Informed Source Separation of Linear Instantaneous Under-Determined Audio Mixtures by Source Index Embedding

Mathieu Parvaix, *Student Member, IEEE*, and Laurent Girin

***Abstract*—In this paper, we address the issue of underdetermined source separation of I nonstationary audio sources from a J-channel linear instantaneous mixture ($J < I$). This problem is addressed with a specific coder-decoder configuration. At the coder, source signals are assumed to be available before the mixing is processed. A time–frequency (TF) joint analysis of each source signal and mixture signal enables to select the subset of sources (among I) leading to the best separation results in each TF region. A corresponding source(s) index code is imperceptibly embedded into the mix signal using a watermarking technique. At the decoder, where the original source signals are unknown, the extraction of the watermark enables to invert the mixture in each TF region to recover the source signals. With such an *informed* approach, it is shown that five instruments and singing voice signals can be efficiently separated from two-channel stereo mixtures, with a quality that significantly overcomes the quality obtained by a semi-blind reference method and enables separate manipulation of the source signals during stereo music restitution (i.e., remixing).**

***Index Terms*—Audio processing, remixing, under-determined source separation, watermarking.**

I. 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 mixture $\mathbf{x} = [x_1, \ldots, x_J]^{\mathrm{T}}$ ($[.]^{\mathrm{T}}$ 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 BSS and related field of ICA developed for (over)determined mixtures can be found in [1]–[3]. 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 music processing since most music 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.

No BSS/ICA algorithm is truly blind, in the sense that a minimal number of assumptions (generally involving some form of prior knowledge) on the sources and/or on the mixture process must be integrated in the algorithms to derive solutions to the separation problem [4].[1] In the underdetermined case, 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 (much more than the time domain where source signals generally strongly overlap) [5], [6]. As a consequence, many USS techniques are based on sparse time–frequency (TF) representations of signals. For example, in [7] the authors make the assumption that the nonstationary 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,

Manuscript received May 21, 2010; revised September 07, 2010; accepted November 15, 2010. Date of publication December 06, 2010; date of current version June 01, 2011. This work was supported by the French National Research Agency (ANR) as a part of the DReaM project (ANR CONTINT program–09 CORD 006). The associate editor coordinating the review of this manuscript and approving it for publication was Mr. James Johnston.

M. Parvaix was with the Grenoble Laboratory of Image, Speech, Signal, and Automation (GIPSA-Lab), Grenoble Institute of Technology, 38402 Grenoble Cedex, France. He is now with Audience, Mountain View, CA 94043-2232 USA (e-mail: mathieu.parvaix@gipsa-lab.grenoble-inp.fr).

L. Girin is with the Grenoble Laboratory of Image, Speech, Signal, and Automation (GIPSA-Lab), Grenoble Institute of Technology, 38402 Grenoble Cedex, France (e-mail: laurent.girin@gipsa-lab.grenoble-inp.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2097250

[1] As a major example underlined in [4], the Bayesian approach to BSS requires to model the PDF of the sources with priors.

from which sources waveforms are reconstructed. In [8], 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. An extension to the anechoic case is proposed in [9]. Stereo anechoic mixtures are also studied in [10], where a large number of speech signals are separated using only two observations. In each TF bin, the mixture is here assumed to be composed of a single predominant source. The relative attenuation and delay histograms are used to determine the mixing parameters, before the source signals are estimated by TF binary masking. This approach is extended to more than two microphones in [11] and [12].

Beyond the "minimal" assumptions/prior knowledge on the sources and mixture process exploited in usual blind (or rather "semi-blind") separation methods, it can be very interesting to exploit additional prior information that can be available in a specific target application. This provides a new perimeter for the concept of *informed* source separation (ISS) discussed in [4]. Music processing offers a particularly interesting framework for such informed approach, since separation methods can exploit pitch and note onset/offset information as provided by score or MIDI information [13], [14], or even by melody humming [15].

In [16] and [17], we introduced an extreme configuration of ISS, in the sense that the source signals are assumed to be perfectly known, but the separation does not exploit those source signals directly: we proposed a system 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 available at the mixing level (so-called here the coder). 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 defaults 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.[2]

As for BSS, different approaches exist for such "source-informed" 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 [16], [17], a *single-channel* LIS mixture of (speech/music) source signals was processed. A joint "source(s)-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 [20], we first addressed the problem for underdetermined LIS *stereo 2-channel* mixtures of music signals. The ISS system proposed in [20] 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 clearly built on [20]. Its first objective is to present the I-ISS framework and method in more details. Its second objective is to present a series of improvements and additional material that were not considered in [20]. First, the core of the method, i.e., the source signals selection-indexation and estimation, is refined. In [20], a sub-optimal (*a priori*) source selection criterion based on source signals energy was used. It is now replaced with an optimal (*a posteriori*) criterion, which is directly inspired by the Oracle estimators developed in [21] and [22] for the evaluation of source separation techniques. The improved I-ISS system can thus be seen as a source separation technique performing optimal estimation of source signals (under the LIS and sparse assumptions) using the parameters of the Oracle estimators, encoded and transmitted within the mixture signal. Second, a refined "high-capacity" watermarking technique is used to embed the side-information used for source separation. It is based on the same basic principle as in our previous works (Quantization Index Modulation of TF coefficients), but it has been improved independently of the ISS application by using a psycho-acoustic model. This new version enables higher maximum capacity (up to 250 kbits/s depending on the musical content) and automatic adjustment of this capacity to the need. The watermarking system has been presented in [23] and [24]; thus, it will not be presented in details in the present paper. We rather focus on the exploitation of the adjustable capacity in relation with the side-information coding and the consequences of watermarking on source separation quality. Finally, we provide in this paper extended results obtained with an extended set of music signals and 5-source mixture configurations that were not considered in [20].

This paper is organized as follows. Section II is a general overview of the proposed method. In Section III, a detailed description of the technical implementation is given, focusing on the sources selection at the coder and the separation process at the decoder. The relationship between side-information coding, watermarking and separation performances is discussed. Separation results for music signals are given in Section IV. Finally, some perspectives are presented in Section V.

[2]From some point of view, the spirit is close to the one of the MPEG Spatial Audio Coding (SAC) system [18], [19], but our goal is here to completely separate the source signals (from uncompressed mixture signals), and not only to resynthesize/respatialize the audio scene (from compressed downmix signals) as is the case for MPEG-SAC. As a result, the nature of transmitted side-information, the way it is transmitted, and the way it is exploited (for separation and not spatialization) are completely different from SAC. Note that, so far, the proposed ISS methods are not robust to compression, they are dedicated to audio-CD/wav music signals (see Section III-D).

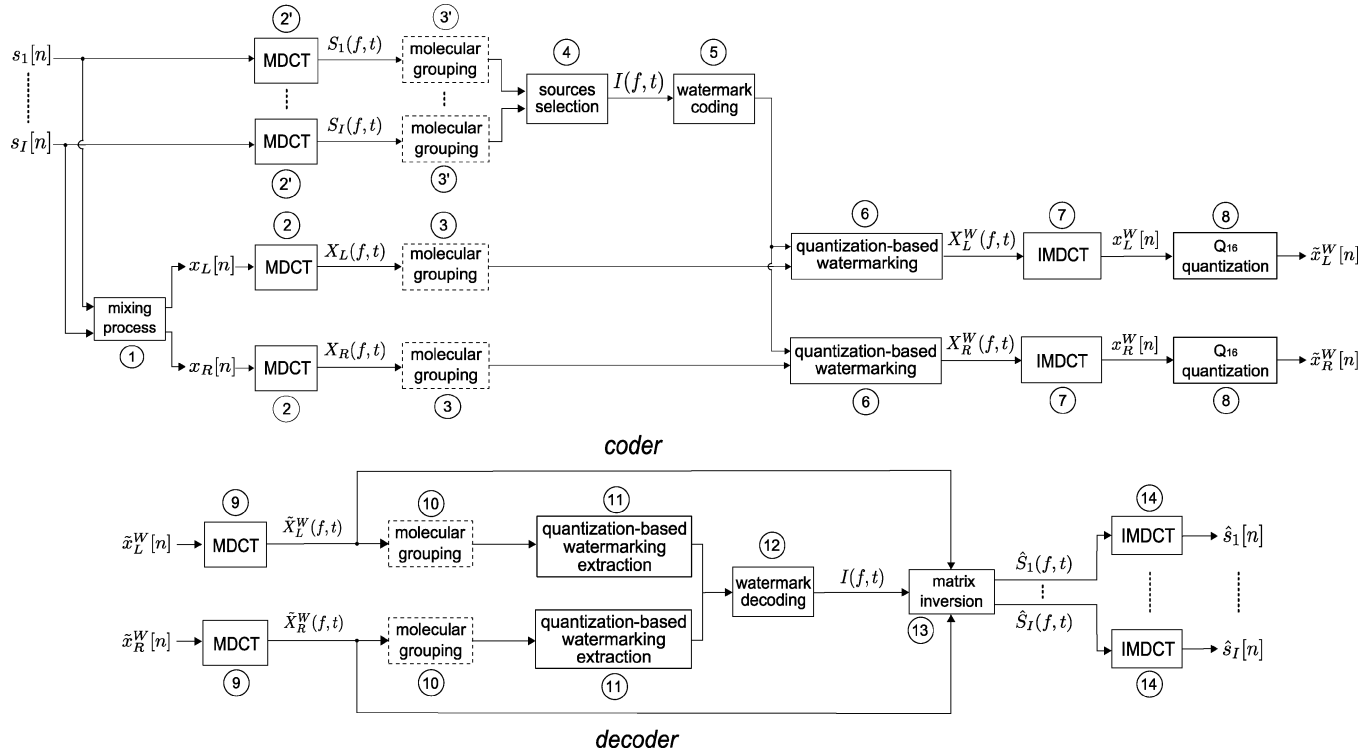

Fig. 1. Detailed structure of both coder and decoder for I-ISS.

II. General Overview of the I-ISS System

Fig. 1 presents the diagram of the proposed stereo Index-based Informed Source Separation (I-ISS) 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], [17] is retained in the present work. However the mixing process at Block 1 of Fig. 1 is here a LIS 2-channel stereo mixture[3] of I nonstationary source signals, as given by (1) for $J = 2$.

Since the I-ISS technique strongly relies on the sparsity of source signals, the overall process is carried out in the TF domain where audio sources are much sparser 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 mixture stereo signal, and at Blocks 2' for the individual source signals (see Section III-A). Within the TF domain, the process can be carried out either for each TF bin, or at a larger scale, referred to as the *molecular* level, depending on the rate of the side-information to be embedded and the settings of the watermarking process (see Section III-E). A molecule is a sub-matrix of a few neighboring TF coefficients. If the process is made at the molecular level, a molecular grouping of MDCT coefficients is required (Blocks 3 and 3'). Since this step is optional, Blocks 3 and 3' are drawn with dotted lines. The core of the method is the analysis carried out at Block 4 of the coder which consists in selecting the most relevant sources in each TF region for further separation by local inversion of the mixture (see Section III-B). The combination of *index* of the selected sources constitutes the side-information to be coded (Block 5) and then embedded (Blocks 6) into the mixture signal by a quantization-based watermarking technique (see Section III-D). The dual operation of Block 2, time-domain signal synthesis by inverse MDCT (IMDCT), is done at the output of the coder (Blocks 7) to provide the time samples of the watermarked mix signal (Section III-A). These samples are finally converted to 16-bit 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 and optional molecular grouping are processed (Blocks 9 and 10) the same way as was done at the coder. Then the watermark is extracted from watermarked MDCT coefficients using quantization (Blocks 11) (Section III-D) and then decoded (Block 12). The resulting combination of source indexes is used to locally invert the mixture (Block 13). This is the core of the I-ISS decoder that will be described in Section III-C. A time-domain synthesis by IMDCT is finally carried out at Blocks 14 to reconstruct the estimated source signals from the separated TF coefficients.

III. Detailed Description of the I-ISS System

In this section, we describe in details the functional blocks of the proposed 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.

[3]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.

A. Time–Frequency Decomposition Using MDCT

The source signals of interest are voice/instrument signals playing a same piece of music (but recorded separately for the sake of the proposed *informed* technique). They are nonstationary, with possibly large temporal and spectral variability, and they generally strongly overlap in the time domain. Using a time–frequency (TF) representation of audio signals has been shown to exhibit natural sparsity, i.e., much lower overlapping of signals in the TF domain, thus leading to sparsity-based separation methods [7]–[17]. As in [16] and [17], MDCT [25] 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 (property exploited in the selection process of Section III-B).

The MDCT is applied at Blocks 2, 2', and 9 on signal time frames of $W = 2048$ samples (46.5 ms for a sampling frequency $f_s = 44.1$ kHz), with a 50%-overlap between consecutive frames. This results in matrices of MDCT coefficients of dimension 1024 frequency bins (denoted by f) by $L/1024$ time bins (denoted by t; L is the total length of each signal). 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 [21] and [22]. The time-domain signals are recovered from processed MDCT matrices at Blocks 7 and 14 by frame-wise inverse transformation followed by overlap-add. Appropriate windowing is applied at both analysis and synthesis to ensure the "perfect reconstruction" property [25].

Detailed description of the MDCT/IMDCT equations will not be given in the present paper, since it can be found in many papers, e.g., [25], including our previous work [17] for its use in ISS. Let us rather focus on the following key point of interest: Since the MDCT is a linear transform, the LIS source separation problem remains LIS in the transformed domain for each TF bin, i.e., (1) can be rewritten as

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

where $\mathbf{X}(f,t) = [X_1(f,t),\ldots,X_J(f,t)]^{\mathrm{T}}$ and $\mathbf{S}(f,t) = [S_1(f,t),\ldots,S_I(f,t)]^{\mathrm{T}}$ denote the mixture and source vectors of MDCT coefficients located at frequency bin f and time bin t. Therefore, the index-based separation process can be carried out in the MDCT domain, as well as the watermarking process.

B. Local Inversion of the Mixture and Sources Selection

In I-ISS as in the semi-blind method of [8], the estimation of source signals is processed by a local inversion of the mixture signal. "Local" means that the process is considered for each TF region, and at this level, only at most J sources are assumed to be relevant, i.e., of significant energy (see below). Therefore, the mixture is locally given by

$$\mathbf{X}(f,t) \approx \mathbf{A}_{\mathcal{I}_{ft}} \mathbf{S}_{\mathcal{I}_{ft}}(f,t) \tag{3}$$

where $\mathcal{I}_{ft}$ denotes the set of $I_{ft} = J$ most active sources at TF bin (f,t), i.e., the set of source signals locally predominant within the mixture. $\mathbf{A}_{\mathcal{I}_{ft}}$ represents the $J \times J$ mixing sub-matrix made with the $\mathbf{A}_i$ columns of $\mathbf{A}$, $i \in \mathcal{I}_{ft}$. If $\overline{\mathcal{I}}_{ft}$ denotes 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[4]

$$\begin{cases} \hat{\mathbf{S}}_{\mathcal{I}_{ft}}(f,t) = \mathbf{A}^{\dagger}_{\mathcal{I}_{ft}} \mathbf{X}(f,t) \\ \hat{\mathbf{S}}_{\overline{\mathcal{I}}_{ft}}(f,t) = 0 \end{cases} \tag{4}$$

where $\mathbf{A}^{\dagger}_{\mathcal{I}_{ft}}$ denotes the inverse of $\mathbf{A}_{\mathcal{I}_{ft}}$. Note that such a separation technique enables to jointly exploit all J mixture channels, and to relax the restrictive assumption of a single active source at each TF bin, as made in [10]–[12].

The side-information that is transmitted between ISS coder and decoder (in addition to the mix signal) mainly consists of 1) the coefficients of the mixing matrix $\mathbf{A}$, and 2) the combination of indexes $\mathcal{I}_{ft}$ that characterizes the "identity" of predominant sources in each local region of the TF plan. This contrasts with blind and semi-blind separation methods 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 (for five source signals we have ten fixed coefficients for each piece of music; if $\mathbf{A}$ is made of normalized column vectors depending on source azimuths, then we have only five coefficients). Therefore, the transmission cost of $\mathbf{A}$ is negligible compared to the transmission cost of $\mathcal{I}_{ft}$, and in the following we consider for simplicity that $\mathbf{A}$ is perfectly known at the ISS decoder.

As for the source indexes, in the specific ISS framework, $\mathcal{I}_{ft}$ is estimated using the source signals: this is done at Block 4 of Fig. 1. The key point is here to define a criterion and the associated optimization process to determine which combination of selected source signals leads to the best global estimation of all source signals using (4). In [20], we considered a raw *a priori* criterion that simply selected the (at most) J most energetic source signals, i.e., the source signals with higher MDCT absolute values (for each TF region). This former criterion was sub-optimal since it did not exploit the mixture signal and the knowledge of $\mathbf{A}$. In the present work, the chosen criterion is an *a posteriori* criterion using the knowledge of the source signals, the mixture signal, and the mixture matrix $\mathbf{A}$ (at the coder) to optimize the estimation of source signals (further carried out at the decoder). This problem is actually similar to the determination of Oracle estimators, as introduced in [21] for the general purpose of evaluating the performances of source separation algorithms, especially in the case of underdetermined mixtures and sparse separation techniques. Depending on the type of separation algorithm, appropriate Oracle estimators provide an upper bound for separation performances, computed by using the available target source signals. In the TF/sparse framework,

[4]An example with five source signals may enlighten the previous notations: $\mathbf{S}(f,t) = [S_1(f,t), S_2(f,t), S_3(f,t), S_4(f,t), S_5(f,t)]^{\mathrm{T}}$. If $\mathcal{I}_{ft} = \{1,3\}$, i.e., if s_1 and s_3 are the predominant sources at TF bin (f,t), then $\mathbf{A}_{\mathcal{I}_{ft}} = [\mathbf{A}_1, \mathbf{A}_3]$, $\mathbf{S}_{\mathcal{I}_{ft}}(f,t) = [S_1(f,t), S_3(f,t)]^{\mathrm{T}}$, and $\mathbf{S}_{\overline{\mathcal{I}}_{ft}}(f,t) = [S_2(f,t), S_4(f,t), S_5(f,t)]^{\mathrm{T}}$.

the authors of [21] established that, since the MDCT is orthogonal, obtaining the best separation results in the time domain according to the mean squared error (MSE) criterion, i.e., minimizing the total distortion $\sum_n \| \hat{\mathbf{s}}[n] - \mathbf{s}[n] \|^2$ between the original and estimated source vectors, is equivalent in the MDCT domain to optimizing the combination of source signals at each TF bin separately (according to the same MSE criterion). The selection of the optimal source combination is thus processed separately at each TF bin by finding

$$\tilde{\mathcal{I}}_{ft} = \arg \min_{\mathcal{I}_{ft} \in \mathcal{P}} \sum_{i=1}^{I} \left(\hat{S}_i(f,t) - S_i(f,t) \right)^2 \tag{5}$$

where $\mathcal{P}$ represents the set of all possible combinations $\mathcal{I}_{ft}$ and the I estimated source signals $\hat{S}_i(f,t)$ are provided by (4). In the case I is limited to a small number of sources (typically about 5 for standard western popular music), $\tilde{\mathcal{I}}_{ft}$ can be found by exhaustive search, and coded with a very limited number of bits before being embedded into the mixture signal (see Section III-E). We found out that using this new *a posteriori* optimal criterion led to an average separation gain of about 1 dB w.r.t. the *a priori* sub-optimal criterion used in [20]. This is because the most relevant sources can here be seen as the ones that better "explain" the mixture signal, and if they generally correspond to the most predominant sources in terms of individual energy, this is not always the case because the mixing matrix coefficients weight the energy of the individual sources in the mixture.

The above source selection and local inversion problem has been presented for $I_{ft} = J = 2$. However, in the present study, we have also considered the option to select a number of locally active sources either lower than J (i.e., $I_{ft} = 1$ for 2-channel stereo mixtures) or greater than J, as a complement to the case $I_{ft} = J = 2$. In case $I_{ft} > J$, the inversion is made in the MSE sense, i.e., $\mathbf{A}_{\mathcal{I}_{ft}}^{\dagger}$ in (4) is the Moore–Penrose pseudo-inverse of $\mathbf{A}_{\mathcal{I}_{ft}}$[26]. This case, which was not considered in [20], can be useful when more than J sources have simultaneously significant energy. In such case, the $I_{ft} \times J$ pseudo-inversion of the mixture can provide a lower MSE than $J \times J$ inversion. In case $I_{ft} = 1$, $\mathcal{I}_{ft}$ is reduced to the singleton $\{i_{ft}\}$, $\mathbf{A}_{\mathcal{I}_{ft}}^{\dagger}$ is equal to $\mathbf{A}_{i_{ft}}^{\mathrm{T}} / \| \mathbf{A}_{i_{ft}} \|_2^2$, and the estimate of the source signal $S_{i_{ft}}(f,t)$ is obtained by $\hat{S}_{i_{ft}}(f,t) = \mathbf{A}_{i_{ft}}^{\mathrm{T}} \mathbf{X}(f,t) / \| \mathbf{A}_{i_{ft}} \|_2^2$. This case can be of interest when the inverse mixing matrices for $I_{ft} > 1$ are ill-conditioned and one of the sources has high energy in comparison to others. Anyhow, when different numbers of active sources are allowed, the selected combination is always the one that provides the lowest MSE in (5), i.e., the best local separation results. Note that tests carried out on 5-source western popular music songs provided the average following distribution: $I_{ft} = 2$ for about 60% of all TF bins, $I_{ft} > 2$ for about 35%, and $I_{ft} = 1$ for less than 5%. However, when weighting such distribution with the source signals energy, it appears that the huge majority of signals energy is processed within the $I_{ft} = 2$ configuration (see Section IV-B).[5]

[5]Note that Vincent *et al*. reported in [21] that Oracle local mixing inversion with a free number of active sources $I_{ft} \leq I$ provided a maximum separation improvement of about 1.5 dB compared to the case where $\mathcal{I}_{ft} = 2$, and we confirm those results in Section IV-D3.

C. Separation Process

The separation is processed at the decoder at Block 13 of Fig. 1. It basically consists of applying (4) using the MDCT coefficients $\tilde{\mathbf{X}}^W(f,t)$ of the transmitted mix signal $\tilde{\mathbf{x}}^W$, calculated at Block 9, instead of the coefficients of the original mix signal $\mathbf{X}(f,t)$

$$\begin{cases} \hat{\mathbf{S}}_{\mathcal{I}_{ft}}(f,t) = \mathbf{A}_{\mathcal{I}_{ft}}^{\dagger} \tilde{\mathbf{X}}^W(f,t) \\ \hat{\mathbf{S}}_{\overline{\mathcal{I}}_{ft}}(f,t) = 0. \end{cases} \tag{6}$$

For this, the watermark is previously extracted at Block 11, and decoded at Block 12 (see Sections III-D and III-E) to provide the combination $\mathcal{I}_{ft}$ controlling the inversion/separation process. While for blind and semi-blind separation, the determination of the optimal combination of sources is very challenging and may be corrupted with many errors, the knowledge of source signals at the coder in the I-ISS system enables to select and transmit the optimal combination $\mathcal{I}_{ft}$. The estimation of source signals at the decoder is thus ensured to be optimal, excepted that the transmitted version of the mixture signal is used as the input of the inversion. Yet, the transmitted mix signal has been watermarked (to embed $\mathcal{I}_{ft}$) and its time waveform has been quantized to 16-bit PCM at Block 8. Both watermarking and 16-bit PCM quantization induce a perturbation of the mixture signal MDCT coefficients, and are thus likely to influence the separation performances (in addition to the degradation induced by the sparsity assumption, i.e., the fact that "residual" non-predominant, but non-null, sources may interfere as noise in the local inversion process). The influence of time-domain 16-bit quantization on MDCT values is assumed to be negligible, especially w.r.t. the influence of the watermarking, since the watermarking is itself configured to be robust to the 16-bit quantization (see Section III-D). The influence of the watermarking on the separation performances is discussed in Section III-E and experimentally evaluated in Section IV-D. We will see that the impact is generally very low with appropriate settings of the whole system.

D. Watermarking Process

The watermarking technique used at Blocks 6 and 11 of Fig. 1 is derived from the Quantization Index Modulation (QIM) technique of [27], applied to the MDCT coefficients. A first basic watermarking scheme based on this technique has been used in our previous works [16], [17], [20]. A refined and more efficient version has been recently proposed in [23] and [24]. We use this last version in the present study. Therefore, we focus on the points that concern the specific use of this technique in the present I-ISS framework. Concerning the watermarking technique in itself, we just present in this section the basic principles and we refer the reader to [23] and [24] for technical details.

For each TF bin, a set of 2^R uniform quantizers is defined, which quantization levels are intertwined, and each quantizer represents a R-bit binary code. Watermarking a R-bit binary code on a given MDCT coefficient is done by quantizing this coefficient with the corresponding quantizer (i.e., the quantizer indexed by the code to transmit; see Fig. 2). At the decoder, recovering the code is done by comparing the transmitted MDCT coefficient (potentially corrupted by transmission noise) with

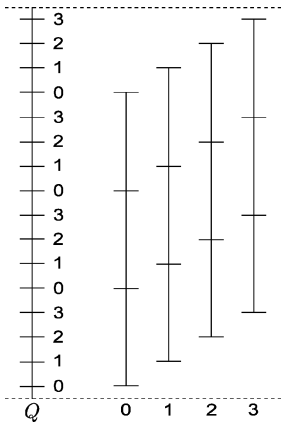

Fig. 2. Example of a set of (here 4) quantizers used for QIM. On the right, the individual quantizers corresponding to four different codes, and on the left, the resulting intertwined quantizer.

the 2^R quantizers (which are assumed to be available at both coder and decoder) and selecting the quantizer with the quantization level closest to the transmitted MDCT coefficient. The complete binary message to transmit (here the set of codes encoding the combinations $\mathcal{I}_{ft}$ for all TF bins) is split and spread across the different MDCT coefficients, so that each coefficient carries a small part of the complete message (see Section III-E). The performance of the watermarking process is determined by two related constraints: it must be robust to the 16-bit PCM conversion of the mixture signal (in other words, the quantization of the original MDCT coefficients at Block 6 of the coder and the quantization of the transmitted MDCT coefficients at Block 11 of the decoder must provide the same result), and it must be inaudible. The first constraint induces a lower bound for the quantization step of the quantizers, since PCM quantization in the time-domain leads to additive Gaussian noise on MDCT coefficients, and given that lower bound, the inaudibility constraint induces an upper bound on the number of quantizers, hence a corresponding upper bound on the individual (MDCT coefficient-wise) capacity R [23], [24].

In [16], [17], and [20], R was determined empirically from listening tests, with a substantial margin that clearly meant sub-optimal choice.[6] Watermarking bit-rates of about 150 kbits/s (depending on the musical content) were obtained. In contrast, in [23] and [24] a psycho-acoustic model (PAM) is introduced in the watermarking scheme. This PAM is calculated for each MDCT frame to control the inaudibility of the R-bit quantization, and therefore leads to an optimal choice for R (for each frame and each frequency bin f) according to a signal-to-mask ratio (SMR) criterion. Because the values of R depend on (f,t), those values must be transmitted to the decoder. For this, a fixed-capacity watermarking "reservoir" is allocated in the higher frequency region of the spectrum. The PAM is inspired from the MPEG-AAC model [28] and adapted to the watermarking problem: the total capacity can be adjusted by shifting the average level of the global masking curve. With this improved version of the watermarking technique, maximum watermarking bit-rates of about 250 kbits/s (depending on the musical content) are obtained. Such rates correspond to the higher level of the masking curve allowed by the PAM; thus, the limit of masking power can be reached. More "comfortable" rates can be set between 150 and 200 kbits/s to guarantee transparent quality for the watermarked signal. This flexibility is used in the present I-ISS system to fit the watermarking capacity to the bit-rate of the $\mathcal{I}_{ft}$ side-information (see Section III-E).

[6]Actually, the watermarking technique was implemented in [16], [17], [20] by gathering all the watermarking quantizers into a single quantizer of resolution R_2. A reference quantizer of resolution R_1 was defined, such that for each TF bin, $R = R_2 - R_1$, and R_1 was fixed to 8 bits for all TF bins; see [17] for details.

E. Coding and Allocation of the Side-Information

In this sub-section, we examine how the side-information $\mathcal{I}_{ft}$ is coded/decoded (Blocks 5 and 12) and how the resulting binary stream is allocated among the different MDCT coefficients in the watermarking process of Block 6. On the way, we propose different possible settings (with different possible consequences on the quality of the source separation). This latter point will be tested experimentally in Section IV-D on the basis of the present discussion.

Let us remind that in the proposed method, the watermark aims at identifying, among I source signals, which ones are selected in each TF region to take part to the local inversion process (see Sections III-B and III-C). In the present study, we consider mixtures of $I = 4$ or $I = 5$ source signals, since it is a reasonable number of simultaneous main musical sources (or coherent groups of musical sources) for many different styles of popular music such as rock, pop, jazz, funk, metal, electro, bossa, fusion, etc. For example, we can consider one singing voice, two or three rhythmic instruments (e.g., piano, bass, drums), and one soloist instrument (e.g., guitar or horn) or choirs, all of them possibly playing at the same time.[7]

The number of possible combinations $\mathcal{I}_{ft}$, $\mathrm{card}(\mathcal{P})$, depends on I_{ft}, the number of sources considered as active in TF bin (f,t). If one single source is assumed to be active, i.e., if $I_{ft} = 1$, $\mathrm{card}(\mathcal{P}) = I$. If two sources are assumed to be active, i.e., if $I_{ft} = J = 2$, $\mathrm{card}(\mathcal{P}) = I(I-1)/2$. If at most two sources are assumed to be active, i.e., if $I_{ft} \leq 2$, $\mathrm{card}(\mathcal{P}) = I(I+1)/2$, and finally, if the number of active sources is let free, i.e., $I_{ft} \leq I$, then $\mathrm{card}(\mathcal{P}) = 2^I$. In the latter case, a fixed-size code of I bits can be used to encode $\mathcal{I}_{ft}$. If $I_{ft} \leq 2$, there are respectively 10 and 15 possible source combinations for $I = 4$ and $I = 5$, respectively, hence a fixed-size 4-bit code is appropriate (although non optimal) for encoding $\mathcal{I}_{ft}$. Therefore, in the following we always consider 4-bit codes in every setting for simplicity, except for the case $I_{ft} \leq I$ with $I = 5$ where a 5-bit code is used (we will see that watermarking at the corresponding bit-rates has poor influence on the separation results). Since the mixture is stereo, half of the side-information can be embedded into each channel. Therefore, if $\mathcal{I}_{ft}$ is provided for each (f,t) bin, the average necessary watermarking capacity for each channel is 2 bits per MDCT coefficient (or 2.5 bits/coefficient when $I_{ft} \leq 5$). In practice, the source separation process can be limited to the [0 16 kHz] bandwidth, because energy of audio signal is generally negligible beyond 16 kHz. Since the MDCT decomposition

[7]More than five instruments can be separated with the present system while keeping $I = 5$ if several instruments that do not play at the same time are on the same audio track, and are thus considered by the algorithm as a single source.

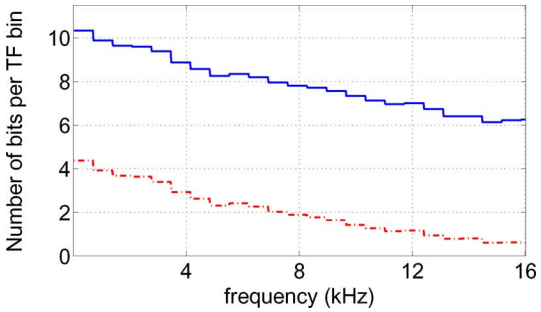

Fig. 3. Example of embedding capacity per TF bin for a given MDCT frame (up to 16 kHz). Continuous line: the PAM is set to enable the maximum embedding capacity under inaudibility constraint; dash-dotted: the PAM is adapted to the required capacity (2 bits/coefficient on the average).

provides as many coefficients as time samples, the side-information bit-rate is $2 \times F_s \times 16,000/(F_s/2) = 64$ kbits/s/channel (or 80 kbits/s/channel when $I_{ft} \leq 5$; $F_s = 44,100$ kHz is the sampling frequency). This is about 1/4 of the maximum capacity of the watermarking process. Therefore, the PAM is automatically tuned so that only the total necessary (fixed) capacity is provided.[8]

In the following, the above settings are referred to as "basic" configuration. Since an important question raised in Section III-C is the influence on separation performance of the watermarking of the side-information in the transmitted mix signal, we define two additional settings for the process.

- A "light-watermark" configuration, where the watermarking bit-rate is intentionally set to half of the value of the basic case, i.e., 32 kbits/s/channel, to limit the influence of the watermarking in the separation process. This is obtained by lowering the level of the masking curve. To compensate for the loss of side-information rate, MDCT coefficients are gathered in 1×2 molecules (optional Block 3) and one single value of $\mathcal{I}_{ft}$ is used for the two consecutive bins (f,t) and $(f,t+1)$ of each molecule. In this configuration, the higher fidelity of transmitted MDCT is balanced by a loss of separation resolution, and we will experiment in Section IV-D3, which parameter is more important for preserving the separation performances.
- An opposite "full-watermark" configuration, where the watermarking bit-rate is intentionally maximized (by setting the masking threshold at its maximum level). Therefore, the bit-rate is here significantly higher than what is needed for the transmission of $\mathcal{I}_{ft}$ (which is anew provided here for each (f,t) bin). The aim is to test if the separation method is robust to high-capacity watermarking, in case users would like to embed additional information[9] for further audio applications, or in future improvements of the present separation application (see Section V). The difference between the "basic" and "full-watermark" configurations is illustrated on Fig. 3 which shows an example of embedding capacity per TF bin of a given MDCT frame for the two configurations.

[8]Or a little more, since in the system of [24], the capacity is defined within subbands, and not for each individual MDCT coefficient.

[9]In the experiments of Section IV, an extra random message (with no interest in the separation process) is added to the useful side-information to fill the watermark channel.

TABLE I
PERCENTAGE OR THE OVERALL ENERGY OF SOURCE SIGNALS DEPENDING ON THEIR RANK WITHIN THE MIXTURE, FOR A 4-SOURCE *JAZZ* MIXTURE AND A 5-SOURCE *POP*-ROCK MIXTURE (40 s OF SIGNAL). (a) JAZZ. (b) POP-ROCK

(a)

Rank	piano	drums	singer	bass
1	89,1	87,3	95,8	84,2
2	9,2	10,0	3,6	13,9
3	1,5	2,2	0,5	1,8
4	0,3	0,4	0,1	0,1

(b)

Rank	guitar	drums	singer	bass	keyboards
1	85,7	71,7	95,6	97,9	37,2
2	12,5	22,4	3,4	1,8	51,9
3	1,6	4,9	0,6	0,3	10,4
4	0,2	0,9	0,1	5.10^{-2}	0,5
5	6.10^{-3}	6.10^{-2}	2.10^{-2}	2.10^{-3}	1.10^{-2}

TABLE II
CONFIGURATION OF TESTED ALGORITHMS

Algorithm	Scale	I_{ft}	Waterm. bitrate (kb/s)
Oracle_O	single TF bin	≤ 2	-
Oracle_M	molecule	≤ 2	-
BZ	single TF bin	2	-
basic I-ISS	single TF bin	≤ 2	64
light-watermark I-ISS	molecule	≤ 2	32
full-watermark I-ISS	single TF bin	≤ 2	250
free I-ISS	single TF bin	$\leq I$	64 ($I = 4$) 80 ($I = 5$)

The characteristics of the three different settings are summarized in Table II. The size of the side-information $\mathcal{I}_{ft}$ code is fixed. However, whatever the PAM adjustment, the available capacity is variable for the different MDCT coefficients of a given frame (see Fig. 3). It is generally larger in the low-frequency region and lower at high frequency. As a consequence, the $\mathcal{I}_{ft}$ codes for the whole frequency bins are concatenated and the resulting bitstream is sliced to fill the non-uniform allocation of the embedding resource. As mentioned before, the bitstream is equally shared by the two channels of the stereo, i.e., half of the total bitstream is allocated to one channel, and the other half is allocated to the other channel, in an arbitrary manner. Similarly, when the 1×2 molecular grouping is used (in the "light-watermark" configuration), half of the per-channel bitstream (i.e., a quarter of the total bitstream) is allocated to the MDCT coefficients at (f,t), and the other half (or other quarter of the total bitstream) is allocated to the neighboring coefficients at $(f,t+1)$, also in an arbitrary manner. Because of this arbitrary aspect of side-information allocation (and because it is a rather trivial task), it will not be presented in further details. Note that Block 12 at the decoder performs inverse concatenation from extracted watermarks, and inverse slicing to recover the $\mathcal{I}_{ft}$ codes for each TF bin.

F. Comparison With ISS Based on Source Coding

In this section, we briefly discuss the differences between the I-ISS configuration presented in this paper, and the SC-ISS (Source-Coding ISS) system of our previous work [16], [17]. Although both approaches are characterized by the assumption of known sources before the mixture, and a coder–decoder structure, there are many fundamental differences. First, the 2-dimension of the mixture signal in I-ISS, as opposed to mono mixture for SC-ISS, enables to estimate the source signals by local inversion while taking advantage of the sparsity of MDCT decomposition. Then, the difference of separation process, and thus the different nature of side-information, directly impacts the embedding resource: the necessary embedding capacity is much lower in I-ISS than in SC-ISS. Indeed, in [16] and [17] a high embedding capacity was expected, to accurately encode the source signals. Consequently, large molecular grouping was necessary (typically 2×4 molecules), and the mixture signal was allowed to be deteriorated (to some extent, and under the constraint of inaudibility): Typically, 8 bits per MDCT coefficient were embedded (i.e., 64 bits per molecule) for accurate source coding. However, the separation process was poorly affected since the mixture signal was not directly used for the source signals estimation. Instead, prototypes issued from codebooks were used. In the present I-ISS method, the modification of the mixture signal induced by the watermark is expected to be significantly lower, because the mixture signal is used in the inversion process. Fortunately, this constraint fits well with the small size of the embedded side-information. As seen in the previous section, this small size enables to reduce molecular grouping to 1×2 molecules, or even to 1×1 molecule, i.e., the whole process is carried out at a single TF bin scale. Therefore, if a tradeoff has to be found in I-ISS between the size of molecules and the deterioration of the mixture induced by the watermark (the bigger a molecule, the lower the deterioration of each TF bin of the mixture, but the higher the risk of source overlapping, and vice-versa), fortunately, this tradeoff is in line with much softer constraints on both the watermarking process and the molecular grouping, as compared to SC-ISS.

IV. Experiments

In this section, we present a series of experiments that we conducted to evaluate the performances of the proposed I-ISS system. We first present the data, then we provide some measure of music signals overlapping/sparsity, and then we provide the results for the source separation itself.

A. Data

Tests have been processed with 44.1-kHz music signals, with 4-source and 5-source singing voice + instruments mixtures. The separation results of Section IV-D have been averaged over five 10-s excerpts of different musical styles (rock, pop, funk, new-wave, and jazz), representing a total amount of 50 s 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 with constant power stereo panning[10] were used to create the stereo test mixtures. One typical example is (for 5-source mixture)

$$\mathbf{A} = \begin{bmatrix} 0.95 & 0.82 & 0.71 & 0.57 & 0.31 \\ 0.31 & 0.57 & 0.71 & 0.82 & 0.95 \end{bmatrix} \tag{7}$$

corresponding to the azimuths (in degrees) $\boldsymbol{\theta} = \{18, 35, 45, 55, 72\}$. For 4-source mixtures, the mixing matrix $\mathbf{A}$ is a sub-matrix formed by the first four columns of (7), and only sources s_1 to s_4 are used. The minimum difference between two azimuths of two different sources is set to 10° in the present experiments, so that the maximum condition number of the 2×2 sub-matrices of A remains limited (to approx. 11.5). Therefore, the impact of the noise due to the watermarking and the sparse assumption on the inversion process (see Section III-C) also remains limited.[11]

B. Source Signals Overlapping

The overlapping of source signals in the TF domain remains a critical issue for sparsity-based source separation techniques, even for informed techniques. In order to assess the relevance of the sparse assumption on music signals, a measurement of source signals overlapping in the TF domain has been carried out, taking into account the energy distribution of the sources. For this, the following energy ratio is calculated for each TF bin of each source signal s_i:

$$R_i(f,t) = \frac{|S_i(f,t)|^2}{\sum\limits_{j \neq i} |S_j(f,t)|^2}. \tag{8}$$

At each TF bin, we then compute the energy distribution of each source with respect to the rank of its power ratio (8), i.e., the percentage of energy for which a given source is ranked first, second, third, and so on. Results are presented in Table I(a) for a 4-source jazz mixture, and in Table I(b) for a 5-source pop–rock mixture.

Those tables obviously show that considering only one active source at each TF bin is a too coarse approximation. Indeed, this would preserve about 96% of the voice energy (in both cases), or about 98% of the bass energy in the pop-rock mixture, but it would also preserve only about 72% of the drums energy and only about 37% of the keyboards energy in the pop-rock mixture (remind that in the source separation process of (6), the energy of sources considered as non-active is set to zero). Therefore, those sources may be severely degraded. Even percentages within the range 85%–90%, as for the bass, drums and piano for the jazz mixture, may not be sufficient to ensure good reconstruction quality. In contrast, even for the 5-source mixture, the assumption of only two active sources can reasonably be made. Indeed, for both examples and for all sources, most of the energy of a source is located at TF bins where this source is within the two most energetic sources of the mixture

[10] i.e., each column is equal to $[\cos(\theta)\ \sin(\theta)]^{\mathrm{T}}$, with the source azimuth θ between 0° (azimuth of the right loudspeaker) and 90° (azimuth of the left loudspeaker).

[11] Let us just mention that we have not observed "unreasonable" estimated values for the MDCT source coefficients in our experiments; a deep investigation of the effects of sub-matrix conditioning on separation performance is out of the scope of the present study.

[according to the ratio (8)]. 89.1% for the keyboards to 99.7% for the bass (in the 5-source mixture) of the overall energy of source signals is concentrated in TF regions where sources are either the most energetic or the second most energetic of all the sources compounding the mixture. This implies that the third (and fourth and, if any, fifth) source is generally of very poor energy compared to the two most predominant sources.[12] Therefore, those remaining sources, not belonging to the two most energetic ones, can reasonably be considered as a noise, and, if the 2×2 inverse matrix $\mathbf{A}_{\mathcal{I}_{ft}}^{-1}$ is not ill-conditioned, the separation process (6) with $I_{ft} = 2$ generally provides (very) good separation results.

C. Quality of Mix Signals

Before we provide separation results, we confirm in this subsection that the watermarking process has no influence on the perceived quality of the mixture signals. This was assessed by extensive informal listening tests, confirming the subjective and objective tests reported in [23] and [24].[13] In fact, for the test signals used in the present study, the watermarking is inaudible in the "full-watermarking" configuration defined in Section III-E, revealing the efficiency of the psycho-acoustic model and associated watermarking process. Therefore, it is guaranteed to be "highly inaudible" for the "basic" and "light-watermark" configurations, since in those cases the masking curve is significantly lowered to fit the required capacity which is much lower than in the "full-watermarking" configuration (see Fig. 3).

D. Separation Results

1) Test Configurations: The different settings presented in Section III-E and summarized in Table II have been tested to evaluate the separation performances of the proposed I-ISS system, including the evaluation of the (cross-)influence of TF resolution (single TF bin or 1×2 molecule) and watermarking bit-rate. In addition, the following three reference configurations were tested. The configuration named Oracle_O refers to the ideal configuration of the oracle estimator, as introduced in [21] (see Section III-B): the optimal combination $\tilde{\mathcal{I}}_{ft}$ is used for separation using (4) instead of (6), i.e., there is no watermarking (hence no separation of the whole process between coder and decoder). As in [21], this configuration is used as an optimal ideal reference that provides the upper bound for the performances of the present (sparse) separation technique. It is thus also used to measure the influence of watermarking in the proposed I-ISS system. The Oracle_M configuration is similar to Oracle_O except that the 1×2 molecular grouping is activated (i.e., one single value of $\tilde{\mathcal{I}}_{ft}$ is used to separate the coefficients of two consecutive TF bins (f, t) and $(f, t+1)$, as for the "light-watermarking" configuration of I-ISS). This configuration is used to measure the influence of molecular grouping alone on the separation process, and also to measure the influence of watermarking in the molecular configuration. Finally, we also implemented the underdetermined blind source separation process of [8], which is also based on local inversion of the mixture, to measure the contribution of the *informed* aspect to the separation performances w.r.t. non-informed separation, in a similar framework of sparse separation techniques.[14] In this reference configuration, further referred to as BZ, the two relevant source signals (out of 4 or 5 here) are selected for each TF bin by finding the linear combination of the two basis vectors that provides the shortest path from the origin to the observed data $\mathbf{x}$. For example, in Fig. 4, the mixture vector $\mathbf{x}$ is assumed to be generated by sources 1 and 2. It can be noticed that such a geometrical method does not provide all the possible source combinations. For instance, if $\mathbf{x}$ is actually a combination of sources 1 and 3, this method will always return the spurious couples of sources (1, 2) or (2, 3). The watermark embedded in the proposed I-ISS method fixes this issue.

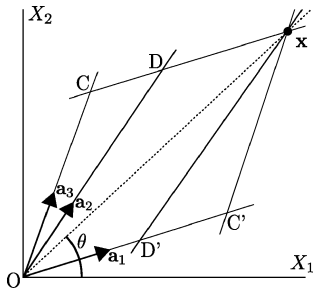

Fig. 4. Geometrical method of the *shortest path* from the origin to the data point $\mathbf{x}$ introduced in [8].

2) Performance Measures: 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 [30]. Basically, 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 (i.e., the level of "self"-degradation in a given source such as musical noise, due to the processing and not to the interfering sources), and the source-to-distortion ratio (SDR) provides an overall separation performance criterion (that gathers the influence of interfering sources and artefacts) [30]. We also provide the input SIR (denoted SIR_in), defined as the (dB) ratio between the power of the considered source and the power of all other sources in the mix to be separated, because all musical sources do not contribute with the same power in a well musically balanced mix (as we tried to generate). Therefore, this input SIR must be taken into account when measuring the rejection power of the method since it characterizes the difficulty of the task: a source with low SIR_in is more difficult to extract than a source with high SIR_in. For the 4-source mixtures, the input SIRs for sources s_1 to s_4, averaged over all tested mixtures, are respectively -8.4, -7.1,

[12] At first sight, this may seem contradictory with the distribution of optimal I_{ft} mentioned in Section III-B, but deeper investigation reveals that the approximate 35% of TF bins where $I_{ft} > 2$ generally contain sources of quite poor energy, hence modestly contributing to the mixture and to the separation process. An additional test will be presented in Section IV-D3 to assess the improvement in separation obtained by considering a free number of active sources (i.e., up to I), in compliance with the results reported in [21].

[13] For example, in [24], objective difference grade (ODG) scores [29] were calculated for a large range of embedding rates and different musical styles. ODG remained very close to zero (hence imperceptibility of the watermark) for rates up to about 260 kbps for musical styles such as pop, rock, jazz, funk, etc. (and "only" up to about 175 kbps for classical music).

[14] The comparison with [8] is here made for the sources estimation step only, i.e., the mixing matrix $\mathbf{A}$ is assumed to be known at the decoder, although in [8] $\mathbf{A}$ is claimed to be accurately estimated by a clustering technique. Therefore, this reference technique is actually a semi-blind technique in the present study.

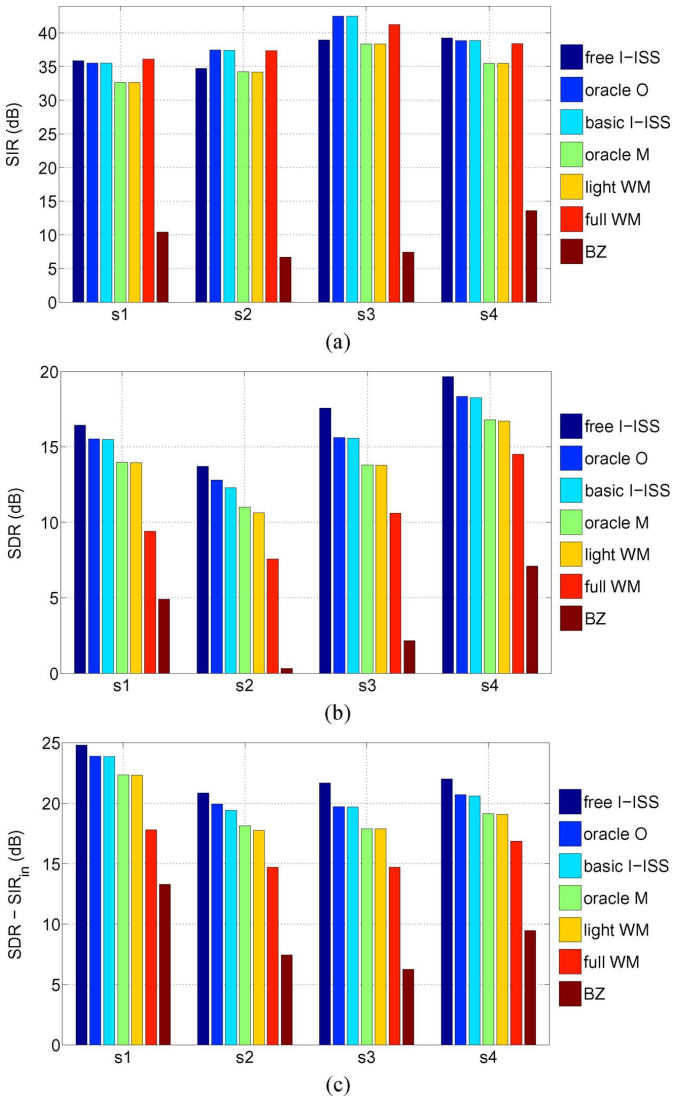

Fig. 5. Separation results of I-ISS for all seven settings of Table II. Average performances over 50 s of five 4-source stereo mixtures of different music styles. Sources s1 to s4 are guitar/piano, drums, singing voice, bass guitar. (a) SIR. (b) SDR. (c) SDR-SIR$_{in}$.

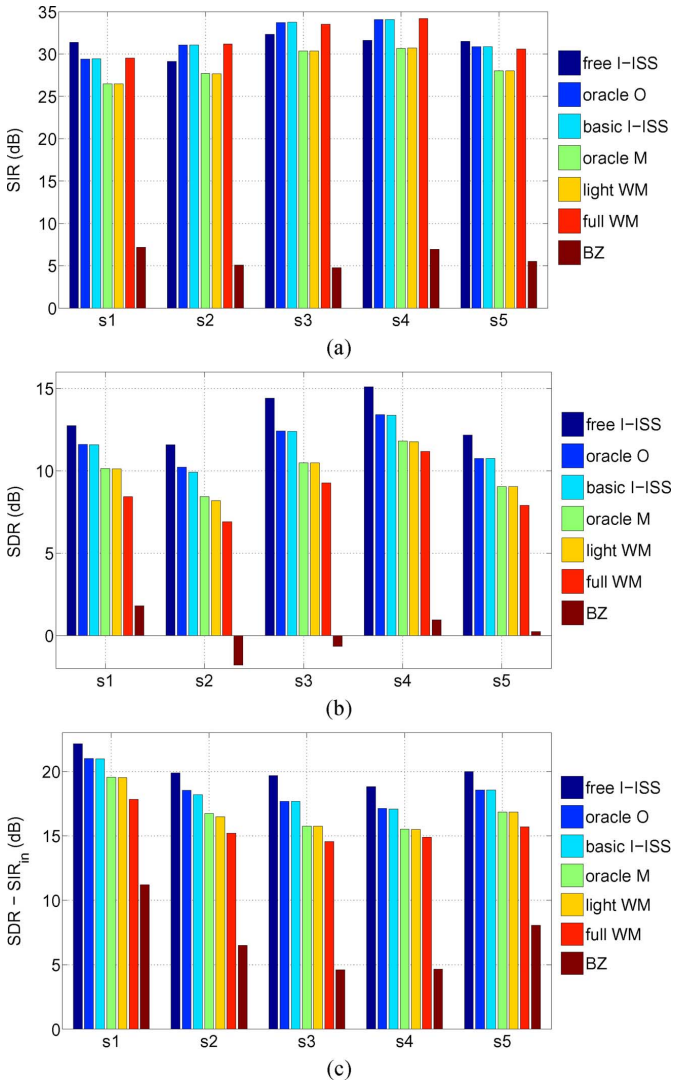

Fig. 6. Separation results for all seven algorithms. Average performances over 50 s of five 5-source stereo mixtures of different music styles. Sources s1 to s5 are guitar/piano, drums, singing voice, bass guitar, horns/choirs/keyboards. (a) SIR. (b) SDR. (c) SDR-SIR$_{in}$.

-4.1, and -2.4 dB. For the 5-source mixtures, the input SIRs for sources s_1 to s_5 are, respectively, -9.4, -8.3, -5.3, -3.7, and -7.8 dB.

3) Separation Performances: Separation results are presented in Fig. 5 for 4-source mixtures and in Fig. 6 for 5-source mixtures. Let us first consider the results of the "basic" I-ISS configuration. A first observation is that, for both 4-source and 5-source mixtures, high separation performances are obtained, in terms of competing sources rejection, as demonstrated by high-output SIR values. SIRs between 35 and 42.5 dB for 4-source mixtures [Fig. 5(a)], and between 29.5 and 34 dB for 5-source mixtures [Fig. 6(a)], show a very good rejection of the interferences for all sources. The source signals are clearly isolated, as confirmed by listening tests (see below). This comforts the validity of the assumption of two predominant sources at each TF bin: since most of the energy of source signals is concentrated in TF bins where the source is within the two predominant sources, the local 2×2 inversion of the mixture enables a very good separation of all source signals.

The results are also very satisfactory in terms of SDR and SAR. Actually, because of high output SIRs, the measured SDRs and SARs are almost identical (pair-wise) for all settings [30]. Therefore, we only provide SDR measures [in Figs. 5(b) and 6(b)], and it can be noted that the overall quality of separated source signals mostly depends on artefacts, i.e., musical noise. SDRs ranging from 12.5 dB to 18 dB for 4-source mixtures, and ranging from 10 to 13.5 dB for 5-source mixtures are obtained.[15] Although those values indicate that the estimated source signals remain noticeably different from the original source signals, they nonetheless confirm the efficiency of the separation in terms of individual source signal reconstruction,

[15]Obviously, separation performances decrease from four-source mixture to 5-source mixture because the presence of an additional source increases the probability of source overlapping. In our experiments, separation performances decrease by about 6.5-dB SIR and 3.5-dB SDR (on the average across all sources), whereas input SIRs only decrease by about 1.2 dB. This may be explained by the fact that, for some of our mixtures, about 10% of the energy of the fifth source signal is located in TF bins where this source is the third most energetic source of the mixture (see Table II(b), thus possibly impairing the local 2×2 mixture inversion.

given the difficulty of such underdetermined mixtures. As for the differences between sources, we remind that input SIRs must be taken into account: for example, SDRs are higher for the bass guitar (s_4) and lower for the drums (s_2), but at the same time the input SIRs are also higher for the bass than for the drums. Therefore, looking at the difference between SDR and SIR_{in} (i.e., a measure of signal enhancement from input to output) in Figs. 5(c) and 6(c) reveal a more balanced picture across sources.[16] Obviously, ratio improvements of about 20 dB and above for 4-source mixtures, and of about 17 dB and above for 5-source mixture confirm the efficiency of the separation.

Altogether, the SIR, SDR, and SDR–SIR_{in} values demonstrate the possibility for individual manipulation of separated signals. This is confirmed by listening tests: as mentioned above, each instrument is clearly isolated, and artefacts are quite limited. Most importantly, 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 (either boosted or subtracted) 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 II can be downloaded at http://www.gipsa-lab.inpg.fr/~mathieu.parvaix/IISS-demo.zip. 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 the watermarking influence. Very interestingly, the "basic" I-ISS system exhibits performances that are almost identical to the Oracle_{O} configuration. Straightly stated, this means that the watermarking at reasonable "basic" bit-rates (i.e., 64 kbits/s/channel here) has negligible influence on the separation process. The watermarked MDCT coefficients are very close to the unwatermarked coefficients; hence, (6) provides results that are almost identical to (4). Those observations are confirmed by the similar separation performances obtained for the Oracle_{M} and "light watermark" configurations (remind that in those cases, the 1×2 molecular grouping is activated, and a 32 kbits/s/channel watermark is embedded into the mixture signal for the "light" configuration). However, when the volume of the watermark strongly increases, as for the "full-watermark" configuration (approx. 250 kbits/s/channel), the effects on the separation performances are significant: an average SDR decrease of about 5 dB (resp. 3 dB) is evidenced for the 4-source mixture (resp. 5-source mixture), as compared to the "basic" configuration. This is because in time–frequency regions where the masking threshold is high, the modification of MDCT coefficients can be high enough to significantly corrupt the inversion process, although remaining inaudible in the mixture signal. Therefore, an end-user of the proposed I-ISS system should be careful when using large watermarking capacity to transmit substantial extra side-information, in addition to the one required for index-based source separation.

As for the influence of the molecular grouping, it is shown by comparing the results of the Oracle_{O} and Oracle_{M} configurations (i.e., without watermarking), and by comparing the results of the basic and light configurations (i.e., with watermarking). In both cases, a decrease of about 3-dB SIR and less than 2-dB SDR can be observed when switching from single TF-bin processing to (1×2) molecular processing (for both 4-source and 5-source mixtures). Therefore, comparing the effects of molecular grouping and watermark size clearly shows that maximizing the resolution of the processing (by working at a single TF bin level) should be preferred to limiting the amount of embedded data (at least from 64 to 32 kbits/s). Indeed, a limited watermark (here 64 kbits/s) does not impair the performances of the inversion/separation process, while gathering TF bins does. The "basic" configuration of the proposed I-ISS system eventually appears to be a very good setting since it conjugates reasonable watermarking rate (with almost no effects) and optimal separation resolution, leading to separation performances almost identical to the optimal oracle configuration.

Let us now briefly see the effect of the parameter I_{ft} on the separation performances. A comparison of separation results between the two basic I-ISS configurations with $I_{ft} = 2$ and $I_{ft} \leq I$ shows an average SDR improvement of 1.3 dB for the 4-source mixture and 1.6 dB for the 5-source mixture, confirming the results of [21]. So the performance gain is noticeable, and it goes together a very reasonable increase of watermarking rate (from 64 to 80 kbits/s/channel for $I = 5$; for $I = 4$, the rate does not change because we use suboptimal 4-bit codes for $I_{ft} \leq 2$). Note that it also goes with an increase of computational complexity, but this issue is out of the scope of the present paper (however, this issue should be considered with attention for real-time implementations if the number of sources increases significantly).

Finally, the comparison of the proposed I-ISS system with the semi-blind method BZ, shows the tremendous gain enabled by the *informed* separation process, for all performance measures, source signals, or mixture size. The SDR gain provided by the transmission of side-information is within the range 10–13.5 dB, and accordingly, the quality of separated signals is much higher for the ISS system. For instance, source signals separated by the BZ method cannot be used for high quality remix/karaoke applications, whereas source signals separated with the I-ISS system clearly can.

V. Conclusion

The Index-based Informed Source Separation system described in this paper is based on the sparsity of source signals in the TF domain and the exploitation of stereophony. This system is characterized by a quite simple separation process and by the fact that the side-information that is embedded to guide the separation process is particularly compact. Therefore,

[16] With a slightly better performances for the first source (guitar/piano). The study of the influence of source signal characteristics on separation performance is beyond the scope of this paper. It may be considered with attention in our future works on ISS.

with appropriate settings, the degradation of the mixture signal by the watermark embedding at the coder has been shown to have negligible effects on the inversion procedure. As a result, the performances of the I-ISS system are comparable to the performances of the optimal Oracle estimator proposed in [21] (for similar LIS/sparse separation configuration). Compared to blind and semi-blind approaches also based on sparsity and local mixture inversion, the informed aspect of separation guarantees optimal combination of sources, leading to a remarkable increase of separation performances. Although it was not much enlighten, another advantage of the I-ISS over blind methods is the knowledge of the mixing matrix $\mathbf{A}$ at the decoder.

The simplicity of the proposed I-ISS system, including the use of a single MDCT transform exploited in both the separation routine and the watermarking routine, has already enabled the realization of a first real-time software implementation of the decoder running on PC/Mac [31]. This software is able to separate 5-source (LIS) stereo mixtures (read from audio-CD or 16-bit PCM wav files) in real-time and it enables the user to remix the piece of music during restitution with basic functions such as volume and spatialization control.

Although it enables basic but efficient left/right spatialization of the sources, the LIS mixture is generally an over-simplistic process when professional/commercial music production is at stake. Moreover, the corresponding sparseness-based separation process has its own limitations: for example, it cannot process two sources located at the same position, since the corresponding submatrix is not invertible. Future works will consider those limitations and deal with going towards more realistic/professional mixtures, involving convolutive filtering (e.g., reverberation) and "true stereo" source signals (e.g., 2-channel synthesizers). A future extension of this work will be the combination of the present 2-channel sparse approach with the source coding ISS approach of [17]. For example, within a mixture of, say, six sources, two of them could be extracted by the coding approach, and the four remaining sources could be estimated by the present sparse method after subtraction of the first two decoded sources to the mixture. A reduction of remaining artifacts is expected. The separation of convolutive and true stereo sources will be considered in such extended framework.

REFERENCES

[1] J. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 9, no. 8, pp. 2009–2025, Oct. 1998.

[2] *Independent Component Analysis*, A. Hyvärinen, J. Karhunen, and E. Oja, Eds.. New York: Wiley, 2001.

[3] *Handbook of Blind Source Separation—Independent Component Analysis and Applications*, P. Comon and C. Jutten, Eds.. New York: Academic, 2010.

[4] K. H. Knuth, "Informed source separation: A Bayesian tutorial," in *Proc. Eur. Signal Conf. (EUSIPCO'05)*, Antalya, Turkey, 2005.

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

[6] N. Mitianoudis and M. E. Davies, "Audio source separation: Solutions and problems," *Int. J. Adapt. Control Signal Process.*, vol. 18, no. 3, pp. 299–314, 2002.

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

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

[9] P. Bofill, "Underdetermined blind separation of delayed sound sources in the frequency domain," *Neurocomputing*, vol. 55, pp. 627–641, 2003.

[10] 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, Jul. 2004.

[11] S. Araki, H. Sawada, and S. Makino, *K-Means Based Underdetermined Blind Speech Separation*, S. Makino, Ed. *et al.* New York: Springer, 2007, Blind Source Separation, pp. 243–270.

[12] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[13] S. Dubnov, "Optimal filtering of an instrument sound in a mixed recording using harmonic model and score alignment," in *Proc. Int. Comput. Music Conf. (ICMC 2004)*, Miami, FL, 2004.

[14] J. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. Int. Conf. Music Inf. Retrieval*, Victoria, BC, Canada, 2006, pp. 314–319.

[15] P. Smaragdis and G. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'09)*, New Paltz, NY, 2009, pp. 69–72.

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

[17] 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, Lang. Process.*, vol. 18, no. 6, pp. 1464–1475, Aug. 2010.

[18] J. Breebaart, J. Herre, C. Faller, J. Rödén, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjörling, and W. Oomen, "MPEG spatial audio coding/MPEG surround: Overview and current status," in *Proc. AES 119th Conv.*, New York, 2005.

[19] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, and F. Myburg, "The reference model architecture for MPEG spatial audio coding," in *Proc. AES 118th Conv.*, Barcelona, Spain, 2005.

[20] 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, TX, 2010, pp. 245–248.

[21] E. Vincent, R. Gribonval, and M. Plumbleys, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Process.*, vol. 87, no. 2007, pp. 1933–1950, 2007.

[22] 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, NV, 2008, pp. 41–44.

[23] J. Pinel, L. Girin, and C. Baras, "A high-rate data hiding technique for uncompressed audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, submitted for publication.

[24] J. Pinel, L. Girin, C. Baras, and M. Parvaix, "A high-capacity watermarking technique for audio signals based on MDCT-domain quantization," in *Proc. Int. Congr. Acoust. (ICA)*, Sydney, Australia, 2010.

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

[26] L. Trefethen and D. Bau, *Numerical Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1997.

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

[28] ISO/IEC 13818–7: Information Technology—Generic Coding of Moving Pictures and Associated Audio Information – Part 7 : Advanced Audio Coding (AAC), 2004.

[29] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, and C. Colomes, "PEAQ – The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1, pp. 3–29, 2000.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech. Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[31] S. Marchand, B. Mansencal, and L. Girin, "Interactive music with active audio CDs," in *Proc. Int. Symp. Comput. Music Modeling Retrieval*, Malaga, Spain, 2010.

Mathieu Parvaix (S'08) was born in Limoges, France, in 1983. He received the M.Sc. degree in signal processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 2008 and the Ph.D. degree in signal processing from Grenoble-INP in 2010. His Ph.D. work was carried out at the Speech and Cognition Department, GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation) and concerned audio/speech processing and underdetermined source separation.

He is now an R&D Staff Member with Audience, Mountain View, CA.

Laurent Girin was born in Moutiers, France, in 1969. He received the M.Sc. and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 1994 and 1997, respectively.

In 1999, he joined the Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble (ENSERG), as an Associate Professor. He is now a Professor at Phelma (Physics, Electronics, and Materials Department, Grenoble-INP), where he lectures (baseband) signal processing, from theoretical aspects to audio applications, including implementation of signal processing algorithms into DSP. His research activity is carried out at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). It concerns different aspects of speech and audio processing (analysis, modeling, coding, transformation, synthesis), with a special interest in joint audio/visual speech processing and speech/audio/music source separation.