Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Hybrid coding/indexing strategy for informed source separation of linear instantaneous under-determined audio mixtures

Mathieu Parvaix (1), Laurent Girin (1), Laurent Daudet (2), Jonathan Pinel (1), Cléo Baras (1)

(1) Grenoble Laboratory of Images, Speech, Signal and Automation (GIPSA-lab) CNRS UMR 5216 , Grenoble Institute of Technology, Grenoble, France
(2) Institut Langevin, CNRS UMR 7587, ESPCI, Paris, France.

PACS: 43.75Zz, 43.75Xz, 43.60Fg, 43.60Uv, 43.60Ek

ABSTRACT

We present a system for under-determined source separation of non-stationary audio signals from a stereo 2-channel linear instantaneous mixture. This system is dedicated to isolate the different instruments/voices of a piece of music, so that an end-user can separately manipulate those source signals. The problem is addressed with a specific informed approach, that is implemented with a coder corresponding to the step of music production, and a separate decoder corresponding to the step of signal restitution. At the coder, source signals are assumed to be available, and are used to i) generate the stereo 2-channel mix signal, and ii) extract a small amount of distinctive features embedded into the mix signal using an inaudible watermarking technique. At the decoder, extracting and exploiting the watermark from the transmitted mix signal enables an end-user who has no direct access to the original source signals to separate these source signals from the mix signal. In the present study, we propose a new hybrid system that merges two techniques of informed source separation: a subset of the source signals are encoded using a "sources-channel coding" approach, and another subset are selected for local inversion of the mixture. The respective codes and indexes are transmitted to the decoder using a new high-capacity watermarking technique. At the decoder, the encoded source signals are decoded and then subtracted from the mixture signal, before local inversion of the remaining sub-mixture signal leads to the estimation of the second subset of source signals. This hybrid separation technique enables to efficiently combine the advantages of both coding and inversion approaches. We report experiments with 5 different source signals separated from stereo mixtures, with a remarkable quality, enabling separate manipulation during music restitution.

INTRODUCTION

Source separation aims at recovering an unobserved vector of I source signals $\mathbf{s} = [s_1, \ldots, s_I]^{\mathrm{T}}$, from J observations of their mixtures $\mathbf{x} = [x_1, \ldots, x_J]^{\mathrm{T}}$ ($[.]^{\mathrm{T}}$ denotes the transpose operator). This problem has a variety of configurations. When both the source signals and the mixing process are unknown, it is referred to as Blind Source Separation (BSS). If at any time index n the mixture signal can be expressed as

$$\mathbf{x}[n] = \mathbf{A} \cdot \mathbf{s}[n] \tag{1}$$

where the $J \times I$ mixing matrix $\mathbf{A}$ is composed of constant gains, the mixture is *linear instantaneous* and *stationary* (LIS). This models the case where all the sources reach the sensors at the same time but potentially with different intensities. If the direct-path delays (resp. multiple propagation delays and attenuations) from sources to sensors are taken into account, the mixture is called *anechoic* (resp. *convolutive*).

The number of source signals and observations also condition the problem. When $J \geq I$, the mixture is said to be (over) determined, and the source signals can be estimated by searching for the inverse (or pseudo-inverse) unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ up to a scaling and permutation of the rows. Major contributions to Blind Source Separation (BSS) and related field of Independent Components Analysis (ICA) developed for (over) determined mixtures can be found in [4] [9] [6]. The *underdetermined* case $J < I$ is more delicate to solve, since the mixing matrix cannot be directly inverted. However, this case is of particular interest in audio (music) processing since most audio mixtures are composed of more than two sources, while the number of observations J is often limited to one or two (respectively for the mono and stereo configurations). Separating source signals from such music mixtures is a major challenge since it would enable to separately manipulate the different elements of the audio scene, e.g., modifying the volume, the color or the spatialization of an instrument, a process referred to as *active listening* or *remixing*. In the present paper, we will focus on the underdetermined source separation (USS) of music signals from LIS stereo mixtures.

To achieve such underdetermined separation, many relevant techniques take advantage of the *sparse* nature of audio source signals. These methods make the (realistic) assumption that, in a given basis, source signals have a parsimonious representation, *i.e.* most of the source coefficients are close to zero. A direct consequence of sparsity is the limitation of sources overlapping in the appropriate basis since the probability that several sources are simultaneously active is low. For most music signals, the time-frequency domain is a natural appropriate domain for exploiting sparsity [22] [7]. As a consequence, many USS techniques are based on sparse time-frequency (TF) representations of signals. For instance, in [11] the authors make the assumption that the non-stationnary source signals

to be separated 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. In [3], LIS stereo mixtures of speech and music signals are decomposed using Short-Time Fourier Transform (STFT). The mixing matrix is estimated using a clustering algorithm, then a shortest path procedure is used to select the two predominant sources for 2×2 mixture inversion in each TF bin. Note that estimating for instance five music source signals from a stereo mixture using "pure" BSS methods is a very complex task. Obtaining a satisfactory audio quality of the individual separated signals is then quite unrealistic.

In [15] [16] we introduced the concept of *Informed* Source Separation (ISS), with a specific coder-decoder configuration corresponding to the distinct steps of signal production (e.g. music recording/mixing in studio) and signal restitution (e.g. audio-CD at home). In addition to the mixture signals at the separation level (so-called here the decoder), source signals are assumed to be available at the mixing level (so-called here the coder). A limited set of parameters are extracted from the source signals at the coder, and are imperceptibly embedded into the mixture signals using a watermarking technique. This latter exploits the masking properties of the human hearing system to insert a high-capacity message into TF coefficients of the mix signal. Extracting and exploiting the watermark at the decoder enables an end-user who has no direct access to the original source signals (but only to the watermarked mixture signals), to separate these source signals from the mixture signals, and thus to manipulate them individually for remixing/active listening[1]. Hence, the informed approach makes the assumption that inserting side-information (even quite limited) into the mixture can heavily increase the separation performances. This appears as a mean to overcome the difficulties of BSS in the complex under-determined configuration.

As for BSS, different approaches exist for ISS, depending on the assumptions made on the source signals (mutual independence, sparsity) and on the mixture (linear, instantaneous, anechoic, convolutive, over/under-determined). As a result, the side-information embedded into the mixture, and the way it is used for the separation process may differ for the different configurations. In [15] [16], a *single-channel* LIS mixture of (speech or music) source signals was processed. A joint "source-channel" coding approach was followed: codebooks of molecular prototypes (i.e. matrices of neighboring TF coefficients) were generated and used to represent the source signals. The codes resulting from encoding the source signals with those prototypes were embedded into the mixture signals. Hence, source separation directly rested upon source encoding/decoding, and we can refer to this method as Source-Coding ISS (SC-ISS). In [14], we first addressed the problem of ISS for underdetermined LIS stereo 2-channel mixtures of music signals. The ISS system proposed in [14] jointly exploits the sparsity of source signals in the TF domain and the spatial information provided by the multi-channel dimension of the mixture. The watermarked side-information is here reduced to the indexes of the locally (i.e. in each TF region) predominant sources, as provided by an analysis of the source signals at the coder. Hence, we call such approach Index-based ISS (I-ISS). At the decoder, extracting the watermarked indexes enables to compute estimates of the source signals by local inversion of the mixing system.

The present paper is based on a combination of both SC-ISS and I-ISS methods. The hybrid method presented in this paper jointly exploits the sparsity of source signals in the TF domain, and the coding approach of source signals detailled in [16]. The main purpose of this combination is to increase the separation performances obtained by both the SC-ISS and I-ISS systems taken separately. Since I-ISS is based on the sparsity of source signals, the less signals overlap, the higher are the separation performances by inversion of the mixture, and a straightforward solution to reduce the overlapping of the sources is to reduce the dimension of the mixture: this task can be processed by SC-ISS. Therefore, source signals are divided into two categories: a subset of source signals is chosen to be encoded using vector quantization, while side-information identifying the locally predominant sources is extracted from the sub-mixture composed by the remaining sources. As in our previous work [13], the indexes of predominant sources (I-ISS) are estimated using an optimal (a posteriori) criterion inspired by the Oracle estimators developed in [20] [12]. Both coding and index watermarks are then embedded using a "high-capacity" Quantization Index Modulation of the TF coefficients of the mixture signals. The use of a psycho-acoustic model, inspired by MPEG-AAC, to determine the available watermarking capacity enables to embed up to about 250kbits/s/channel depending on the musical content. Note that constraints on robustness and embedding capacity of the watermark are here different from Digital Right Management (DRM) watermarking. In the present case, watermarking is used for metadata transmission, and thus offers a low robustness to malicious attack (irrelevant in the present configuration). Moreover, this capacity can be automatically adjusted to the need. The watermarking system is presented in the same congress [17], thus it is not detailed in the present paper. At the decoder, a first step in the separation process consists in decoding the sources encoded by SC-ISS, before removing these estimates from the (watermarked) mixture signal. A local inversion of the resulting sub-mixture by I-ISS is then processed using the decoded index of the locally predominant sources. By reducing the dimension of the mixture, the risk of source overlapping is decreased and the performances of source estimation by I-ISS are consequently increased.

This paper is organized as follows. First, a general overview of the proposed method is given. A detailed description of the technical implementation is then presented. Separation results for music signals are then given, and finally, some conclusions and perspectives are presented.

GENERAL OVERVIEW OF THE SYSTEM

Fig. 1 presents the diagram of the proposed stereo Index/Source-Coding-based Informed Source Separation technique. In this section, we first present a general overview of the entire system before presenting the functional blocks in more details in the next section.

The general principle of a coder-decoder configuration introduced in the mono configuration of [16] and also used in [14] is retained in the present work. The mixing process at Block 1 of Fig.1 is a LIS 2-channel stereo mixture[2] of I non-stationnary source signals, as given by (1) for $J = 2$. At the coder, source signals are first divided into two categories. I_c sources, denoted s_1 to s_{I_c}, are chosen to be encoded by SC-ISS (the core of this process lies in Block 3). The remaining $I - I_c$ sources, denoted s_{I_c+1} to s_I, are estimated by I-ISS (the core of this process lies in Blocks 4 and 5). Since the I-ISS strongly relies on the sparsity of source signals, the overall process is carried out in the TF domain where audio sources are much sparser

[1] Note that, so far, the proposed ISS methods are not robust to compression (bitrate reduction or dynamic compression), they are dedicated to audio-CD/wav music signals

[2] In this paper we focus on 2-channel mixture since it is of particular interest in music processing. However, the main principles of the process remain valid for $2 < J < I$, and we use the general notation J for preserving this generality when possible.

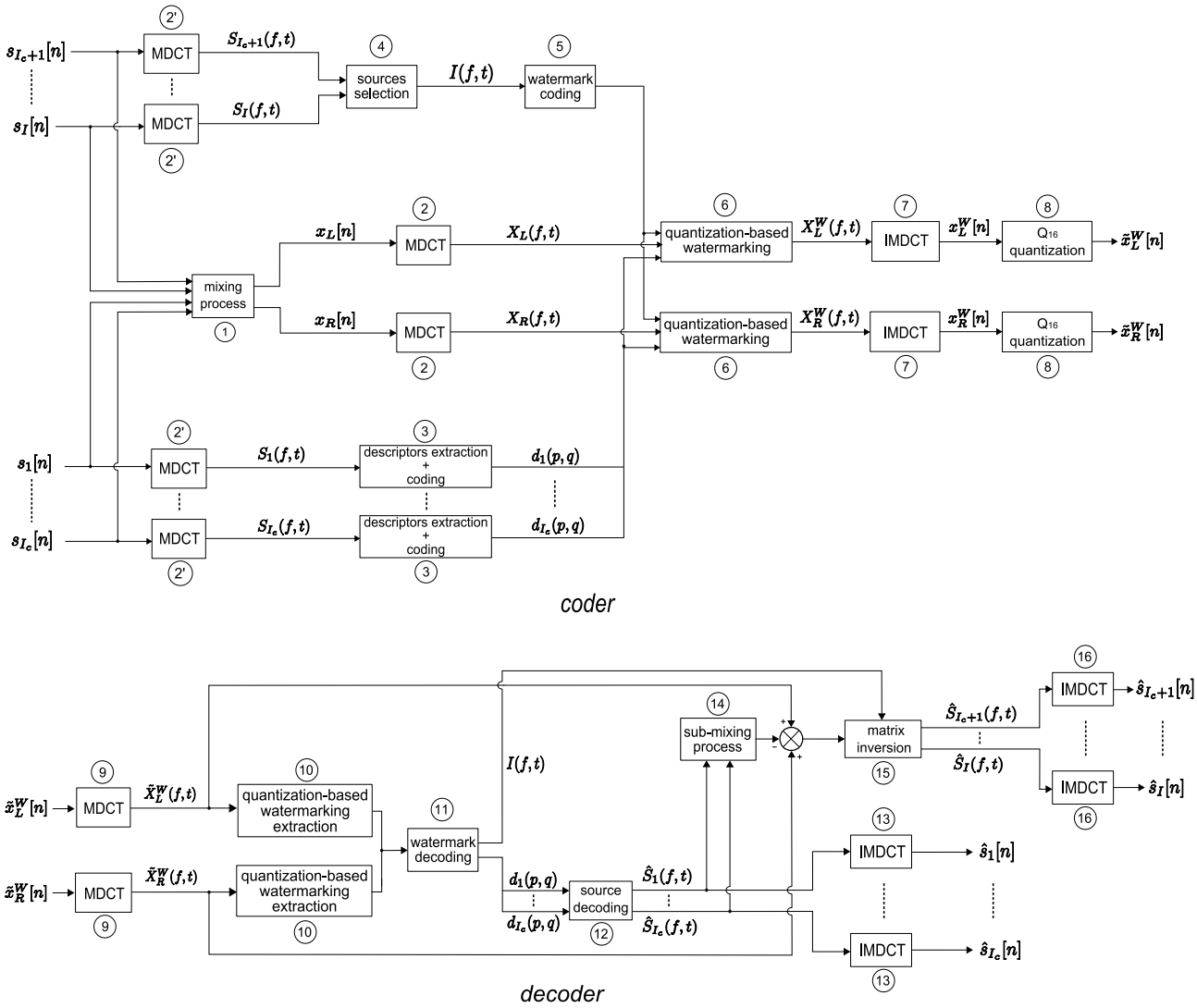


Figure 1: Detailed structure of both coder and decoder for SC-I-ISS.

than in the time domain. Therefore the Modified Discrete Cosine Transform (MDCT) is used at the input of the coder, at Blocks 2 for the stereo mixture signal, and at Blocks 2' for the individual source signals. Encoding parameters for sources s_1 to s_{I_c} (referred to as descriptors) are extracted from each source signal and coded (Block 3). In parallel, the I-ISS process at Block 4 consists in selecting the most relevant sources in each TF bin for further separation by local inversion of the mixture. The combination of *indexes* of the selected sources constitutes the side-information to be coded (Block 5) in the I-ISS part of the process. Then both watermarks of descriptors (SC-ISS) and indexes (I-ISS) are gathered before being embedded into the mixture signal (Blocks 6) by a quantization-based watermarking technique. The dual operation of Block 2, time-domain signal synthesis by inverse MDCT (IMDCT), is carried out at the output of the coder (Blocks 7) to provide the time samples of the watermarked mix signal. These samples are finally converted to 16-bits PCM (uniform quantization) at Blocks 8, since audio-CD / wav format application is targeted.

At the decoder, only the (watermarked) mix signal is available. MDCT decomposition is processed (Block 9) the same way as was done at the coder. Then the watermark is extracted from watermarked MDCT coefficients using quantization (Blocks 10) and then decoded (Block 11). Both the index watermark and the descriptors watermark are recovered. The TF coefficients of the I_c encoded source signals are reconstructed at Block 12, and the corresponding time estimates $\hat{s}_1$ to $\hat{s}_{I_c}$ are provided by IMDCT at Block 13. The MDCT coefficients of source signals estimated by SC-ISS are then removed from the mixture at Block 14. The combination of source indexes is used to locally invert the resulting sub-mixture (Block 15). A time-domain synthesis by IMDCT is finally carried out at Blocks 16 to reconstruct the estimated source signals $\hat{s}_{I_c+1}$ to $\hat{s}_I$ from the separated TF coefficients.

DETAILED DESCRIPTION OF THE PROPOSED ISS SYSTEM

In this section, we describe in details the functional blocks of the proposed hybrid SC-I-ISS system. When the role of a block is similar at the coder and at the decoder, it is only described once for concision. The articulation between blocks has been given in the previous section.

A sparse TF decomposition using MDCT

The target source signals in ISS are voice/instrument signals playing a same piece of music (but recorded separately for the sake of the proposed *informed* technique). Using a time-frequency (TF) representation of audio signals has been shown to take into consideration both spectral and temporal variability of the signals, as well as to exhibit their natural sparsity, i.e. much lower overlapping of signals in the TF domain than

in the time domain [11] [21] [2] [16]. As in [16] [14], the Modified Discrete Cosine Transform (MDCT) [18] is used as the TF decomposition since it presents several properties very suitable for the present problem: good energy concentration (hence emphasizing audio signals sparsity), very good robustness to quantization (hence robustness to quantization-based watermarking), orthogonality and perfect reconstruction.

The MDCT is applied at Blocks 2, 2' and 9 on signal time frames of W=2048 samples (46.5ms for a sampling frequency $f_s = 44.1$kHz), with a 50%-overlap between consecutive frames. The frame length W is chosen to follow the dynamics of music signals while providing a frequency resolution suitable for the separation, in accordance with the results established in [20] [12]. The time-domain signals are recovered from processed MDCT matrices at Blocks 7, 13 and 16 by frame-wise inverse transformation followed by overlap-add. Appropriate windowing is applied at both analysis and synthesis to ensure the "perfect reconstruction" property [18].

Since the MDCT is a linear transform, the initial LIS source separation problem remains LIS in the transformed domain for each TF bin, i.e. (1) can be rewritten:

$$\mathbf{X}(f,t) = \mathbf{A} \cdot \mathbf{S}(f,t) \tag{2}$$

where $\mathbf{X}(f,t) = [X_1(f,t), \cdots, X_J(f,t)]^{\mathrm{T}}$ and $\mathbf{S}(f,t) = [S_1(f,t), \cdots, S_I(f,t)]^{\mathrm{T}}$ denote the mixture and source vectors of MDCT coefficients located at frequency bin f and time bin t.

Source-Coding ISS

The SC-ISS technique is based on coding the source signals by a set of descriptors, denoted $d_i(p,q)$, aimed at characterizing the TF structure of sources (Block 3). Because of the limitation of the watermarking capacity, descriptors have to be extracted and then encoded at a molecular level, i.e. for groups of neighbouring MDCT coefficients [16] (for instance, a molecule is a 1×4 matrix of MDCT coefficients). We adopt the mean-gain-shape vector quantization technique [8], since it has proven its ability to optimize the trade-off between coding quality and bit-rate: For each molecule $M_{pq}^{s_i}$ of source signal $s_i, i \in [1, \ldots, I_c]$, its mean $\mu_{pq}^{s_i}$ and its gain (i.e. standard deviation) $g_{pq}^{s_i}$ are encoded using scalar quantizers, and its shape $\phi_{pq}^{s_i}$, defined as the normalized set of amplitudes of the molecule MDCT coefficients, is encoded using vector quantization[3]

For each type of instrument, and for each descriptor, codebooks of prototypes are designed at each frequency bin using the Linde-Buzo-Gray algorithm [10] applied on a large training database of signal molecules. The mean/gain/shape descriptor of a source molecule is the closest mean/gain/shape prototype in the corresponding codebook, according to the Euclidean distance. The codebooks are assumed to be known at the decoder, and the index of the prototype in the codebook is embedded as mean/gain/shape information (Block 6). Note that the resolutions of descriptor quantizers are determined using bit allocation tables known at both coder and decoder levels. The allocation process is not detailed in the present paper, because we rather focus on the combination of the SC-ISS approach with I-ISS.

At the decoder, after the extraction and decoding of the watermark at Blocks 10 and 11, the molecules of the source signals $\hat{s}_i, i \in [1, \ldots, I_c]$ are estimated at Block 12 by:

[3] Actually, if the embedding capacity is too low to embed all three descriptors, only the gain is coded and the source modecule is estimated at the decoder by weighting the mix molecule using this gain (see [16] for details).

$$\hat{M}_{pq}^{s_i} = \check{g}_{pq}^{s_i} \times \check{N}_{l_q} + \check{\mu}_{pq}^{s_i}. \tag{3}$$

where $\check{N}_{l_q}$ denotes the molecule of the shape codebook (adapted to the "instrument" category of source signal s_i) closest to the normalized version of molecule $M_{pq}^{s_i}$, and $\check{}$ denotes the quantized version of the descriptors.

Index-based ISS

In this subsection, we first briefly describe the principle of I-ISS for a general mixture such as (2) (see [14] for details). We will see in the next section how it is combined with the SC-ISS system.

In I-ISS, the estimation of source signals is processed by an inversion of the mixture signal in each TF region. In each TF region, only J sources are assumed to be relevant (for instance $J = 2$ for our stereo mixtures), i.e. of significant energy. The possibility to consider J active sources relaxes the restrictive assumption of a single active source at each TF bin, made in [21] [2]. Equation (2) can thus be locally reduced to:

$$\mathbf{X}(f,t) \approx \mathbf{A}_{\mathscr{I}_{ft}} \mathbf{S}_{\mathscr{I}_{ft}}(f,t) \tag{4}$$

where $\mathscr{I}_{ft}$ denotes the set of $I_{ft} = J$ most active source signals at TF bin (f,t), *i.e.* the set of source signals locally predominant within the mixture. $\mathbf{A}_{\mathscr{I}_{ft}}$ represents the $J \times J$ mixing submatrix made with the $\mathbf{A}_i$ columns of $\mathbf{A}$, $i \in \mathscr{I}_{ft}$. We denote $\overline{\mathscr{I}}_{ft}$ the complementary set of non-active (or at least poorly active) sources at TF bin (f,t). The source signals at bin (f,t) are estimated by local inversion of the mixture:

$$\begin{cases} \hat{\mathbf{S}}_{\mathscr{I}_{ft}}(f,t) &= \mathbf{A}^{\dagger}_{\mathscr{I}_{ft}} \mathbf{X}(f,t) \\ \hat{\mathbf{S}}_{\overline{\mathscr{I}}_{ft}}(f,t) &= 0 \end{cases} \tag{5}$$

where $\mathbf{A}^{\dagger}_{\mathscr{I}_{ft}}$ denotes the inverse of $\mathbf{A}_{\mathscr{I}_{ft}}$.

The side-information that is transmitted between I-ISS coder and decoder (in addition to the mix signal) consists of i) the coefficients of the mixing matrix $\mathbf{A}$, and ii) the combination of indexes $\mathscr{I}_{ft}$ that identify the predominant sources in each local region of the TF plan. This contrasts with blind and semi-blind separation methods (e.g. [3]) where those both types of information have to be estimated from the mix signal only, generally in two steps which can both be a very challenging task and source of significant errors. As for the mixing matrix, the number of coefficients to be transmitted is quite low in the present LIS stereo configuration, and the transmission cost of $\mathbf{A}$ is assumed to be neglectible compared to the transmission cost of $\mathscr{I}_{ft}$. In the following, we thus consider that $\mathbf{A}$ is perfectly known at the decoder.

A crucial step of I-ISS relies in the determination of $\mathscr{I}_{ft}$. For this, an Oracle estimator using the knowledge of the source signals, the mixture signal, and the mixing matrix $\mathbf{A}$ is used, similar to the one introduced in [20] for the general purpose of evaluating the performances of source separation algorithms. The principle is straightforward: All possible combinations of separated source coefficients (5) are calculated, and the combination that provides the lower mean squared error between original and estimated source coefficients (averaged across all sources) is selected. The side-information transmitted via watermark is the code corresponding to $\mathscr{I}_{ft}$ for each TF bin. The size of this code only depends on the number of sources to es-

timate by inversion (the number of possible combinations $\mathscr{I}_{ft}$ for I sources is $\binom{I}{J}$).

Combined SC-ISS/I-ISS process

In the present system, the I-ISS process only concerns the subset of $I - I_c$ source signals s_{I_c+1} to s_I. The index $\mathscr{I}_{ft}$ of the predominant source that drives the separation process is thus determined from this subset of source signals, the corresponding "submix" matrix, and resulting "submixture" signal. Therefore, the estimation of source coefficients $[\hat{S}_{I_c+1}, \ldots, \hat{S}_I]$ is processed at Block 15 of the decoder using I-ISS after that sources $[\hat{S}_1, \ldots, \hat{S}_{I_c}]$ estimated by SC-ISS have been removed from the mixture signal at Block 14 (using the mixing matrix $\mathbf{A}$; we remind that $\mathbf{A}$ is supposed to be transmitted at the decoder). In other words, the I-ISS process of (5) is applied at Block 15 using $\dot{\mathbf{X}}(f,t) = \mathbf{X}^W(f,t) - \mathbf{A}_{SC}\ \hat{\mathbf{S}}_{SC}(f,t)$ instead of $\mathbf{X}(f,t)$, where *SC* denotes the subset of sources coded by SC-ISS and corresponding submix matrix. $\mathbf{X}^W(f,t)$ denotes the MDCT coefficients of the mixture after watermarking and time-domain PCM conversion.

To illustrate the overall SC-I-ISS process, let us consider an example of a 5-source mixture signal, *i.e.* $\mathbf{S}(f,t) = [S_1(f,t), S_2(f,t), S_3(f,t), S_4(f,t), S_5(f,t)]^{\mathrm{T}}$, If s_1 and s_2 are encoded using SC-ISS , and that $\mathscr{I}_{ft} = \{3,5\}$, *i.e.* if s_3 and s_5 are the predominant sources at TF bin (f,t), then $\mathbf{A}_{\mathscr{I}_{ft}} = [\mathbf{A}_3, \mathbf{A}_5]$, $\mathbf{S}_{\mathscr{I}_{ft}}(f,t) = [S_3(f,t), S_5(f,t)]^{\mathrm{T}}$, and the non predominant sources among $[s_3, s_4, s_5]$ is reduced to the singleton $\mathbf{S}_{\overline{\mathscr{I}}_{ft}}(f,t) = \{S_4(f,t)\}$.

The inversion process can be affected by a double deterioration: the deterioration $\mathbf{X}(f,t) - \mathbf{X}^W(f,t)$, due to the watermark embedding, but also the deterioration due to the coding noise $\mathbf{S}_{SC}(f,t) - \hat{\mathbf{S}}_{SC}(f,t)$ which affects the mixture when sources estimated by SC-ISS are removed. The influence of the watermark on the inversion results has been studied in [14], and it appeared that for a limited watermark (up to, say, 80kbits/s), the separation performances were not impaired by the use of an altered version of the mixture. Furthermore, if the encoding of sources s_1 to s_{I_c} is accurate enough, the effect of the coding noise onto the (reduced-size) inversion process is also assumed to be neglectible. Obviously, the coding process is expected to improve the overall estimation (by reducing the dimension of the mixture inversion)[4].

The watermarking process

The watermarking technique used at Blocks 6 and 10 of Fig. 1 is derived from the Quantization Index Modulation (QIM) technique of [5], applied to the MDCT coefficients. The embedded message is carried by a quantization of the MDCT coefficients on uniform scalar quantizers. For each TF bin, an overall quantizer is defined, which is the result of the intertwining of a set of 2^R uniform quantizers, each quantizer representing a distinct R-bits binary code. Watermarking a R-bits binary code on a given MDCT coefficient is done by quantizing this coefficient with the corresponding quantizer. More details about the watermarking technique can be found in [17]. While in [14] the quantizers resolution was fixed empirically by listening tests, in the present paper, it is locally determined by the use of a psycho-acoustic model (PAM) so that the watermarking remains inaudible. The inaudibility constraint gives an upper bound to the determination of R while the robustness of the watermark to the 16-bits PCM conversion of Block 8 gives a lower bound. Using a PAM enables to gain significant embedding capacity for the side-information: while in [14] an embedding bitrate of about 150kbits/s/channel was obtained, the PAM inspired by the MPEG-AAC model [1] enables to reach a bitrate up to 250kbits/s/channel. The embedding bitrate depends on the musical content: the richer the signal, the higher the embedding capacity. Hence, the capacity is well suited with the amount of side information needed to describe the content of the signal. Furthermore, an adjustment of the masking curve in the PAM enables to find a trade-off between a large embedding capacity, useful for the SC-ISS approach, and a limited alteration of the mixture, to ensure good estimation results by I-ISS.

Allocation of the embedding capacity

The allocation of the embedding capacity is, in SC-I-ISS, a multi-constraints issue. In SC-ISS, the capacity determines the settings of the descriptors. Once computed, the capacity must be distributed among 1) the sources and 2) the different descriptors for each source. On the contrary, in I-ISS, since the side-information has a fixed size, depending only on the number of sources to be estimated by inversion, the requested capacity is fixed. The choice is made to embed the I-ISS watermark in the high frequencies since 1) the capacity is larger in low frequencies, and thus will be used to encode the SC-ISS watermark, 2) the I-ISS watermark requires only a limited ressources.

In the present paper, the bit allocation is done according to diverse strategies which are not detailed here. Different configurations are presented in the experiments section. Let us just mention here a basic example. If we consider a mixture of 5 source signals, with 2 sources encoded by SC-ISS, a 2-bit code is sufficient to encode the I-ISSwatermark ($\binom{3}{2} = 3$). This corresponds to a bitrate of 32kbits/s/channel for the I-ISS watermark to estimate source signals in the frequency range [0-16kHz]. If the PAM is set to provide an overall capacity of 150kbits/s/channel (as presented in Table 2), a resulting bitrate of 118kbits/s/channel is available to encode each source signal estimated by SC-ISS. Note that, in anticipation of real-time implementation of the decoder, the whole embedding process is made for each time frame independently, to allow a frame-by-frame decoding of the watermark, and subsequent streaming source separation for active listening of music.

EXPERIMENTS

In this section, we present a series of experiments that we conducted to evaluate the performances of the proposed SC-I-ISS system on music signals. After presenting the data, we consider the advantage of a reduced sub-mixture in I-ISS. Finally we provide the results for the combined SC-I-ISS system, and we compare it to performances obtained by separate I-ISS and SC-ISS systems on the same signals.

Data

Tests have been processed with 44.1kHz-music signals, with 5-source singing voice + instruments mixtures. The separation results have been averaged over five 10-seconds excerpts of different musical styles (rock, pop, funk, new-wave and jazz), representing a total amount of 50s of music. Sources are: s_1 = guitar or piano, s_2 = drums (one track for the overall drum set), s_3 = singing voice (from a male or female singer), s_4 = bass guitar, s_5 = horns or choirs or keyboards. Different LIS mixing matrices were used to create the stereo test mixtures. One typical example corresponding to the azimuths vector (in degrees) $\theta = [-30, -10, 0, 10, 30]$ is:

$$\mathbf{A} = \begin{bmatrix} 0.95 & 0.82 & 0.71 & 0.58 & 0.32 \\ 0.32 & 0.58 & 0.71 & 0.82 & 0.95 \end{bmatrix} \qquad (6)$$

[4] A more detailed study of the deterioration of the mixture signal in the present SC-I-ISS framework will be carried out in a future work.

Source signals overlapping

Remember that the main purpose of the SC-I-ISS is to reduce the overlapping of source signals into the MDCT domain to improve the separation performances. In order to assess the relevance of our approach, a measurement of the source signals overlapping, similar to the one presented in [14] , is carried out. At each TF bin, a ratio between the energy of a target source signal and the energy of the other source signals within the mixture is measured, then source signals are ranked in a descending order. The energy distribution of each source signal with respect to this rank is computed at each TF bin, and results are presented in Table 1. Table 1a presents the (average) percentage of energy of each source signal, at TF bins where this source is ranked first, second, third, and so on, within the mixture. It can be noticed that 89.7 to 98.8% of the energy of the source signal are concentrated within TF bins where the sources are among the first two most energetic sources. However, 9.5, 1.2 and 10.3% of the energy of sources s_1, s_3 and s_5 respectively are not reconstructed by I-ISS since it corresponds to TF bins where these sources are not among the two most energetic within the mixture. Table 1b presents overlapping results in the (simulated) SC-I-ISS configuration were s_2 and s_4 are estimated by SC-ISS and removed from the original mixture before s_1, s_3 and s_5 are estimated by I-ISS. The removal of two sources from the initial mixture reduces the percentage of each source s_1, s_3 and s_5 which is not reconstructed by I-ISS. This percentage decreases to 3.4, 0.3 and 3.6% respectively.

Table 1: Average percentage or the overall energy of source signals depending on their rank within a) the original 5-source mixture, b) the sub-mixture composed of the 3 sources s_1, s_3 and s_5, for 5 mixtures of 10 seconds each.

(a) Original mixture.

Rank	s_1	s_2	s_3	s_4	s_5
1	69.1	86.6	92.1	87.6	65.4
2	21.3	10.7	6.7	10.7	24.3
3	7.3	2.0	1.0	1.5	8.5
4	1.9	0.5	0,2	0.2	1.7
5	0.3	0.1	3.10^{-2}	3.10^{-2}	0.2

(b) Sub-mixture of s_1, s_3 and s_5.

Rank	s_1	s_3	s_5
1	80.7	93.5	79.8
2	15.9	6.2	16.7
3	3.4	0.3	3.6

Comparative tests

Different configurations of ISS are tested to evaluate the separation performances of the proposed hybrid SC-I-ISS method. The settings of the different configurations are summarized in Table 2.

The SC-ISS configuration corresponds to the ISS technique only based on the source coding. The coding information is distributed across the 2 channels of the mixture. The average embedding bitrate of the watermark is 290kbits/s/channel (capacity is similar on the two channels) and the size of the molecules is 1×4 [16]. Five source signals are to be encoded into the two channels of the mixture. The two sources presenting the larger spectral variability are encoded into one channel with a bitrate of about 290/2 = 145kbits/s/source, while the three remaining sources are encoded into the second channel with a bitrate of $290/3 \approx 97$kbits/s/source. In the following experiments, sources s_1 and s_2 are encoded into one channel, while sources s_3 ,s_4 and s_5 are encoded into the second channel.

Table 2: Configuration of tested algorithms

Algorithm	I-ISS code (bits)	I-ISS bitrate (kb/s)	SC-ISS bitrate (kb/s)	Total waterm. (kb/s)
SC-ISS	-	-	290	290
I-ISS	4	64	-	64
SC-I-ISS	2	32	118	150
SC-I-ISS opt.	2	32	118	150

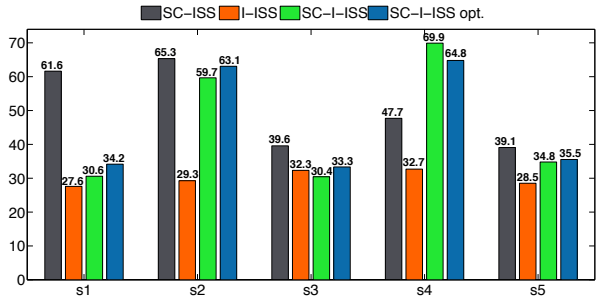
The I-ISS configuration corresponds to the ISS technique only based on local mixture inversion. The process is carried out at a single TF bin level. Since the number of combinations of $J = 2$ sources among the 5 initial source signals is $\binom{5}{2} = 10$, a 4 bits code per TF bin is sufficient to represent the set $\mathscr{I}_{ft}$ of predominant sources, which corresponds to an embedding bitrate of 64kbits/s/channel [14] .

Two configurations of the new hybrid SC-I-ISS method are also tested. For both of them, SC-ISS concerns 2 out of the 5 initial sources and is carried out with a 1×4 molecule while the I-ISS of the 3 remaining sources is processed at a single TF bin level. The bitrate of the sum of the I-ISS watermark and the SC-ISS watermark is set for the two configurations at 150kbits/s/channel. This bitrate is fixed empirically as a trade-off between the need of a large capacity to encode sources by SC-ISS, and the constraint of a small deterioration of the mixture imposed by the inversion process of the I-ISS [13]. Since a 2-bit code is sufficient for the encoding of all the possible combinations of $J = 2$ active sources among the 3 sources separated by I-ISS, the bitrate for I-ISS is 32kbits/s/channel, and thus the remaining 118kbits/s are used on each channel to encode one source signal. The two SC-I-ISS configurations are distinguished by the allocation of the coding ressource between the sources to encode by SC-ISS. In the SC-I-ISS configuration, the same coding ressource is allocated to the different encoded sources, while in the SC-I-ISS opt. configuration, the allocation is determined by the spectral content of the source signal. For instance, in the following experiments sources s_2 (drums) and s_4 (bass guitar) are encoded by SC-ISS. The TF spectrum of the bass, concentrated in the low frequencies ([0-4kHz]) is much sparser than the TF spectrum of the drums spread in all the frequency bandwidth [0-16kHz]. In order to encode the descriptors of the drums in all the bandwidth [0-16kHz], a larger embedding capacity is allocated to the drums: an average bitrate of 150kbits/s is allocated to s_2 while 85kbits/s are allocated to s_4. As a consequence, a significant improvement is expected by this "optimal" allocation of the coding ressources, especially if sources to encode present a large spectral diversity.

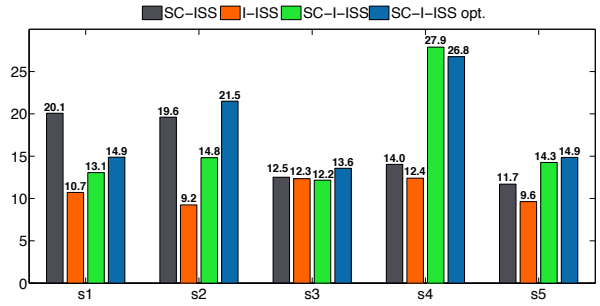
Separation results

The quality of separated sources has been assessed by both informal listening tests with high-quality headphones, and performance measures (log power ratios), as defined in [19]. The source-to-interferences ratio (SIR) measures the level of interferences from the other sources in a source estimate, the source-to-artefacts ratio (SAR) measures the level of artefacts in a source estimate, and the source-to-distortion ratio (SDR) provides an overall separation performance criterion (that gathers the influence of interfering sources and artefacts). Because

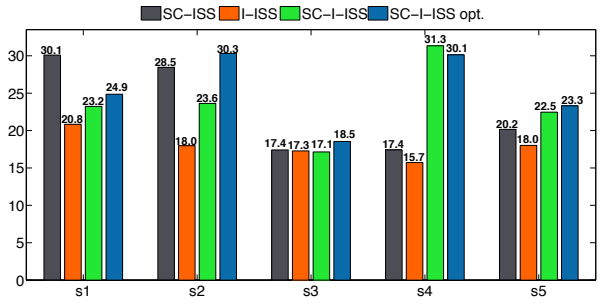
of high output SIRs, the measured SDRs and SARs are almost identical (pair-wise) for all settings [19]. Therefore, we only provide SDR measures. Furthermore, since all musical sources do not contribute to the mixture with the same power in a well musically balanced mix, we also provide the input SIR of each source. The rejection power of the method is revealed by the difference between the (output) SDR and the input SIR which characterizes the difficulty of the task: a source with low input SIR is more difficult to extract than a source with high input SIR. The input SIRs for sources s_1 to s_5 are respectively -9.4, -8.3, -5.3, -3.7 and -7.8dB.



(a) SIR (dB)



(b) SDR (dB)



(c) SDR-SIR_{in} (dB)

Figure 2: Separation results for all 4 settings of Table 2. Average performances over 50 seconds of five 5-source stereo mixtures of different music styles. Sources s_1 to s_5 are guitar/piano, drums, singing voice, bass guitar, horns/choirs/keyboards. For the two SC-I-ISS configurations, sources s_2 and s_4 are estimated by SC-ISS, while sources s_1, s_3 and s_5 are estimated by I-ISS.

Separation results averaged over five 5-source mixtures of 10 seconds each are presented in Fig.2. A first observation is that high separation performances are obtained for all ISS techniques in term of competing sources rejection, as demonstrated by high output SIR values. SIRs between 27.6 and 69.9dB show a very good rejection of the interferences for all sources. The source signals are clearly isolated, as confirmed by listening tests. High SDR values show a (very) good individual signal reconstruction, and a (very) good overall quality which confirm the previous results obtained in [16] [14] . In addition to a very good isolation and (very) limited artefacts for each source, the quality of the isolated source signals makes them usable to clearly enhance or on the contrary turn down a source in the mixture (by simple time-domain or MDCT-domain addition or subtraction), possibly until complete suppression. Although this should be confirmed by dedicated formal listening tests, when remixing a given estimated source within the mix signal, the artefacts coming from this estimated source appear to be efficiently masked by the other sources. This clearly opens the way for generalized remix/karaoke "real-world" applications. Sound samples for the different configurations of Table 2 can be downloaded at `http://www.gipsa-lab.inpg.fr/~mathieu.parvaix/SC-I-ISS-demo.rar`. The package includes original and watermarked mixtures, and original and separated source signals. All signals are correctly scaled and mixing matrix values are given in an accompanying file so that the interested reader can directly process its own remix using the mixture signal and separated sources.

Let us now consider in more details the performances for each of the four configurations tested. SDRs ranging from 11.7 to 20.1dB confirm the efficiency of the SC-ISS approach in terms of individual signal reconstruction, especially given the low input SIR of each source signal. The difference of coding performances between sources s_1/s_2 and sources $s_3/s_4/s_5$ can be explained by the ressource allocation (sources s_1 and s_2 were allocated about 1.5 times more ressource s_3, s_4 and s_5). Consequently s_1 and s_2, are encoded with more precision than sources s_3, s_4 and s_5. Furthermore, when the set of descriptors mean-gain-shape is used in SC-ISS, the mixture is not used in the source estimation process (but instead molecule prototypes from shape codebooks), which results in a high rejection of competing sources and a better reconstruction (with large codebooks ensuring good coding performances) as shown by SIRs over 60dB for s_1 and s_2 while SIRs are ranging from 39.1 to 47.7dB for s_3 to s_5 and SDRs over 19.5dB for s_1 and s_2 while SDRs are ranging from 11,7 to 14,0dB for s_3 to s_5. Note that SIR_{in}s slightly lower for sources s_1 and s_2 widen the gap in the global measure SDR-SIR_{in} between s_1/s_2 and the other sources.

The performances provided by the configuration I-ISS are lower than the performances obtained by SC-ISS (or the hybrid SC-I-ISS method, see later). This can mostly be explained by the overlapping of source signals within the TF plane, even when assuming J sources active at each TF bin. However, the reader should be aware that the separation performances for the I-ISS configuration are similar to the Oracle performances presented in [19], which correspond to the best performances achievable by local inversion of the mixture when J sources are supposed simultaneously active, see [13] for more details on this point. These good performances validate the assumption of 2 simultaneously predominant sources which enables to estimate the large majority of the energy of each source (cf Fig. 1b). Since no distinction in the ressource allocation between sources is done in I-ISS, separation performances are quite balanced across the different sources.

The interest of the hybrid SC-I-ISS method clearly appears on Fig. 2. The two sources s_2 and s_4 encoded by the SC-ISS system of SC-I-ISS show very good SDRs and SIRs, for both SC-I-ISS configurations. The sub-mixture to separate being composed of 3 sources instead of 5, the separation performances of s_1, s_3 and s_5 by the I-ISS part of the hybrid system are significally increased (compared to the single I-ISS system), in compliance with the energy distribution presented Fig. 1b. While s_4

was encoded with the lowest capacity in SC-ISS (≈97kbits/s), in the SC-I-ISS configuration, it is allocated 118kbits/s, which results in a large SDR increase of about 14dB, and an increase of about 20dB for the SIR[5]. On the contrary, for s_2, the ressource allocation between SC-ISS and SC-I-ISS configurations decreased from 145kbits/s to 118kbits/s, which results in a 4.8dB drop of the SDR performances between those two configurations. Finally, the interest of an allocation of the capacity adapted to the spectral content of a source signal appears when comparing the configurations SC-I-ISS and SC-I-ISS opt. The larger capacity allocated to s_2 (from 118 to 150kbits/s) results in an increase of 6.7dB for SDR, and a 3.4dB for SIR, while the results for s_4 are just slightly decreased (resulting from a drop of the coding capacity from 118 to 85kbits/s). Although not truly "optimal", the second configuration demonstrates the importance of adequately allocating the coding ressource among the sources concerned by the SC-ISS process. This point will be further investigated in future works.

CONCLUSION

The hybrid system of Informed Source Separation described in this article is based on both the sparsity of source signals in the TF domain and the coding of source signals by an appropriate set of TF descriptors. The quite simple separation process of I-ISS is here enhanced by the use of a coding approach already validated in SC-ISS. The reduction of the dimension of the original mixing by coding a subset of sources decreases the overlapping between sources. Significantly better separation performances by local inversion of the mixture are consequently obtained. This system appears as a satisfactory tradeoff between the coding approach, efficient for a reasonable number of sources, but computationally demanding, and the sparsity-based approach, computationally very light, but limited by the overlapping of competing sources. A significant gain on performances of both the I-ISS and the SC-ISS approach was obtained. Note that the audio quality obtained with our hybryd ISS method enables a direct use of individual estimated signals for remix/karaoke applications. In this sense, the *informed* source separation approach presented in this paper provides separation results that strongly outperform the results achievable with classical BSS techniques. Future works will deal with going towards more realistic/professional mixtures, involving convolutive filtering (e.g. reverberation), "true stereo" source signals (e.g. 2-channel synthesizers), and a potentially significantly larger number of sources.

ACKNOWLEDGEMENTS

This work is supported by the French National Research Agency (ANR) CONTINT program, as a part of the DReaM project.

REFERENCES

[1] Iso/iec 13818-7: Information technology - generic coding of moving pictures and associated audio information - part 7 : Advanced audio coding (aac), 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing*, 87(8):1833–1847, 2007.

[3] P. Bofill and M. Zibulevski. Underdetermined blind source separation using sparse representations. *Signal Processing*, 81(11):2353–2362, 2001.

[4] J.F. Cardoso. Blind signal separation: Statistical principles. *Proc. IEEE*, 9(10):2009–2025, 1998.

[5] B. Chen and G. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inform. Theory*, 47(4):1423–1443, 2001.

[6] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation - Independent Component Analysis and Applications*. Academic Press, 2010.

[7] M. E. Davies. Audio source separation. *Mathematics in Signal Processing*, 2002.

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

[9] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley and Sons, 2001.

[10] Y. Linde, A. Buzo, and R. M. Gray. Algorithm for vector quantizer design. *Trans. IEEE Commun.*, 28(1):84–95, 1980.

[11] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim, and B. Boashash. Separating more sources than sensors using time-frequency distributions. *J. Applied Signal Process.*, 2005(17):2828–2847, 2005.

[12] A. Nesbit and M. Plumbley. Oracle estimation of adaptive cosine packet transforms for underdetermined audio source separation. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, Nevada, 2008.

[13] M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Trans. Audio, Speech, and Language Process.*, 2010. submitted.

[14] M. Parvaix and L. Girin. Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, Texas, 2010.

[15] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for single-channel audio source separation. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 101–104, Taipei, Taiwan, 2009.

[16] M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. *IEEE Trans. Audio, Speech, and Language Process.*, 2010. to be published.

[17] J. Pinel, L. Girin, C. Baras, and M.Parvaix. A high-capacity watermarking technique for audio signals based on mdct-domain quantization. In *Int. Congress on Acoustics (ICA)*, Sydney, Australia, 2010.

[18] J.P. Princen and A.B. Bradley. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Trans. Acoust., Speech, Signal Process.*, 64 (5):1153–1161, 1986.

[19] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. Speech Audio Process.*, 14(4):1462–1469, 2005.

[20] E. Vincent, R. Gribonval, and M.D. Plumbleys. Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(2007):1933–1950, 2007.

[21] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process.*, 52(7):1830–1847, 2004.

[22] M. Zibulevsky, B. A. Pearlmutter, P. Bofill, and P. Kisilev. Blind source separation by sparse decomposition in a signal dictionary. *Independent Component Analysis : Principles and Practice*, pages 181–208, 2001.

[5] The better coding resolution of descriptors and a regular use of the mean-gain-shape set of descriptors explain these results