Informed audio source separation from compressed linear stereo mixtures

Laurent Girin, Jonathan Pinel

Informed Audio Source Separation from Compressed Linear Stereo Mixtures

Laurent Girin[1], and Jonathan Pinel[1]

[1]*Grenoble Laboratory of Images, Speech, Signal and Automation (GIPSA-lab), Grenoble, France*

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

ABSTRACT

In this paper, new developments concerning a system for informed source separation (ISS) of music signals are presented. Such system enables to separate $I > 2$ musical instruments and singing voices from linear instantaneous stationary stereo (2-channel) mixtures, based on audio signal natural sparsity, pre-mix source signal analysis, and side-information embedding (within the mix signal). The foundations of the system have been presented in previous papers, within the framework of uncompressed (16-bit PCM) mix signals. In the present paper, we extend the study to compressed mix signals. For instance, we use a MPEG-AAC codec and we show that the ISS process is quite robust to compression, opening the way for "real-world" karaoke/soloing/remixing applications for downloadable music.

1. INTRODUCTION

Nowadays, the public is more and more eager to interact with musical pieces during restitution. For example, more and more commercial hits are distributed with several versions, some being instrumental versions (for karaoke), other being remixes. The karaoke phenomenon gets generalized from voice to instruments in musical video games such as *Rock Band* and *Guitar Hero*. In this case, the interaction with the musical piece is based on the inclusion of the multitrack recording in the video game. Indeed, multitrack formats (i.e. concatenation of the separate instruments/voices recordings in a single large file) offer the possibility to manipulate separately the different sound sources while remixing the musical piece in real-time, e.g. controling their volume and spatialization, or adding individual audio effects, leading to a new "active" mode of music listening.

Several multitrack formats are present on the music market, but their use remains marginal for the general public since those formats are confronted to the predominance of the stereo (2-channel) format. Moreover, the music industry is still reluctant to release the multitrack version of musical hits. Because of this historical inertia of usage and technology, it seems quite unrealistic to think that most of the music listeners would be ready to switch rapidly from stereo to multitrack usage if they were offered new interactive possibilities during music listening. In contrast, it is likely that a significant amount of music listeners would entertain the active listening experience if they could access to it from usual stereo format.

This is the goal of the DReaM project.[1] This project aims at developping source separation methods that enable the efficient separation of many different sound sources from only 2-channel mixture signals. Hence we enter the so-called underdetermined difficult (because non-invertible) configuration of the source separation problem. Moreover, the mixture signals should be realistic music pieces, approaching professional music production, and the separation should be processed in real-time with reasonable computation cost, so that real-time source manipulation and remixing can follow. Such ambitious goal is out of reach for current Blind (or semi-blind) Source Separation (BSS) techniques and related Independent Components Analysis (ICA), i.e. separation techniques that only proceed from (determined and overdetermined) mixture signals with very few information about the sources and mixture process characteristics [1][2][3]. It is even still a quite difficult task for informed source separation (ISS) methods, where additionnal prior "external" information about the sources is provided, for example with midi/score files [4][5] or melody humming [6].

Therefore, in the DReaM project, an extreme ISS configuration is considered, in the sense that the source sig-

[1]Disque Repensé pour l'ecoute active de la Musique - This project is supported by the French National Research Agency (ANR) - Grant CONTINT 2009 CORD 006

nals and the mixture process are assumed to be perfectly known. However, the separation does not exploit those information directly: we rather propose a two-step specific coder-decoder configuration. The coder corresponds to the music signal production level (e.g. music recording/mixing in studio) where the source signals are assumed to be available (because of separate recordings) and the mixing process is controlled. The decoder corresponds to the music signal restitution level (e.g. audio-CD at home or mobile player), where only the stereo mix signal is available. Parameters that characterise the source signals and the mixing process are embedded into the mixture signal at the coder level, so that they can be retrieved at the decoder and exploited for source signals separation from the mix signal.

In [7][8], we considered linear instantaneous stationary stereo (LISS) mixtures, a.k.a. constant-gain stereo panning. Basically, the embedded information consists of i) the mixture matrix, and ii) the indexes of the two predominant sources in each time-frequency (TF) region as provided by an analysis of the source signals at the coder. Thus, source separation at the decoder uses those information to perform local 2×2 inversion of the mixture in each TF region. The separation performances obtained by this system are quite impressive: Gains of about 20dB-SDR [9] are obtained for all sources of 5-source mixtures, enabling realistic applications such as generalized karaoke/soloing. Another implementation of the ISS principle within the DReaM project can be found in [10] within the general framework of Non-negative Matrix Factorization (NMF).

The mixture signals considered in [7][8] were 16-bit PCM uncompressed signals, and the side-information was embedded using a high-capacity watermarking technique [11] well suited for such signals: straightly stated, the watermarking is shown to be inaudible and have negligible effects on separation performances. Therefore, the resulting ISS system can be applied to audio-CD and .wav music, and a real-time software demonstrator of the decoder has already been achieved and presented in [12]. Since audio compression [13] has become a very popular tool for ubiquitous listening of "dematerialized" music, audio applications must consider compressed audio formats. The goal of the present study is to demonstrate the feasability of ISS within the audio compression framework. Therefore, in the present paper we still consider LISS mixtures, but the mix signal is compressed, for instance using MPEG-AAC [14]. Because audio compression and watermarking are very difficult to conciliate, the side-information is here transmitted using metadata segments of the compressed binary stream, or adjunct channels, such as in MPEG-SAC/SAOC [15][16]. We show that, although the separation performances are degraded compared to the use of PCM mix signals, the sparseness-based index-embedding ISS technique is remarkably robust to audio compression of the mix signal at reasonable bitrates. We also show that using a refined coding of the side-information enables a quite reasonable additional transmission cost compared to the bitrate of the compressed mix.

This paper is organized as follows. In Section 2, we present an overview of the ISS system based on audio signals sparsity in the TF domain and local inversion of the mixture. The general principle of the system is common for both uncompressed and compressed cases, but some implementation details in Section 2 rather concern the uncompressed case. The adaptation of the system to the new case of compressed mix signals is provided in Section 3. Results obtained for music mixtures are presented in Section 4, before concluding in Section 5.

2. ISS BASED ON SIGNAL T-F SPARSITY AND LOCAL MIX INVERSION

2.1. General Principle

The functional schema of the proposed ISS method is represented on Fig. 1 for the case of uncompressed signals (as previously presented in [7][8]). The decoder is the "active player", that can process separation only on mix signals that have been generated by the corresponding coder. At the coder, the mix signal is generated as a linear instantaneous stationary stereo (LISS) mixture, i.e. summation of individual source signals with constant-gain panning coefficients. Then, the system looks for the two sources that better "explain" the mixture (i.e. the two source signals that are predominant in the mix signal) at different time intervals and frequency channels, and the corresponding source indexes are encoded and embedded into the mixture signal as side-information. At the decoder, the only available signal is the (embedded) mix signal. The side-information is extracted from the transmitted mix signal, decoded and used to separate the source signals by a local time-frequency mixture inversion process. In [7][8], the embedding of the side-information is done by using a high-capacity watermarking method similar to the one presented in [11], and the

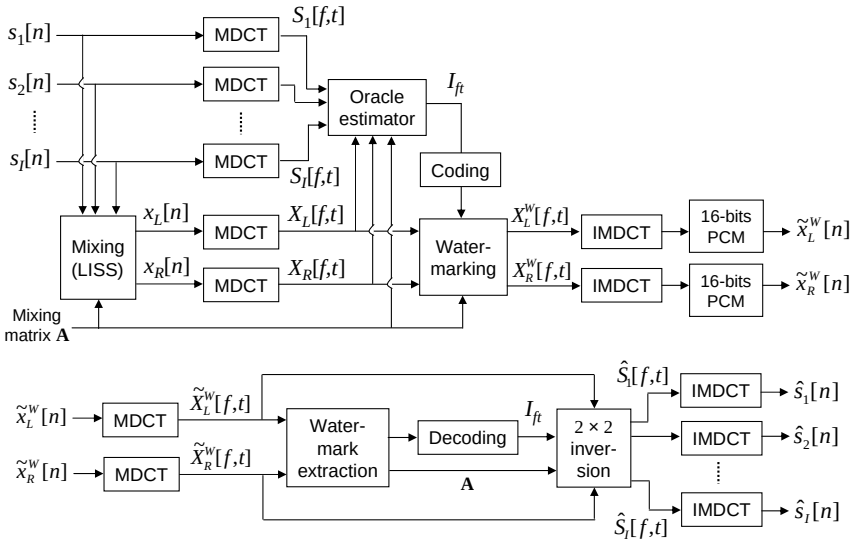

Fig. 1: Functional diagram of the ISS system for uncompressed stereo mix.

(watermarked) mix signal is quantized to 16-bits PCM (for audio-CD / wav file compatibility).

2.2. Time-Frequency Decomposition

The source signals of interest are voice/instrument signals playing a same piece of music (but recorded separately). They are non-stationary, with possibly large temporal and spectral variability, and they generally strongly overlap in the time domain. Decomposing the signals in the time-frequency (TF) domain leads to a sparse representation, i.e. few TF coefficients have a high energy and the overlapping of signals is much lower in the TF domain than in the time domain [17]. Therefore, the separation of source signals can be carried out more efficiently in the TF domain [18][19][20]. The Modified Discrete Cosine Transform (MDCT) [21] is used here 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. We will see in the following that the MDCT is also particularly related to the problem of mix compression.

On Fig. 1, the MDCT is applied on the source signals and on the mixture signal at the input of the coder to enable the selection of predominant sources in the TF domain. In the case of uncompressed signals, watermarking of the resulting side-information is applied on the MDCT coefficients of the mix signal (see [11]) and the time samples of the watermarked mix signal are provided by inverse MDCT (IMDCT). At the decoder, the (PCM-quantized) mix signal is MDCT-transformed and the side-information is extracted from the resulting coefficients. We will see that in the case of compressed signals, no IMDCT reconstruction is necessary at the coder since the compressed binary stream is transmitted (together with the side-information; see Section 3.2). In both uncompressed and compressed cases, source separation is carried out at the decoder also in the MDCT domain, and the resulting separated MDCT coefficients are used to reconstruct the corresponding time-domain separated source signals by IMDCT.

We do not provide here the detailed MDCT equations, since they can be found in many dedicated papers, e.g. [21]. Let us just mention here that the MDCT / IMDCT is applied on signal time frames of $W = 2048$ samples (46.5ms for 44.1kHz-sampling), with a 50%-overlap between consecutive frames (of 1024 frequency bins). The frame length W is chosen to follow the dynamics of music signals while providing a frequency resolution suitable for the separation. Appropriate windowing is applied at both analysis and synthesis to ensure the "perfect reconstruction" property [21].

2.3. Local mix inversion and sources selection

Since the MDCT is a linear transform, the LISS source separation problem remains LISS in the transformed domain. For each frequency bin f and time bin t, we thus have:

$$\mathbf{X}[f,t] = \mathbf{A} \cdot \mathbf{S}[f,t] \tag{1}$$

where $\mathbf{X}[f,t] = [X_1[f,t], X_2[f,t]]^T$ denotes the stereo mixture coefficients vector, $\mathbf{S}[f,t] = [S_1[f,t], \cdots, S_I[f,t]]^T$ denotes the I-source coefficients vector, and $\mathbf{A}$ denotes the mixing matrix that gathers the stereo panning coefficients. Because of audio signal sparsity in the TF domain, only at most 2 sources are assumed to be relevant, i.e. of significant energy, at each TF bin (f,t). Therefore, the mixture is locally given by:

$$\mathbf{X}[f,t] \approx \mathbf{A}_{\mathscr{I}_{ft}} \mathbf{S}_{\mathscr{I}_{ft}}[f,t] \tag{2}$$

where $\mathscr{I}_{ft}$ denotes the set of 2 relevant sources at TF bin (f,t), *i.e.* the two source signals that locally "better explain" the mixture. $\mathbf{A}_{\mathscr{I}_{ft}}$ represents the corresponding 2×2 mixing sub-matrix. If $\overline{\mathscr{I}}_{ft}$ denotes the complementary set of non-active (or at least poorly active) sources at TF bin (f,t), the source signals are estimated by [18]:

$$\begin{cases} \hat{\mathbf{S}}_{\mathscr{I}_{ft}}[f,t] &= \mathbf{A}^{-1}_{\mathscr{I}_{ft}} \mathbf{X}[f,t] \\ \hat{\mathbf{S}}_{\overline{\mathscr{I}}_{ft}}[f,t] &= 0 \end{cases} \tag{3}$$

where $\mathbf{A}^{-1}_{\mathscr{I}_{ft}}$ denotes the inverse of $\mathbf{A}_{\mathscr{I}_{ft}}$ (which is assumed to be invertible and well-conditioned). Note that such separation technique exploits the 2-channel spatial information of the mixture signal and relaxes the restrictive assumption of a single active source at each TF bin, as made in [19][20].

The side-information that is transmitted between coder and decoder (in addition to the mix signal) consists of the coefficients of the mixing matrix $\mathbf{A}$ and the combination of indexes $\mathscr{I}_{ft}$ that identifies the predominant sources in each TF bin. This contrasts with classic blind and semi-blind separation methods where those 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 LISS configuration (only I fixed coefficients for each piece of music, if $\mathbf{A}$ is made of normalized column vectors depending on source azimuths). Therefore, the transmission cost of $\mathbf{A}$ is negligible compared to the transmission cost of $\mathscr{I}_{ft}$, and in the following we do not detail the encoding and transmission of $\mathbf{A}$.

As for the source indexes, the optimal $\mathscr{I}_{ft}$ is estimated at the ISS coder using the source signals, the matrix $\mathbf{A}$, and the mixture signals. This is done using an Oracle estimator, as introduced in [22] for providing upper bounds for the performances of source separation algorithms. Exploiting the reconstruction properties of the MDCT, the overall best separation in the time domain in the mean squared error (MSE) sense is obtained by finding the optimal combination of source signals in the MSE sense at each TF bin separately [22]:

$$\tilde{\mathscr{I}}_{ft} = \arg \min_{\mathscr{I}_{ft} \in \mathscr{P}} \sum_{i=1}^{I} \left(\hat{S}_i[f,t] - S_i[f,t] \right)^2 \tag{4}$$

where $\mathscr{P}$ represents the set of all possible combinations $\mathscr{I}_{ft}$ and the I estimated source signals $\hat{S}_i(f,t)$ are provided by (3). In the case I is limited to a small number of sources (typically about 5 for standard western popular music), $\tilde{\mathscr{I}}_{ft}$ can be found by exhaustive search, and coded with a very limited number of bits before being embedded into the mixture signal (see Section 3.3).

2.4. Source separation at the decoder

At the coder, $\tilde{\mathscr{I}}_{ft}$ is determined from the "original" mix signal (Equation (4)). In contrast, at the decoder, only the transmitted mix signal is available, and the source separation is obtained by applying Equation (3) using $\tilde{\mathscr{I}}_{ft}$ and using the MDCT coefficients of the transmitted mix signal instead of the MDCT coefficients of the original mix signal. In the uncompressed case, these MDCT coefficients are the watermarked coefficients recalculated from the watermaked (and PCM quantized) time-domain signal. It has been shown in [7][8] that the influence of the watermarking (and PCM quantization) on separation performance is negligible (for 5-source mixtures). This is because the side-information bitrate that is used to encode the optimal combination for each TF bin (64kbps per channel; see [8] and below) is significantly lower than the maximal bitrate allowed by the high-capacity watermarking technique (approx. 250kbps per channel [11]). For such bitrate the watermarked MDCT coefficients remain very close to the original MDCT coefficients, hence the separation performances are almost

identical for unwatermarked and watermarked mix signals. In fact, the degradation is mainly due to the "failure" of the sparsity assumption, i.e. the fact that non-predominant but non-null sources interfere in the local inversion process.

Those performances are described in details in [8] for "real-world" 5-source LISS music mixtures of different musical styles. Basically, source enhancement from input (mix) to output (separated) ranges from 17dB to 25dB depending on sources and mixture, which is remarkable given the difficulty of such underdetermined mixtures. The rejection of competing sources is very efficient and the source signals are clearly isolated, as confirmed by listening tests. Artefacts (musical noise) are present but are quite limited. The quality of the isolated source signals makes them usable for generalized karaoke/soloing/remix applications.

3. ISS FOR COMPRESSED MIX SIGNAL

In the present study, MPEG-AAC compression is applied to the mix signal at the output of the coder, and corresponding decoding is applied at the input of the decoder, as illustrated on Fig. 2. This combination of ISS with compression has two main consequences: First, it induces a perturbation of the mixture signal MDCT coefficients that is likely to influence the separation performances; and second it modifies the side-information embedding process. We develop those two points in the next subsections, and we propose a new coding scheme for the side-information that ensures both lower side-info bitrate and more efficient source separation as compared to our previous works [7][8].

3.1. Separation from compressed signal

As for the uncompressed case, the separation at the decoder consists of applying (3) using the MDCT coefficients of the transmitted mix signal instead of the coefficients of the original mix signal. In the new compressed case, the MDCT coefficients $\mathbf{X}^C[f,t]$ are recalculated from the compressed (decoded) mix signal $\tilde{x}^C$. Therefore, we have:

$$\begin{cases} \hat{\mathbf{S}}_{\tilde{\mathscr{I}}_{ft}}[f,t] &= \mathbf{A}^{-1}_{\tilde{\mathscr{I}}_{ft}} \cdot \mathbf{X}^C[f,t] \\ \hat{\mathbf{S}}_{\overline{\tilde{\mathscr{I}}}_{ft}}[f,t] &= 0 \end{cases} \quad (5)$$

As opposed to the watermarking process of the uncompressed case, the compression of the mix signal is here expected to modify the MDCT coefficients $\mathbf{X}^C[f,t]$ in a significant way, i.e. with potential perturbation of the separation process (compared to the use of the original MDCT coefficients $\mathbf{X}[f,t]$). The degree of degradation is expected to depend on the compression bitrate, and this dependancy may not be straightforward: on the one hand, a poor-quality low-bitrate mix signal is expected to lead to degraded source separation, but on the other hand the good quality of a "reasonable-rate" mix ensured by the "global" perceptual masking of quantization error does not guarantee to offer good quality of separated sources. The purpose of this paper is precisely to report the results of an experimental study of the robustness of separation performances to compression, for real-world music signals mixed with the LISS configuration and compressed at different rates (see Section 4).

An important remark can be made here. In the present study we use an MPEG-AAC codec [14]. It happens that MPEG-AAC compression is based on the quantization of MDCT coefficients. Therefore, in a more "compact" structure, the quantization, source selection and separation processes could be applied on the same MDCT coefficients, without recalculation of these coefficients from the decoded signal at the decoder. In the present study we do not achieve such integrated system. Rather, we use an "external" AAC codec, because this approach is the "easy way" to carry first experiments on the merging of ISS and compression without having to investigate any codec source code. This suboptimal approach has the advantage to be independent of the codec: in the diagram of Fig. 2, the AAC codec can be replaced by any other codec without changing the global ISS structure. The development of a (computationaly more efficient) fully integrated AAC-ISS system where the separation is directly carried out from the quantized coefficients is part of our current works. The present study can be seen as a preliminary feasability study for such advanced AAC-ISS merged technology, assuming that the degradation of the MDCT coefficients recalculated from the decoded mix is similar to the degradation of the MDCT coefficients quantized in the compression process.

3.2. Side-information embedding

As mentioned before, in [7][8] a high-rate watermarking technique [11] was used to embed the side-information ($\mathbf{A}$ and $\tilde{\mathscr{I}}_{ft}$) within the mix signal waveform. In the present case, audio compression and watermarking are quite antinomic: it is difficult to simultaneously encode a

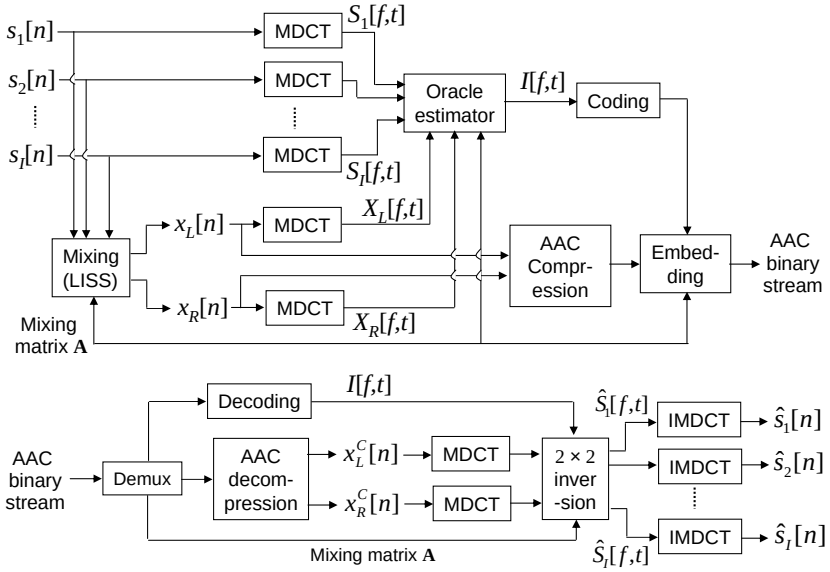

Fig. 2: Functional diagram of the ISS system for compressed stereo mix.

signal or a set of coefficients with as few binary resource as possible, and embed additional information within the same signal or set of coefficients. Therefore, we consider that the side-information is here transmitted using metadata segments of the compressed binary stream. It can also be transmitted using a dedicated channel.[2] In the present paper, we do not detail the metadata segment format and the corresponding embedding of the side-information. However, we provide in Section 4 the side-info bitrates obtained for our test signals, as derived from the variable-rate encoding process described below.

Overall, the spirit of the present ISS-AAC system is connected to the spirit of MPEG-SAC/SAOC [15][16] and related pioneering works on parametric multi-channel joint audio coding (e.g. [25]) where spatialization parameters that enable to respatialize a multi-channel audio scene from a mono/stereo downmix signal are transmitted with the latter. However, it must be noted that in contrast these systems, the goal of the proposed ISS system is to completely separate the source signals and not only to resynthesize/respatialize the audio scene from downmix signals. As a result, the nature of transmitted side-information and the way it is exploited are quite different from spatial audio coding systems. In particular, the spatialization parameters in SAC/SAOC are used to "redistribute" the content of spectral bands of the downmix signal across the different output channels, but they cannot separate the contribution of two different sources that are present within the same band (hence the sources are "respatialized together" and not clearly separated; see, e.g., [25]). In contrast, the separation of two overlapping sources is precisely the goal of our local inversion process, that is carried out for each TF bin.

3.3. Side-information coding

In [7][8], the total number of sources I was fixed for all MDCT frames and frequencies of a given piece of music. Therefore, the number of source combinations was also fixed and $\tilde{\mathscr{I}}_{ft}$ was coded with a fixed number of bits (e.g., 10 combinations of 2 active sources out of 5, coded with 4 bits). In the present study, we refine the coding of $\tilde{\mathscr{I}}_{ft}$ by exploiting the fact that musical sources generally have some temporal sparsity, *i.e.* they sometimes are silent. Therefore, before making the mix, each source signal is labeled into non-silent / silent sections, and for each MDCT frame (i.e. each time bin t), a I-bit code c_1 is transmitted to provide the combination of

[2]Note that the combination of compression and watermarking has been tackled in several prospective papers, e.g., [23][24], but this remains a difficult task, especially for "high" data rates. Moreover, in this specific ISS framework, joint watermarking/compression of MDCT coefficients is likely to provide a higher degradation of the separation than compression alone. This point is part of our future investigations.

non-silent sources. The estimation of $\tilde{\mathscr{I}}_{ft}$ by (3) is then carried out for each frequency bin f only among the non-silent sources, leading to a shorter $\tilde{\mathscr{I}}_{ft}$-code c_2 compared to the fixed-size code of our previous work. Let us give the following example: if only 4 sources out of 5 are non-silent on a given frame, the combination of active sources at each frequency bin only concerns 2 sources out of 4, hence 6 combinations instead of 10, and a 3-bit code can be used instead of a 4-bit code for each frequency bin of the considered frame. Note that i) when only two sources are non-silent on a given frame, the analysis of (3) is useless, the 2 active sources at each frequency bin being the 2 non-silent sources, ii) when one source is non silent, the inversion process for this source is reduced to dividing either the left or right mixture coefficient by the corresponding mixing coefficient, and iii) obviously when all sources are silent, all estimated source coefficients are set to zero. In all those three cases, c_2 is not transmitted, since the decoding of active sources can be made from c_1 alone, leading to a particularly important bit saving.

Finally, this new improved coding scheme has a noticeable consequence on the quality of the separated sources: in our previous works, silent source signals happened to be estimated as non-silent in TF bin where all sources had very low energy and the "two predominant sources" assumption was not verified (a situation that happens more frequently in the higher region of the spectrum) . This resulted in musical noise for the reconstructed sources in sections where the original sources were silent. With the new coding scheme, silent sections are ensured to be reconstructed as silent sections. This provides more natural reconstructed sources and this is a convenient property for karaoke/soloing applications.

4. EXPERIMENTS

4.1. Data and performance measures

Tests have been processed with 44.1kHz-sampled music signals. We report the results for 3 complete songs made of LISS mixtures of 5 sources each (recorded separately in studio conditions), from different artists and different musical styles (see Fig. 3). The mixture matrices $\mathbf{A}$ were set arbitrarily for each piece of music, so that the resulting mixture seems (to us) well musically balanced and the 2×2 submatrices are well-conditioned, e.g.:

$$\mathbf{A} = \begin{bmatrix} 0.93 & 0.80 & 0.71 & 0.60 & 0.50 \\ 0.37 & 0.60 & 0.71 & 0.80 & 0.87 \end{bmatrix} \quad (6)$$

The quality of separated sources has been assessed by both informal listening tests with high-quality headphones, and log power ratios, as defined in [9]. For clarity, we only report here the source-to-distortion ratio (SDR) that provides an overall separation performance criterion (that gathers the influence of interfering sources and artefacts [9]). We also provide in the legends of Fig. 3 typical input signal-to-interference ratios ($\mathrm{SIR_{IN}}$) to underline that musical sources do not contribute with the same power in a well musically balanced mix (note that those input SIR are averaged on sections where all 5 sources are active to avoid under-representation biais). Hence, gains of the form $\mathrm{SDR}-\mathrm{SIR_{IN}}$ are good indicators of separation efficiency.

4.2. Separation results

Fig. 3 provides the average separation SDR obtained with the 3 test signals, for each source signal (averaged over the total duration of the piece of music), and for AAC bitrates ranging from 192 to 64kbps. The results for PCM mix signal are also plotted for comparison (with no watermarking). It can be seen that, as could be expected, the compression of the mix signal leads to a degradation of the output SDR, and this degradation increases as the compression bitrate decreases (since compressed and PCM mix signals differ more and more). However, the degradation is quite regular and moderate untill 128kbps: for example, only 0.5-2dB SDR are lost from PCM to 192kbps for test signals 1 and 3, depending on the source signals, and another 0.5-2dB interval is lost from 192 to 160kbps for all 3 test signals (for test signal 2, the transition from PCM to 192kbps is more abrupt, indicating a possible dependancy of the compression effects on musical style and source composition; this point will be investigated in details in future studies). Since the PCM-reference separation performances range within 15-24dB gains ($\mathrm{SDR}-\mathrm{SIR_{IN}}$), source separation can be extended to compressed source signal at reasonable bitrates, i.e. from 192 to 128kbps, with performances that remain remarkable, given the low quantity of transmitted side-information (see below for additional side-information rates). Below 128kbps, the degradation gets worse, but this is not surprising, since it is in line with the degradation of compressed signal quality as binary resource gets sparse (going from 192 to 160kbps is easier than going from 96 to 64kbps).

Those measures are confirmed by informal listening tests. Although not perfect, the separation quality

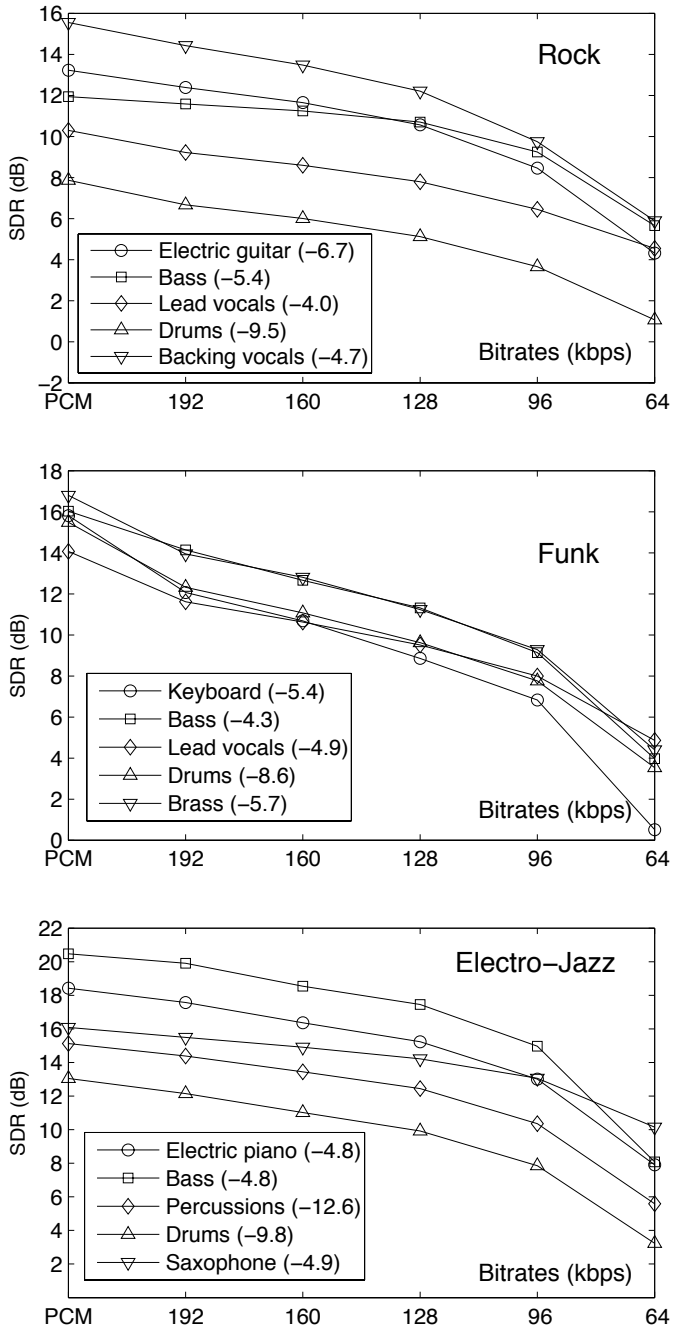

Fig. 3: Separation results for the 3 test signals (resp. durations: 4'39, 4'55, and 5'35). Typical $\mathrm{SIR_{IN}}$ (in dB) are provided in the legends.

obtained with PCM mix signals enables major applications such as remixing or generalized karaoke (i.e. complete suppression of any given source). This has been assessed in our previous studies [7][8]. This result also stands for compressed mix at reasonable bitrates (say, over 128kbps): although the quality of the separated signals from the compressed mix is lower than from the uncompressed mix, this quality is sufficient for individual source enhancement or subtraction. For lower bitrates, a more or less annoying level of musical noise corrupts the estimated sources. In some cases, this noise is correlated with one of the sources, for example it can be correlated with the beat of the drums. The degree of pollution of this noise depends on the considered source signal and on the application. For example, in a remix application (where all sources are remixed with different spatialization and possible addition of audio effects), the noise on a given separated source can be masked by the other remixed sources. In the case of generalized karaoke however, at low compression rates, the noise can prevent the complete suppression of a given source. Demo samples can be downloaded at `http://www.gipsa-lab.fr/~laurent.girin/demo/LISS-AAC-demo.rar`.

4.3. Side-information bitrates

The distribution of the number of non-silent sources and the corresponding side-information bitrates (average coding cost of c_1 and c_2) are provided in Table 1 for the 3 test signals. The additional bitrate is about 100kbps for the Rock and Electro-Jazz examples and it is 77kbps for the Funk piece (revealing its sparser musical arrangement). Even if it not much lower than the mix bitrate, this is quite moderate since it enables to recover 5 source signals from the mix. Note that the side-information bitrate was 128kbps in the basic configuration of our previous works [7][8]. Therefore, the two-step c_1-c_2 coding leads to about 1/3 bit saving. A deepened investigation of the distribution of code c_2 in the TF plan reveals that this distribution is highly structured: c_2 is often identical for large subblocks of the TF plan, revealing the "clustering" organization of the spectral content of the sources in the TF plane. Therefore, the side-information can be coded even more efficiently. This is part of our future works.

Table 1: % of frames with i non-silent sources, $i = 1$ to 5, and corresponding side-information bitrates.

Signal	1	2	3	4	5	bitrate (kbps)
Rock	1.9	0.3	22.3	28.7	46.8	102
Funk	5.1	11.6	27.5	38.2	17.6	77
Elec.-Jazz	3.8	3.9	13.2	31.0	48.0	100

5. CONCLUSION

The sparse ISS method described in [7][8] has been extended to compressed audio signals, with quite robust separation performances at reasonable compression bitrates, and a moderate side-information bitrate (e.g. about 100kbps to separate 5 sources). Specific applications such as active-listening from audio-CD [12] are thus likely to be extended to dematerialized compressed music. The simplicity of the side-information is compensated by an efficient exploitation of signals sparsity and spatial information, as long as the MDCT coefficients are not too much degraded by the compression. This study proves the feasability of a smarter and more computationaly efficient "integrated" ISS-AAC system, where the quantization and separation processes would be carried out on the same MDCT coefficients, as discussed in Section 3.1. In the present study, we did not provide a detailed characterization of the quantization effects on the separation performance (e.g. in different subbands, at different rates, etc.) We plan to do that within the framework of this integrated system, and derive strategies for the control of the bit allocation process by target separation performance. Future extensions of this work will also deal with more complex types of mixture such as binaural/convolutive mixtures, a larger number of sources, and a more efficient coding of the side-information.

6. REFERENCES

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

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

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

[4] S. Dubnov. Optimal filtering of an instrument sound in a mixed recording using harmonic model

and score alignment. In *Int. Computer Music Conf.*, Miami, USA, 2004.

[5] J. Woodruff, B. Pardo, and R. B. Dannenberg. Remixing stereo music with score-informed source separation. In *Int. Conf. on Music Information Retrieval*, pages 314–319, Victoria, Canada, 2006.

[6] P. Smaragdis and G. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications Signal Proc. to Audio and Acoustics*, New Paltz, NY, USA, 2009.

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

[8] 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 Proc.*, Accepted.

[9] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. Speech Audio Proc.*, 14(4):1462–1469, 2006.

[10] A. Liutkus, R. Badeau, and G. Richard. Informed source separation using latent components. In *Int. Conf. on Latent Variable Analysis and Signal Separation*, Saint-Malo, France, 2010.

[11] 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*, Sydney, Australia, 2010.

[12] S. Marchand, B. Mansencal, and L. Girin. Interactive music with active audio CDs. In *Int. Symp. on Computer Music Modeling and Retrieval*, Malaga, Spain, 2010.

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

[14] ISO/IEC JTC1/SC29/WG11 MPEG. Information technology—Generic coding of moving pictures and associated audio information—Part 7: Advanced Audio Coding (AAC), IS13818-7(E), 2004.

[15] J. Herre, K. Kjörling, J. Breebaart, and collegues. MPEG surround - the ISO/MPEG standard for efficient and compatible multi-channel audio coding. In *AES 122nd Convention*, Vienna, Austria, 2007.

[16] J. Herre and S. Disch. New concepts in parametric coding of spatial audio: from SAC to SAOC. In *IEEE Int. Conf. Multimedia and Expo*, Beijing, China, 2007.

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

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

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

[20] 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.

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

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

[23] R. Geiger, Y. Yokotani, and G. Schuller. Audio data hiding with high data rates based on intmdct. In *IEEE Int. Conf. on Acoust., Speech and Signal Proc.*, Toulouse, France, 2006.

[24] F. Siebenhaar, C. Neubauer, and J. Herre. Combined compression/watermarking for audio signals. In *AES 110th Convention*, Amsterdam, The Nederlands, 2001.

[25] C. Faller. Parametric joint-coding of audio sources. In *AES 120th Convention*, Paris, 2006.