy ratio for the sub-molecule with higher frequency. NS is the number of source signals to separate.

	2 signals				3 signals				4 signals			
C_M/NS	S	G	G'	M	S	G	G'	M	S	G	G'	M
1	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-
3	-	3	-	-	-	3	-	-	-	3	-	-
4	-	4	-	-	-	4	-	-	-	4	-	-
5	-	5	-	-	-	5	-	-	-	5	-	-
6	-	6	-	-	-	6	-	-	-	6	-	-
7	-	7	-	-	-	4	3	-	-	4	3	-
8	-	4	4	-	-	4	4	-	-	4	4	-
9	-	5	4	-	-	5	4	-	-	5	4	-
10	-	5	5	-	-	5	5	-	-	5	5	-
11	-	6	5	-	-	6	5	-	-	6	5	-
12	-	6	6	-	-	6	6	-	6	3	-	3
13	-	7	6	-	-	7	6	-	6	4	-	3
14	6	4	-	4	6	4	-	4	7	4	-	3
15	7	4	-	4	7	4	-	4	8	4	-	3
16	7	4	-	5	7	4	-	5	9	4	-	3
17	7	5	-	5	8	4	-	5	-	-	-	-
18	8	5	-	5	9	4	-	5	-	-	-	-
19	8	5	-	6	9	5	-	5	-	-	-	-
20	8	6	-	6	10	5	-	5	-	-	-	-
21	9	6	-	6	-	-	-	-	-	-	-	-
22	9	6	-	7	-	-	-	-	-	-	-	-
23	9	7	-	7	-	-	-	-	-	-	-	-
24	10	7	-	7	-	-	-	-	-	-	-	-
25	10	7	-	8	-	-	-	-	-	-	-	-
26	10	8	-	8	-	-	-	-	-	-	-	-

IV. EXPERIMENTS AND RESULTS

A. Data and performance measurements

Tests have been processed with both 16kHz-speech signals and 44.1kHz-music signals with voice+voice mixtures, and singing voice+instruments mixtures, with a number of sources varying from two to four. Speech signals are from the TIMIT database. Two male and two female English speakers were used for the tests. Ten sentences for each of 279 speakers (122 females, 157 males) were used to build the codebooks. For the music signals, instruments are a bass guitar, a piano, and drums (one track for the overall drum set), and the singing voice is from a female singer. Six pieces of music played together by all four sources (recorded separately in studio conditions) were used. Four of them were used to build the codebooks, and excerpts from the two remaining pieces were used to test

the separation (a slow piece and a fast one were chosen to ensure diversity of test signals). Details about the training data are given in Table II (note that the difference in data durations between instruments results from the suppression of silent sections). The ratio between the size of training data and the maximum size of the codebooks, is assumed to be always greater than 100. In all the following, *N*I+1S denotes a music mixture signal of *N* instruments + 1 singing voice, and *M*S denotes a speech mixture signal of *M* speakers.

TABLE II: Characteristics of the training corpus used to compute the shape codebooks.

source signal	duration (minutes)	size (number of molecules)	corpus size / codebook size
bass guitar	18.1	186732	182.3
female singer	10.8	109568	107.1
drums	16.7	171080	161.1
piano	15.1	154836	151.2
female speech	64.1	239961	234.3
male speech	80.1	300580	293.5

The quality of separated sources has been assessed by both listening tests in anechoic chamber and ISNR measures, as defined in [37]. ISNR is the difference (Improvement) of log-Signal-to-Noise power ratios between input and output of the separation process, where Signal is the source signal s_i to be separated, at the input Noise is the sum of all other signals in the mixture $\sum_{k=1,k\neq i}^{I} s_k$, and at the output Noise is the difference between the original signal s_i and estimated source signal $\hat{s}_i$:

$$\text{SNR}_{in}^{i} = 10log_{10}\left(\frac{\sum_n (s_i\,[n])^2}{\sum_{k\neq i}\sum_n (s_k\,[n])^2}\right) \quad (12)$$

$$\text{SNR}_{out}^{i} = 10log_{10}\left(\frac{\sum_n (s_i\,[n])^2}{\sum_n (s_i\,[n]-\hat{s}_i\,[n])^2}\right) \quad (13)$$

and for each source

$$\text{ISNR} = \text{SNR}_{out} - \text{SNR}_{in}. \quad (14)$$

B. Molecule size, embedding capacity and bit allocation

Concerning the size of a molecule, a trade-off has to be found between a sufficient capacity that enables to encode descriptors properly, and a correct resolution in the TF domain. Table III shows the separation results obtained for 5 different sizes of molecule, containing from 8 to 16 MDCT coefficients. For this experiment, only the unquantized version of (8) was used as a source descriptor, since we only test the "separation power" of the molecular decomposition. A 2×4 molecule size appears to be the best trade-off. Results are slightly lower for a 4×2 molecule, which confirms the importance of spectral dynamics for audio signals.

The top figure of Fig. 5 gives an example of embedding capacity (4) for each MDCT coefficient as a function of frequency bin f, obtained on a L-block of signal ($L \approx$ 1.5s), in the case of a mixture of four musical source signals.

TABLE III: Separation results vs. molecule size for 3 mix signals : 2 speakers (2S), 4 speakers (4S), and 3 instruments + 1 singing voice (3I+1S).

Mix	ISNR (dB)				
	2x4	4x2	3x4	4x3	4x4
2S	10.05	10.02	9.52	9.38	9.15
4S	10.40	10.17	9.93	9.88	9.65
3I+1S	14.97	14.50	14.54	14.27	14.11

The figure below gives the resulting capacity (5) of a 2×4 molecule. As for many music signals, most of the energy of the mixture is located in a specific frequency range, mostly the low frequencies. This results in a large embedding capacity, up to about 60 bits per molecule. In higher frequencies, fewer bits are available to insert the watermark (although in this example energy and capacity have a local raise around 9kHz). In high frequencies, the human ear is less sensitive, and source signals can be represented with a lower accuracy. Thus, the ratio between the accuracy of the separation information to be embedded and the embedding capacity is satisfying along all frequency bins.

Fig. 5 also shows the corresponding results of the bit allocation for the descriptors, in the case of the separation of three signals from the 4-sources mixture, using the bit allocation table of Table I (note that the three sources are allocated the same amount of bits; in a future work, bit allocation may be adapted to each kind of source signal). As a result of the energy distribution, shape codebooks are used to describe low frequencies molecules, here from 0 to 2kHz (molecular frequency bin 12), and also the medium frequencies around 8 to 10kHz (molecular frequency bin 47 to 58). From 2 to 8kHz and then from 10 to 16.5kHz the embedding capacity only allows to encode the gain of 1×4 sub-molecules (see Section III-E). From 16.5 to 19.5kHz (bin 96 to 114), only the gain of full 2×4 molecules can be encoded, and for higher frequencies, no information on the source signals can be embedded. Above 15kHz, it is likely that signal components are very low and thus inaudible, hence source separation may not be necessary.

Table IV shows the total capacity for several music and speech mixtures, summed over frequencies, and averaged over time. Hence, we obtain *average watermarking bitrates*. Rates of about 140 to 180 kbits/s are obtained for music and rates over 70 kbits/s are obtained for speech mixtures[4]. Those rates confirm the possibility to embed a large amount of information into audio signals.

TABLE IV: Embedding capacity for several music and speech mixtures.

mixture	3I+1S	3I	2I+1S	2I	1I+1S	4S	3S	2S
capacity (kbits/s)	177.5	159.6	181.4	178.6	144.0	73.0	72.3	73.3

[4]Note that the capacity does not necessarily increase with the number of sources because mixture signals are normalized in amplitude within $[-1,1[$. Adding a source to an existing mixture may lead to actually decrease the energy of some TF regions after renormalization.

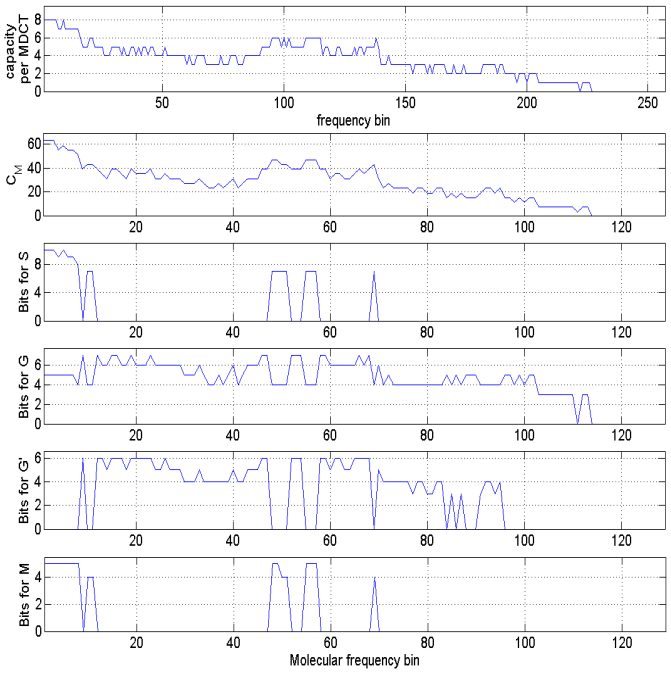

Fig. 5: Example of embedding capacity (per MDCT coefficient and per molecule for the two top figures) and bit allocation (per source) for a mixture of a bass guitar, a female singer, a piano, and a drum set, for the separation of 3 sources with molecules of size 2×4.

C. The complete separation system

In this section we report separation results obtained with the complete system. Note that it was verified that watermarking MDCT coefficients with $Q_2(l,f)$ does not impair the audio quality of the mix signal (this is not surprising since it was the case with $Q_1(l,f)$ in which $Q_2(l,f)$ is entwined). A first series of tests were processed to quantify the effects of using codebooks on separation results. A comparison was made between two configurations: in the first one, only the energy ratio descriptor (8) is used for all molecules, even if the embedding capacity is high (the ratio is then simply quantized with more accuracy). In the second configuration, all descriptors defined in (8) and (10) are used, and the embedding capacity is allocated according to Table I. Table V provides the results obtained for these two configurations (energy ratio (ER) *vs.* mean-gain-shape (MGS)), for the separation of four music signals from their mixture, and for the separation of two speech signals from a mixture of two female and two male speakers. In each case, two different test signals were used, with a duration of approximately 8s each. Table V exhibits ISNR scores from about 7dB up to about 20dB depending on the configuration, hence a good separation of both speech and instrument source signals. Listening tests performed in the ER configuration reveal remaining interferences on the separated sources. Table V proves the significant improvement obtained by using shape codebooks, with an average ISNR increase across configurations of 4.5dB (from +2.2dB for the piano which is the instrument with the "sparsest playing", to +7dB and +7.2dB for the drums and the bass respectively). This gain results in a very significant improvement in the quality of separated signals.

TABLE V: Performances for the separation of 4 signals in a 4-sources music mixture and 2 signals in a 4-sources speech mixture in the ER and MGS configurations.

Mix	Signals	ISNR (dB)	
		ER	MGS
Music	Bass	9.6	16.8
	Voice	7.3	11.7
	Drums	12.7	19.7
	Piano	12.1	14.3
Speech	Female1	12.0	15.0
	Male1	9.3	12.1

Now that the interest of using codebooks has been clearly assessed, we present extended results within the functional configuration of the proposed system (variable MGS allocation using Table I). Table VI gives the separation performance of the presented method for different settings. Two to four signals are separated among mixtures of three or four source signals. Hence, this table illustrates the influence of the number of sources in the mixture and the number of sources to separate. Note that the more numerous are source signals in a mixture, the lower is their input SNR, and the bigger is the risk these sources overlap in the TF plane. For the mixture of three sources, the input SNR is -1.9dB, -5.8dB, and -1.8dB, for the bass guitar, the drums and the piano, respectively, but when the singing voice is added with an input SNR of -1.1dB, the input SNRs of the other sources decrease to -5.8dB, -9.0dB and -5.7dB, respectively.

The main result emerging from Table VI is a good separation for all sources and from every mixture, including the most challenging case of 4 sources from a 4-sources mixture. In this latter case, ISNRs range from 11.7dB for the singing voice (this lowest ISNR is mainly due to a relatively high input SNR compared to other sources) to 19.7dB for the drums (which have the lowest input SNR). ISNR scores for the other instruments are 14.3dB and 16.8dB. For the case of 3 sources to be separated from the same 4-sources mixture, ISNR scores range within 15–22.3dB. The noticeable increase for both output SNRs and ISNRs from the separation of 4 sources to the separation of 3 sources (from 3I+1S) reveals the influence of sharing the embedding capacity between sources. Altogether, those results demonstrate a substantial separation, despite the difficulty of the task. Obviously, this is made possible by the informed aspect of the separation which ensures a good coding of the source signals, as opposed to usual source separation systems. An additional illustration is given in Fig. 6, which provides the 4 original and separated waveforms in the "4 out of 4" configuration. This figure shows that the separated waveforms are very close to the corresponding original waveforms.

The good separation is confirmed by listening tests, that show a very good rejection of competing sources: each instrument is clearly isolated. Of course, the quality is not perfect, and some level of musical noise remains (more or

less significant, depending on the source signal and on the number of sources to separate: for instance, the musical noise is lower for 3 sources than for 4 sources to separate). However, those results provide a very promising basis for the active listening application: separate signals can be added (up to a reasonable extend) and even subtracted from the mixture, directly in the time domain. The resulting quality of the "remastered" mixture is quite encouraging even if it is still far from transparency (i.e., as it would result from direct remastering of original source signals). The isolated source signals can be clearly enhanced or canceled, and the musical noise appears to be partly masked by the mixture. Sound samples for both "3 out of 4" and "4 out of 4" configurations (and also "2 out of 4", see below) can be downloaded at `http://www.icp.inpg.fr/~girin/WB-ISS-demo.rar`. The package includes original and watermarked mixtures, and original and separated source signals. All signals are correctly scaled so that the interested reader can directly process its own remastering using the mixture signal and separated sources.

As for the other settings, analyzing Table VI into more details shows an improvement of the ISNR scores from the extraction of 2 or 3 sources from a 3-sources mixture (3I) to the extraction of the same sources from a 4-sources mixture (3I+1S). This is mainly due to a drop of input SNRs with a larger number of sources within the mixture (see above). On the other hand, output SNRs are approximately the same for the separation of 2 or 3 sources from 3I+1S and from 3I. Observations made during the tests provide two explanations for this latter result: either the embedding capacities for the 3I+1S mixture and the 3I mixture are similar, or the embedding capacity for 3I+1S is larger but the descriptors coding accuracy reaches an upper limit (e.g., the limitation in codebooks size). This also explains the currently low (but real) increase of output SNRs, and corresponding modest increase in signal quality, from the separation of 3 sources to the separation of 2 sources (see Table VI and the samples). This suggests that using larger codebooks, or more refined coding schemes, when the number of sources to separate is small compared to the overall embedding capacity would result in better separation results. Note that in the case of the audio-CD application, the separation of more than 2 or 3 sources is targeted. Hence, we do not focus on improving the separation quality of a small number of source signals, but we rather tend to increase the number of sources to separate. However, the large room for improvement in the "2 out of 4" case is a very encouraging and reassuring point within the audio-CD framework, since a very basic way to exploit stereophony would be for example to embed one different pair of sources in each channel and use the knowledge of the mixture process to control the contribution of each source signal in both channels, leading to 4 high-quality separate source signals. This principle is valid for a larger number of sources as long as their single-channel coding quality is satisfying. This is part of our future works.

TABLE VI: Separation results for music mixtures in term of ISNR (dB).

sources to separate	mix of 4 sources (b,s,d,p)		mix of 3 sources (b,d,p)	
	SNR_{out}	ISNR	SNR_{out}	ISNR
bass guitar (b)	11.1	16.8	-	-
female singer (s)	10.6	11.7	-	-
drums (d)	10.7	19.7	-	-
piano (p)	8.6	14.3	-	-
bass guitar (b)	13.1	18.9	13.1	15.0
female singer (s)	14.4	15.0	-	-
drums (d)	13.2	22,3	13.1	18.9
piano (p)	-	-	10.3	12.1
bass guitar (b)	13.4	19.1	13.4	16.3
piano (p)	11.3	17.0	11.3	14.0
female singer (s)	15.1	15.7	-	-
drums (d)	14.7	23.7	-	-

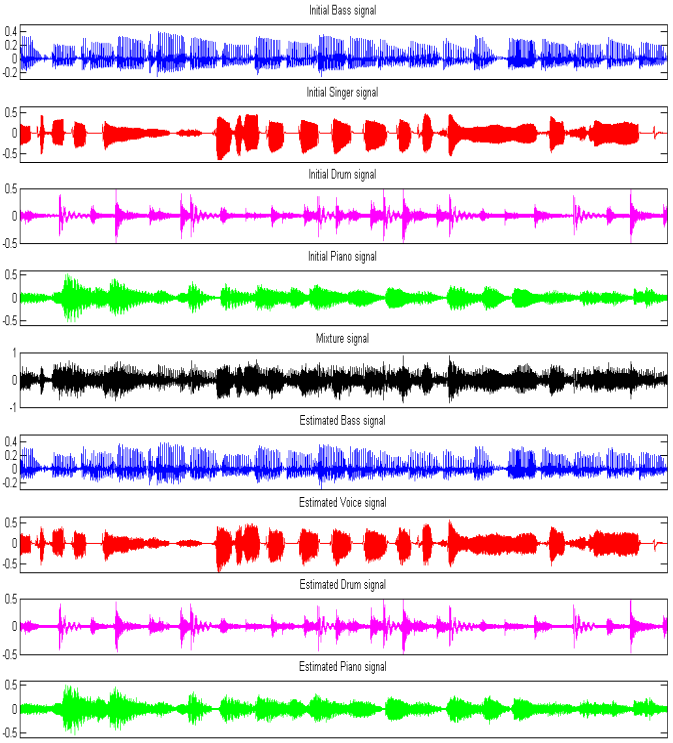

Fig. 6: Waveforms of 4 original and separated signals from the 3I+1S mixture (the mixture is in the center plot).

V. CONCLUSION

The separation approach described in this paper does not belong to classical source separation methods. Contrary to the BSS framework, in the Informed Source Separation, source signals are available before the mix is processed, and specific (but important) applications such as audio-CD "active-listening" are targeted. In the present study, promising preliminary results on speech and music signals separation from a single-channel mixture have been reported.

Many improvements can be made to the proposed method. Future work will first focus on a larger range of music instruments to separate. In the audio-CD application, codebooks could be learned using the source signals later mixed into the

tracks of the CD (i.e. training data = test data), whereas in the present study, test signals were not included in the codebooks training set. A better reconstruction of each source signal is then expectable. Also, the use of refined TF decomposition algorithm such as Molecular Matching Pursuit [27] [28] and/or multi-resolution decomposition will also be considered to better take benefit of both the audio signals sparsity and the properties of the auditory system, and provide a more accurate separation. Those decomposition techniques can be combined with refined strategies for source signal coding (including variable bit allocation among the sources) depending on the local composition of the mixture in the TF plane and not only on the local capacity, as was the case in the present work. As for the watermarking part, a psychoacoustic model will also be used to determine the resolution of quantizer $Q_1(l, f)$ for every L-block of the mixture signal. This is expected to significantly increase the embedding capacity. Finally, the ISS framework can be extended to the multi-channel source separation framework, beginning with the stereo case for the audio-CD application. Beyond the basic schema mentioned at the end of Section IV-C, the combination of the proposed source coding/watermarking method with spatial processing inspired from, e.g., [7] [11] [25] is a promising trail that is currently being investigated.

THIS WORK IS SUPPORTED BY THE FRENCH NATIONAL RESEARCH AGENCY (ANR) CONTINT PROGRAM, AS A PART OF THE DREAM PROJECT.

REFERENCES

[1] J. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[2] C. Jutten and J. Herault, "Blind separation of sources, part 1: An adaptative algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[3] L. D. Lathauwer, D. Callaerts, B. D. Moor, and J. Vandewalle, "Fetal electrocardiogram extraction by source subspace separation," in *IEEE Workshop on High Order Stat.*, 1995, pp. 134–138.

[4] V. Zarzoso and A. K. Nandi, "Noninvasive fetal electrocardiogram extraction: Blind source separation versus adaptative noise cancellation," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 1, pp. 12–18, 2001.

[5] K. Abed-Meraim, S. Attallah, T. J. Lim, and M. Damen, "A blind interference canceller in DS-CDMA," in *IEEE Int. Symp. Spread Spectrum Techniques and Applicat.*, vol. 2, 2000, pp. 358–362.

[6] S. Shamsunder and G. Giannakis, "Modeling of non-gaussian array data using cumulants: DOA estimation of more sources with less sensors," *Signal Process.*, vol. 30, no. 3, pp. 279–297, 1993.

[7] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[8] S. Winter, W. Kellermann, H. Sawada, and S. Makino, *Underdetermined Blind Source Separation of Convolutive Mixtures by Hierarchical Clustering and L1-Norm Minimization*. in S. Makino and al. (Eds), Blind Source Separation, Springer, 2007, pp. 271–304.

[9] C. Fevotte, *Bayesian Audio Source Separation*. in S. Makino and al. (Eds), Blind Source Separation, Springer, 2007, pp. 305–335.

[10] M. Zibulevski and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[11] P. Bofill and M. Zibulevski, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[12] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using time-frequency distributions," *J. Applied Signal Process.*, vol. 2005, no. 17, pp. 2828–2847, 2005.

[13] D. Rosenthal and H. Okuno, Eds., *Computational Auditory Scene Analysis*. Inc. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1998.

[14] D. Kirovski and H. Malvar, "Spread-spectrum audio watermarking: Requirements, applications, limitations," in *IEEE Int. Workshop on Multimedia Signal Process.*, vol. 51, no. 4, 2001, pp. 219–224.

[15] R. Garcia, "Digital watermarking of audio signals using psychoacoustic auditory model and spread spectrum theory," in *107th Conv. Audio Eng. Soc. (AES)*, 1999.

[16] I. Cox, F. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, 1997.

[17] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proc. IEEE*, vol. 87, no. 7, pp. 1127–1141, 1999.

[18] R. Tachibana, "Audio watermarking for live performance," in *SPIE Electron. Imaging: Security and Watermarking of Multimedia Content V*, vol. 5020, 2003, pp. 32–43.

[19] B. Chen and C.-E. Sundberg, "Digital audio broadcasting in the FM band by means of contiguous band insertion and precanceling techniques," *IEEE Trans. Commun.*, vol. 48, no. 10, pp. 1634–1637, 2000.

[20] T. Nakamura, R. Tachibana, and S. Kobayashi, "Automatic music monitoring and boundary detection for broadcast using audio watermarking," in *SPIE Electron. Imaging: Security and Watermarking of Multimedia Content IV*, 2002.

[21] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 64, no. 5, pp. 1153–1161, 1986.

[22] A. Aïssa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, "Underdetermined blind separation of nondisjoint sources in the time-frequency domain," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 897–907, 2007.

[23] J. Woodruff and B. Pardo, "Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings," *EURASIP J. Advances in Signal Process.*, vol. 2007, 2007, article ID 86369.

[24] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models," in *IEEE Workshop Applicat. Signal Process. to Audio and Acoust.*, 2005.

[25] H. Viste and G. Evangelista, "A method for separation of overlapping partials based on similarity of temporal envelopes in multichannel mixtures," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 3, pp. 1051–1061, 2006.

[26] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–112, 2003.

[27] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 5, 2006.

[28] M. Parvaix, S. Krishnan, and C. Ioana, "An audio watermarking method based on molecular matching pursuit," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 1721–1724.

[29] B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[30] M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[31] K. Brandenburg and M. Bosi, "Overview of MPEG audio: Current and future standards for low bit-rate audio coding," *J. Audio Eng. Soc.*, vol. 45, no. 1, pp. 4–21, 1997.

[32] J. S. Garofolo, L. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993, linguistic Data Consortium, Philadelphia.

[33] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Pub., 1992.

[34] Y. Linde, A. Buzo, and R. M. Gray, "Algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[35] N. Cho, Y. Shiu, and C.-C. J. Kuo, "Audio source separation with matching pursuit and content-adaptative dictionaries," in *IEEE Workshop Applicat. of Signal Process. to Audio and Acoust.*, 2007.

[36] K. L. Oehler and R. M. Gray, "Mean-gain-shape vector quantization," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1993.

[37] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, pp. 1462–1469, 2005.