

Informed source separation through spectrogram coding and data embedding

Antoine Liutkus, Jonathan Pinel, Roland Badeau, Laurent Girin, Gael Richard

▶ **To cite this version:**

Antoine Liutkus, Jonathan Pinel, Roland Badeau, Laurent Girin, Gael Richard. Informed source separation through spectrogram coding and data embedding. Signal Processing, 2012, 92 (8), pp.1937-1949. 10.1016/j.sigpro.2011.09.016 . hal-00643957

HAL Id: hal-00643957
https://hal.science/hal-00643957

Submitted on 23 Nov 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Informed source separation through spectrogram coding and data embedding

Antoine Liutkus[a], Jonathan Pinel[b], Roland Badeau[a], Laurent Girin[b], Gaël Richard[a]

[a] *Institut Telecom, Telecom ParisTech, CNRS LTCI, 37/39 rue Dareau 75014 Paris, France*
[b] *Grenoble Institute of Technology, 38402 Grenoble Cedex, France*

Abstract

We address the issue of underdetermined source separation in a particular *informed* configuration where both the sources and the mixtures are known during a so-called *encoding* stage. This knowledge enables the computation of a *side-information* which is small enough to be inaudibly embedded into the mixtures. At the *decoding* stage, the sources are no longer assumed to be known, only the mixtures and the extracted side-information are processed for source separation. The proposed system models the sources as independent and locally stationary Gaussian processes (GP) and the mixing process as a linear filtering. This model allows reliable estimation of the sources through generalized Wiener filtering, provided their spectrograms are known. As these spectrograms are too large to be embedded in the mixtures, we show how they can be efficiently approximated using either Nonnegative Tensor Factorization (NTF) or image compression. A high-capacity embedding method is used by the system to inaudibly embed the separation *side-information* into the mixtures. This method is an application of the Quantization Index Modulation technique applied to the time-frequency coefficients of the mixtures and permits to reach embedding rates of about 250kbps. Finally, a study of the performance of the full system is presented.

Keywords: Audio Source Separation, Wiener Filtering, Data Embedding, NTF

1. Introduction

Source separation has been a popular field of research in signal processing for about 20 years (see e.g. [1] for a review and [2] for fundamental principles). Its goal is to recover several signals called *sources* that were mixed together in one or several signals called *mixtures*.

[☆]This work is partly funded by the French National Research Agency (ANR) as a part of the DReaM project (ANR-09-CORD-006-03) and partly supported by the Quaero Programme, funded by OSEO, French State agency for innovation.

One possible classification of audio source separation systems is given by the value of the ratio r between the number of observed mixture signals, K, and the number of sources, M. If this ratio is greater than 1, the problem is called *overdetermined* and source separation systems may then rely on microphone array algorithms and can often reach excellent performance. When the ratio decreases, the problem becomes *determined* ($r = 1$) and can be solved efficiently, for example through *Independent Component Analysis* [3]. When $r < 1$, the problem gets *underdetermined* and systems' performances are hardly predictable and possibly very poor because the source separation problem is then particularly difficult. 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 K is often limited to one or two (respectively for mono and stereo recordings). Separating source signals from such music mixtures is of great interest because it would enable to isolate the different elements of the audio scene, leading to *karaoke* applications, or to separately manipulate them, e.g. by 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.

In the case of USS, there are less observable signals than necessary to solve the underlying mixing equations and separation cannot be achieved without some *prior information* about the signals that permits to eliminate ambiguity. Research on the subject is hence largely focused on how to model any such knowledge and take it into account in the inference process. We can roughly distinguish between three kinds of commonly used prior information.

First, inspired by Auditory Scene Analysis [4], some authors have proposed generative models for the signals of interest, ranging from parametric harmonic models [5, 6] to acoustic generative models inspired from the speech processing community [7]. In this case, source separation consists in estimating the best parameters of these generative models from the observed mixtures. In the Bayesian framework, informative prior distributions are furthermore assigned to these parameters. Such priors can be obtained from query signals as in [8] or from third party Music Information Retrieval (MIR) methods such as fundamental frequency estimation [9, 10] or onsets detection.

Second, much research has also focused on algorithms using prior knowledge about the mixing process.

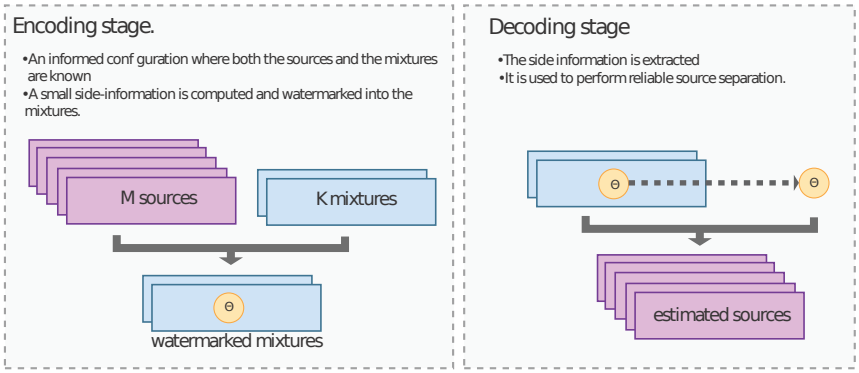


Figure 1: General architecture of an ISS method

In particular, it is common practice in musical production to assign different spatial locations to the sources in multichannel mixtures. Algorithms can then be designed that exploit this spatial distribution of the sources to perform separation [11, 12]. A fundamental property of audio sources is that they are often *sparse* in the Time-Frequency (TF) domain (see [13] and chapter *Sparse Component Analysis* in [1]), which means that few sources are usually active in the same time-frequency bin. Therefore, they can be distinguished by different directions of arrival (DOA), and a filtering of the mixtures depending on the estimated DOA can lead to very good separation performance.

Finally, other methods were proposed in the last few years that decompose the spectrograms of the mixtures into additive elements corresponding to the sources. Decomposition through Nonnegative Tensor Factorization (NTF) techniques [14, 15, 16] was proved to be adequate to this end. In this framework, the spectrograms of the mixtures are decomposed as a time-weighted sum of fixed spectral basis corresponding to the sources. Extensions of NTF methods were subsequently proposed in a Bayesian framework that make use of informative conjugate priors for parameters estimation [17, 18].

In any case, all these methods for USS rely on a TF representation of the mixtures and their objective can be understood as determining the relative contribution of each source in each time-frequency bin. Estimates of the sources are then obtained either through a binary [19] or a soft [20] TF masking strategy. The latter is equivalent to a Wiener filter applied in each frame of the mixtures [21, 14, 22], thus generalizing classical denoising methods to the case of more than only two sources.

In this study, we focus on a special case of USS, called *Informed Source Separation* (ISS) and depicted in Fig. 1. ISS can be understood as an encoding/decoding framework in which both the sources and the

mixtures are available at the *encoder*, but only the mixtures are available at the *decoder*, as well as some *side-information* that has been generated by the encoder and transmitted along with the mixtures to assist the separation process. ISS thus aims at making source separation robust by providing adequate and case-specific prior knowledge to the separation algorithms. Its main advantage is that it permits to reliably recover the separated tracks from mixtures with only a reasonable amount of side-information. This approach was initially proposed in [23, 24] for monophonic linear instantaneous mixtures using Modified Discrete Cosine Transform (MDCT) domain matrix quantization techniques. It has been extended to stereo linear instantaneous mixtures in [25, 26], using local inversion of the mixtures in the TF domain. Even if these studies settle the foundations of the informed approach for source separation within a two-step coding-decoding process, they are confronted to at least two notable limitations: first, they are limited to linear instantaneous mixtures, which is too restrictive if professional music production is to be considered; and second, their performance relies either on a proprietary (i.e. over-specific) source encoding technique or on a "weak form" of the sparsity assumption (in [26], at most two sources are assumed to be predominant in each TF bin) which may fail for a large number of sources or for sources that significantly overlap in the TF plan.

In the present paper, we introduce a new framework for ISS which is based on modeling the sources as *independent* and *locally stationary* GPs that are mixed together through linear filtering into the mixtures [22, 27], thus addressing the more general case of convolutive mixtures. Moreover, the sources are not assumed to be sparse and can possibly significantly overlap in the TF domain. We will show in section 2 that source separation can be achieved very reliably in this framework via generalized Wiener filtering [28, 21, 29, 14], provided the power spectrograms of the sources that are used to build those Wiener filters are available. As those spectrograms are too large to be used directly as the side-information to be transmitted to the decoder, we propose to use dimension reduction techniques such as Nonnegative Tensor Factorization (NTF) or image compression to encode them in a more concise way. This new framework enables to perform ISS with any kind of sources and realistic mixing processes, thus significantly extending the range and potential impact of the informed approach of audio source separation as compared to previously proposed ISS methods.

As for the side-information embedding, we use a high-capacity data embedding technique based on the combination of Quantization Index Modulation (QIM) [30] applied on time-frequency coefficients of the mixture signals, and a Psycho-Acoustic Model (PAM) used to control the inaudibility of the process. This embedding technique has been presented in [31, 32], and already exploited for ISS in [25, 26]. Therefore, we only provide here a general overview of this technique and focus on its use within the new ISS framework. In particular, a thorough evaluation section allows assessing the consequences of the embedding process on source separation quality. Note that this study only focuses on uncompressed, PCM signals.

This paper is organized as follows. In section 2, we detail the model used for the sources and the mixing process as well as the corresponding source separation method. In section 3, we introduce two different approaches for the choice of the parameters to be inaudibly hidden in the mixtures through the high-capacity data embedding technique which is presented in section 4. Finally, we give some experimental results in section 5 in which we study the influence of both the dimensionality reduction technique and data embedding on source separation quality. Finally, we draw some conclusions in section 6.

2. Source separation model

In this section, we make use of the Gaussian Process framework for Source Separation (GPSS) that is presented in full generality in [22]. In the case of audio processing and locally stationary signals, its parameters reduce to the spectrograms of the sources and separation is equivalent to generalized Wiener filtering.

2.1. Notations

Locally Stationary Gaussian Processes (LSGP) [22] are a common model for source separation. An LSGP is a signal whose restriction to a small portion of time called *frame* is a stationary Gaussian process and is independent from the other frames. It is important to note that an LSGP is not necessarily stationary, but only its restrictions to small durations are stationary.

Let $\mathbf{f} \triangleq [f(x_1) \cdots f(x_n)]^\top$ be an LSGP observed on n samples. $\mathbf{f}$ is split into a set $\{\mathbf{f}_t\}_{t=1\cdots N_T}$ of N_T overlapping frames of length L. The Short Term Fourier Transform (STFT) of $\mathbf{f}$ is defined as the

matrix $\mathbf{F}$ whose t^{th} column contains the Fourier transform of $\mathbf{f}_t$. Since the different frames are supposed independent[1], and since the coefficients of the Fourier transform of each frame t are asymptotically independent because $\mathbf{f}_t$ is stationary, all coefficients of the matrix $\mathbf{F}$ are assumed independent. In the following, $\mathbf{F}_{\omega,t}$ will denote the frequency bin ω of frame t.

Let us define the (power) spectrogram $\mathbf{S}$ of $\mathbf{f}$ as $\mathbf{S}_{\omega,t} = |\mathbf{F}_{\omega,t}|^2$. It is readily shown that if f is an LSGP, $\mathbf{F}_{\omega,t}$ follows a centered complex Gaussian distribution of variance $\mathbf{S}_{\omega,t}$ [34, 14, 21].

2.2. Separation of one mixture of locally stationary GPs

Suppose we observe the sum $y(x)$ of M locally stationary signals $f_m(x)$: $y(x) = \sum_{m=1}^{M} f_m(x)$ for n samples. The STFT $\mathbf{Y}$ of the mixture is hence the sum of the STFTs $\{\mathbf{F}_m\}_{m=1\cdots M}$ of the sources. Since we suppose that the sources are LSGPs, each time-frequency bin of $\mathbf{Y}$ is the sum of M independent complex coefficients with Gaussian distribution and can thus be modeled as a complex Gaussian Mixture Model. This model is equivalent to the Gaussian Scaled Mixture Model introduced by Benaroya [20] and further studied by many others [21, 29, 15, 14, 22]. Let $\mathbf{S}_m$ be the power spectrogram of source m. It can be shown that the Minimum Mean Squared Error (MMSE) estimate $\hat{\mathbf{F}}_{m_0}$ of the STFT $\mathbf{F}_{m_0}$ of source m_0 is:

$$\hat{\mathbf{F}}_{m_0} = \frac{\mathbf{S}_{m_0}}{\sum_m \mathbf{S}_m} \cdot \mathbf{Y} \tag{1}$$

where $\frac{A}{B}$ and $A \cdot B$ respectively stand for component-wise division and multiplication of matrices A and B. In audio source separation, this result is referred to as generalized — or adaptive — Wiener filtering.

2.3. Separation of multichannel mixtures

Considering now the more general multichannel case where K mixtures $\{y_k\}_{k=1\cdots K}$ are available (in practice we generally have two channels, for stereo signals). We can assume that each mixture y_k is the sum of filtered versions of the M sources [35]:

$$y_k(x) = \sum_{m=1}^{M} f_{km}(x) \tag{2}$$

[1] Assuming overlapping frames to be independent is classical in the source separation literature. As demonstrated in [22], it leads to overconfidence in the estimates, but provides smooth transitions between the frames. Recent studies such as [33] focus on this issue, which is out of the scope of this paper.

where $f_{km}(x)$ is called the *contribution* of source m to mixture k for time x and is defined as the convolution of source m with a *stable linear filter* a_{km}, that is called the *mixing filter* from source m to mixture k:

$$f_{km}(x) \triangleq (a_{km} * f_m)(x) \triangleq \sum_{\tau=0}^{P} a_{km}(\tau) f_m(x-\tau) \tag{3}$$

We will only consider causal and Finite Impulse Response (FIR) filters of order P here. Provided that P is sufficiently small compared to the length L of each frame, (3) can be simply written in the frequency domain. Typical values for P are 150 in the following, whereas $L \approx 3000$ in our implementation. Let A_{km} be the Fourier transform of the mixing filter a_{km}, $\mathbf{F}_{km}$ be the STFT of f_{km} and $\mathbf{S}_m$ be the power spectrogram of the source m. It is readily shown that the MMSE estimate $\hat{\mathbf{F}}_{km_0}$ of $\mathbf{F}_{km_0}$ is approximately given by[2]:

$$\hat{\mathbf{F}}_{km_0} = \frac{\left(\operatorname{diag}|A_{km_0}|^{\cdot 2}\right)\mathbf{S}_{m_0}}{\sum_{m=1}^{M}\left(\operatorname{diag}|A_{km}|^{\cdot 2}\right)\mathbf{S}_m}.\mathbf{Y}_k \tag{4}$$

This permits to efficiently compute the estimates of the contributions of all the sources within the mixtures provided the mixing filters a_{km} and the spectrograms $\mathbf{S}_m$ are available.

3. Strategies for the side-information

3.1. Introduction

After having introduced source separation for locally stationary GPs, we now focus on the particular *informed* configuration depicted in Fig. 1, where the source signals f_m as well as the mixtures y_k are perfectly known at the encoder. We are then left to produce a set of *parameters* Θ, that we can inaudibly embed into the mixtures to assist source separation at the decoder. In all the following, the signals are assumed locally stationary as in section 2.

The model we introduced in section 2 provides simple ways to estimate the sources in the mixtures. To this end, the separation method given in (1) and (4) requires:

- the mixing filters $\{a_{km}\}_{k=1\cdots K, m=1\cdots M}$,
- the spectrograms $\{\mathbf{S}_m\}_{m=1\cdots M}$ of the sources.

[2] If V is a vector, diag$V$ is the matrix whose diagonal elements are composed of the elements of V. If $\mathbf{M}$ is a matrix, diag$\mathbf{M}$ is the column vector containing the diagonal elements of $\mathbf{M}$.

In the remaining of this section, we will thus devise different possible strategies that allow recovering estimates of $\{a_{km}\}_{k=1\cdots K,m=1\cdots M}$ and $\{\mathbf{S}_m\}_{m=1\cdots M}$ given Θ at the decoder. Note that the informed approach highly contrasts with blind or semi-blind source separation or spectral subtraction methods based on Wiener filtering where the parameters of the Wiener filter (mixing filters and power spectrograms of the sources) have to be estimated from the observations only.

3.2. Definition or estimation of the mixing filters

First of all, there are basically two possibilities for the mixing filters at the encoder. Those filters are either defined during the encoding process of the informed mixtures and thus perfectly known, or the mixing process is made separately (e.g. by a professional sound engineer) and they have to be estimated at the encoder before the encoding process. Previous studies in the automatic mixing literature [35] have focused on the latter case. Given some fixed length $P+1$ for their impulse response, estimation of the mixing filters $\{a_{km}\}_{k=1\cdots K,m=1\cdots M}$ at the encoder is straightforward. Indeed, for every mixture k, everything is available at the encoder to estimate the filters $\{a_{km}\}_{m=1\cdots M}$ that minimize the mean squared error between y_k and $\sum_{m=1}^{M} a_{km} * f_m$ through standard least squares method.

It is noticeable that in real-world scenarios, the mixing process may include very long mixing filters such as reverberations or even nonlinear processing such as compression. In this study, we nonetheless approximate the mixing as a linear filtering of the sources by finite impulse response filters of length $P \approx 150$. As demonstrated in section 5, this approximation yields good results even in nonlinear mixtures, thus suggesting that it is a reasonable simplifying assumption.

In the following discussion, we will thus simply assume that the mixing filters are readily available at the encoder and included in the side information Θ. Note also that in the present study, the mixing filters are assumed to be constant over each whole piece of music to process, hence the amount of side-information reserved to encode those filters is very reasonable (at least compared to the spectrograms information that is computed for each frame).

3.3. Oracle configuration

The proposed separation method requires the spectrograms of the sources at the decoder. A first idea is to simply embed the whole "raw" set of source spectrograms. This would result in $\Theta = \Big\{ \{a_{km}\}_{k=1\cdots K,m=1\cdots M}\,,\ \{\mathbf{S}_m\}_{m=1\cdots M}$

As this setting of Θ is the one that guarantees MMSE estimates, we will call it the *Oracle* configuration in the following. Unfortunately, such a scenario is actually not an option, since we cannot afford to embed the complete spectrograms $\{\mathbf{S}_m\}_{m=1\cdots M}$ inaudibly within the mixtures. More precisely, the total number $\#\Theta_{\text{oracle}}$ of parameters included in Θ in this case is:

$$\#\Theta_{\text{oracle}} = M \times \underbrace{N_\omega \times N_T}_{\text{for each source}} + \underbrace{M \times K \times (P+1)}_{\text{for } \mathbf{A}} \tag{5}$$

which clearly leads to an excessive bitrate considering the embedding capacity: the typical bitrate necessary to embed the parameters in this configuration is 20000kbps[3], which is larger than available: in the embedding method presented in section 4, typical capacity is about 200kbps per mixture signal, which is much insufficient for this purpose. Still, as the Oracle configuration is guaranteed to produce the best results the model is capable of, it will serve as a reference method for evaluation in section 5.

3.4. Dimension reduction techniques

As the spectrograms $\{\mathbf{S}_m\}_{m=1\cdots M}$ of the sources cannot be directly embedded into the mixtures because of insufficient capacity, realistic but effective models have to be devised to encode them concisely. For that purpose, a simple idea we proposed in [27] is to approximate the 3-dimensional tensor $\underline{\mathbf{S}}$ containing the stacked spectrograms of the sources[4].

A first solution is to approximate $\underline{\mathbf{S}}$ as the product of low-rank nonnegative matrices, thus leading to an NTF model [16] that greatly reduces the number of parameters to include in Θ. More specifically, we can chose the Canonical Polyadic (CP) decomposition[5] and approximate $\underline{\mathbf{S}}$ as:

$$\underline{\mathbf{S}}_{\omega,t,m} \approx \hat{\underline{\mathbf{S}}}_{\omega,t,m} = \sum_{r=1}^{R} \mathbf{Q}_{mr}\mathbf{W}_{\omega r}\mathbf{H}_{rt} \tag{6}$$

where $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ are the new parameters Θ_{NTF} of the model, i.e. $\Theta_{\text{NTF}} = \left\{\{a_{km}\}_{k=1\cdots K, m=1\cdots M}, \mathbf{Q}, \mathbf{W}, \mathbf{H}\right\}$. The approximation is graphically illustrated in Fig. 2.

In [27], we have shown that this technique is equivalent to modeling the source signals as a sum of R independent LSGPs with constant normalized power spectrum called *latent components*. This approach leads to finding Θ_{NTF} such that the Itakura-Saito distance[6] $D_{\text{IS}}\left(\underline{\mathbf{S}}|\hat{\underline{\mathbf{S}}}\right)$ between the spectrograms

[3] This approximation is obtained using $M = 5$, $N_\omega = 1024$ and $N_T = 100$frames/s.
[4] That is, $[\underline{\mathbf{S}}]_{\omega,t,m} = [\mathbf{S}_m]_{\omega,t}$.
[5] CP is also called CANDECOMP or PARAFAC [16].
[6] $D_{\text{IS}}(\mathbf{A}|\mathbf{B}) \triangleq \sum_{\omega,t,m}\left[\frac{\mathbf{A}_{\omega,t,m}}{\mathbf{B}_{\omega,t,m}} - \log\frac{\mathbf{A}_{\omega,t,m}}{\mathbf{B}_{\omega,t,m}} - 1\right]$

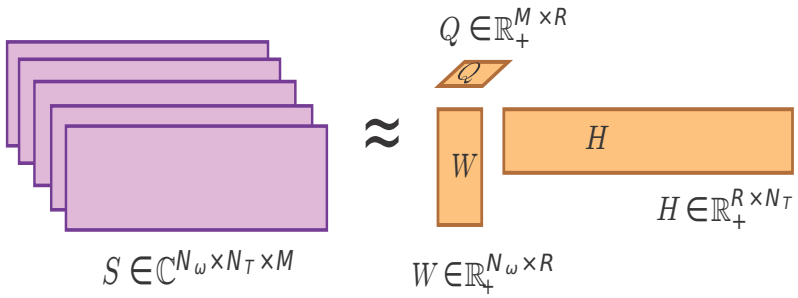


Figure 2: Canonical Polyadic model. The spectrograms of the sources are jointly modeled as an NTF model.

and their reconstruction is minimal. The main advantage of this approach over the Oracle solution as presented in section 3.3 is that the number $\#\Theta_{\text{NTF}}$ of parameters included in Θ_{NTF} becomes:

$$\#\Theta_{\text{NTF}} = \underbrace{N_\omega \times R}_{\text{for } W} + \underbrace{N_T \times R}_{\text{for } H} + \underbrace{M \times R}_{\text{for } Q} + \underbrace{M \times K \times (P+1)}_{\text{for } \mathbf{A}} \tag{7}$$

which is much smaller than $\#\Theta_{\text{oracle}}$ given in (5) as we show in section 5. Typical bitrates necessary to convey Θ_{NTF} indeed drop to approximately 150kbps. It is very important to note that the latent variables $\mathbf{W}$ and $\mathbf{H}$ are the same for all sources, whereas the specific structure of a given source is modeled by $\mathbf{Q}$, hence the very efficient compression power of this representation.

This model has been thoroughly discussed in [29]. For only one source, it is equivalent to the NMF approach that was popularized by [36] when using IS divergence as a cost function, which is a special case of β-divergence for $\beta = 0$ (see [14, 16] on this point). Algorithms in the aforementioned papers can be generalized to the case of 3-dimensional tensors of M channels and the corresponding update rules for the parameters are summarized in Algorithm 1 for any β-divergence[7]. The main idea with those iterative algorithms is to randomly initialize the parameters of the model with nonnegative values, and then iteratively update each matrix by multiplying it component-wise in order to ensure a diminution of the cost function. In Alg. 1, we have also included component-wise exponentiation of the update matrices through an *exponent stepsize* $\eta(\beta)$ as suggested by recent convergence analysis of tensor decompositions

[7] Notations :

- $\mathbf{B}_{m\cdot}$ is the m^{th} row of matrix $\mathbf{B}$,
- $\mathbf{S}_m$ is the observed spectrogram of source m,
- $\hat{\mathbf{S}}_m = \mathbf{W}\text{diag}(\mathbf{Q}_{m\cdot})\mathbf{H}$ is the estimated spectrogram of source m with current model parameters.

Algorithm 1 Update rules for the parameters $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ of the source model (6) for one iteration.

- $\mathbf{Q}_{m.} \leftarrow \mathrm{diag}\left(\mathrm{diag}(\mathbf{Q}_{m.}).\left[\frac{\mathbf{W}^\top\left(\hat{\mathbf{S}}_m^{.\beta-2}.\mathbf{S}_m\right)\mathbf{H}^\top}{\mathbf{W}^\top\hat{\mathbf{S}}_m^{.\beta-1}\mathbf{H}^\top}\right]^{.\eta(\beta)}\right)$
- $\mathbf{W} \leftarrow \mathbf{W}.\left[\frac{\sum_{m=1}^{M}\left(\hat{\mathbf{S}}_m^{.\beta-2}.\mathbf{S}_m\right)(\mathrm{diag}(\mathbf{Q}_{m.})\mathbf{H})^\top}{\sum_{m=1}^{M}\hat{\mathbf{S}}_m^{.\beta-1}(\mathrm{diag}(\mathbf{Q}_{m.})\mathbf{H})^\top}\right]^{.\eta(\beta)}$
- $\mathbf{H} \leftarrow \mathbf{H}.\left[\frac{\sum_{m=1}^{M}(\mathbf{W}\mathrm{diag}(\mathbf{Q}_{m.}))^\top\left(\hat{\mathbf{S}}_m^{.\beta-2}.\mathbf{S}_m\right)}{\sum_{m=1}^{M}(\mathbf{W}\mathrm{diag}(\mathbf{Q}_{m.}))^\top\hat{\mathbf{S}}_m^{.\beta-1}}\right]^{.\eta(\beta)}$

[37, 38, 39]. $\eta(\beta)$ is defined the following way:

$$\eta(\beta) = \begin{cases} \frac{1}{2-\beta} & \text{if } \beta < 1 \\ 1 & \text{if } 1 < \beta \leq 2 \\ \frac{1}{\beta-1} & \text{if } \beta \geq 2 \end{cases}$$

When the parameters have been estimated at the encoder and transmitted to the decoder, separation can then be performed by replacing $\mathbf{S}_m$ in (1) or (4) by its CP approximation (6). Evaluation of this technique for several orders R of the CP approximation is given in section 5.

3.5. Spectrogram compression

Alternatively to NTF decomposition, another original idea that we propose in the present paper to concisely encode the source spectrograms is to compress them using appropriate compression techniques borrowed from the image processing literature [40, 41]. Indeed, each spectrogram $\mathbf{S}_m$ for a given source m is a matrix of nonnegative numbers and can hence be understood (and compressed) as a large image.

Since 8-bit image compression cannot encode more than 256 levels for a grayscale image and since the dynamic range found in typical spectrograms is generally much larger, applying compression algorithms directly on $\mathbf{S}_m$ leads to poorly encoded spectrograms. It was found that instead of encoding $\mathbf{S}_m$, one could rather encode $\log\mathbf{S}_m$, which exhibits much less dynamic variations. In addition, the log-spectrograms can be normalized between $\max(\log\mathbf{S}_m)$ and $\max(\log\mathbf{S}_m) - X\mathrm{dB}$ where X is typically set to 80dB for music spectra (values below the new min value are set to this new min value). Corresponding normalization factors are transmitted within the side-information and used to denormalize the spectrograms at the decoder, but these factors occupy a very limited embedding capacity. Of course, at the decoding stage,

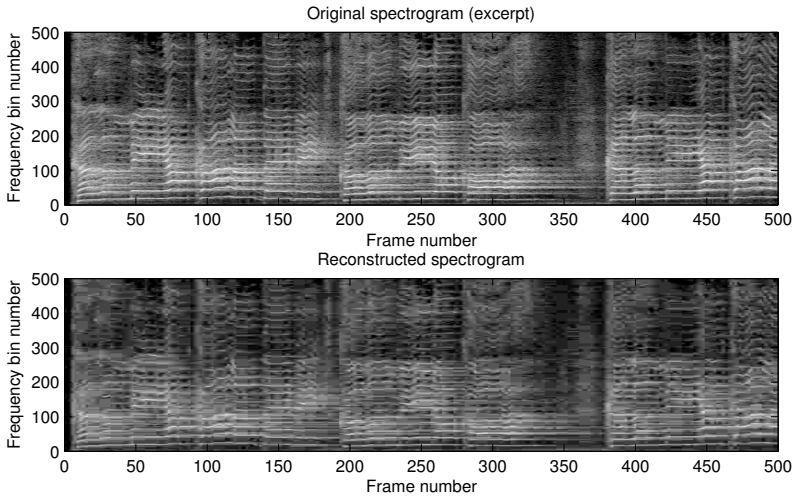


Figure 3: Example of estimated spectrogram using the JPEG algorithm with Quality=10.

log-spectra are converted back to linear scale.

In any case, we consider a compression algorithm denoted $\mathcal{C}$ in the following that produces a buffer of data θ_m which permits to concisely encode the spectrogram $\mathbf{S}_m$ of each source through an image compression algorithm:

$$\theta_m = \mathcal{C}\left\{\mathbf{S}_m\right\}. \tag{8}$$

An estimate $\hat{\mathbf{S}}_m$ of the spectrogram is then obtained by applying the inverse transformation:

$$\hat{\mathbf{S}}_m = \mathcal{C}^{-1}\left\{\theta_m\right\}. \tag{9}$$

Once such an estimate has been obtained, separation can be performed by replacing $\mathbf{S}_m$ in (1) or (4) by its estimate (9). In Fig. 3, we show an excerpt of an original log-spectrogram and its corresponding estimation using the classical JPEG algorithm [41] with a very low quality setting.

When using such an Image Compression (IC) technique, the *side-information* Θ_{IC} to be inaudibly embedded in the mixtures becomes

$$\Theta_{\text{IC}} = \left\{\left\{\theta_m\right\}_{m=1\cdots M},\ \left\{a_{km}\right\}_{k=1\cdots K, m=1\cdots M}\right\} \tag{10}$$

and the corresponding number of parameters and bitrate then depends on the setting of the compression quality in the chosen image compression routine $\mathcal{C}$. In section 5, we see that the capacity required for the transmission of Θ_{IC} lies between 20kbps and 200kbps depending on the Quality setting of the image

compression algorithm. In the following, we will use the standard JPEG compression algorithm for still images [41].

3.6. Conclusion

In this section, we have presented two methods that can be used by the encoder to encode the parameters needed by the decoder to perform source separation in the framework introduced in section 2. These methods lead to a sufficiently small amount of side-information parameters to allow efficient inaudible embedding within the mixtures. The uncompressed, full-accuracy, version of the parameters can be used as a reference oracle configuration providing upper bounds for the separation results.

4. High-capacity data embedding

4.1. Introduction

In this subsection, we present the high-capacity embedding technique that we use to embed the side-information Θ that contains all the hyperparameters needed to perform separation of the mixtures. As mentioned in the introduction, this method has already been presented in [31, 32], and we thus only provide the general lines here. We refer the reader to these references for technical details. The basic principle of the method is that if time-frequency coefficients can be quantized with a limited amount of binary resource without noticeable quality impairments in perceptual audio coding, they can also be modified to embed data. Note that the complete side-information Θ to transmit has to be split and spread across the different TF bins of the different mixtures depending on capacity values, so that each of them carries a small part of the complete message. Conversely, the decoded elementary messages have to be concatenated to recover the complete side-information. For simplicity of presentation, we do not focus here on such implementation details, that are trivial and arbitrary.

4.2. Time-Frequency Transform

As briefly mentioned in section 1, the data hiding process used at the encoder to embed the side-information Θ needed by the decoder to perform source separation is an application of the Quantization Index Modulation (QIM) technique introduced in [30] and is applied to the coefficients of the TF representation of the mixture.

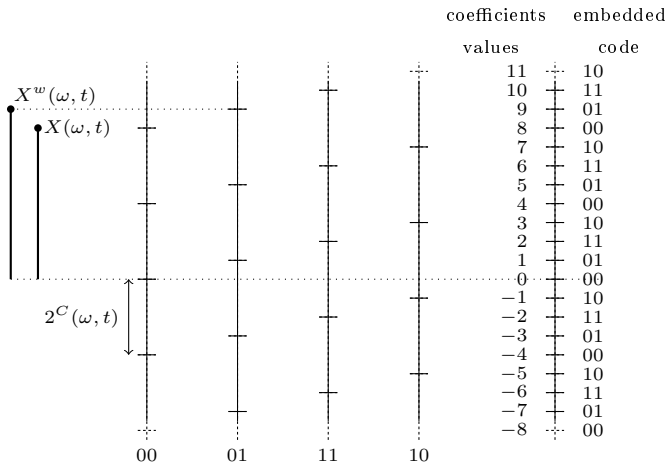


Figure 4: Example of QIM using a set of quantizers for $C(\omega,t)=2$ with their respective Gray code index and a global grid. The binary code 01 is embedded into the IntMDCT coefficient $X(\omega,t)$ by quantizing it to $X^w(\omega,t)$ using the quantizer indexed by 01.

In [31] the Modified Discrete Cosine Transform (MDCT) was used, and in [32] an integer version of the MDCT (IntMDCT, see [42]) has shown to provide improved embedding rates. In addition, IntMDCT maps integer values in the time-domain with integer values in the frequency domain. Therefore, it has the advantage to provide an embedded mix signal which is fully compliant with the PCM format, while an explicit conversion to this format was necessary with the MDCT and introduced noise robustness issues. For these reasons, we use IntMDCT in the present system. In our implementation, the frames are W=2048 samples long (46.5ms for a sampling frequency $f_s = 44.1$kHz), with a 50% overlap between consecutive frames. This results in matrices of IntMDCT coefficients of 1024 frequency bins (denoted by ω) times $n/1024$ time bins (denoted by t; n is the total length of each signal). The time-domain signals are recovered from embedded IntMDCT matrices by frame-wise inverse transformation followed by overlap-add.

4.3. Embedding through IntMDCT quantization

We now present the principle of the embedding process. Let $C(\omega,t)$ denote the capacity at TF bin (ω,t), *i.e.* the size of the binary code to be embedded in the IntMDCT coefficient at that TF bin (under inaudibility constraint). We will see below how $C(\omega,t)$ is determined for each TF bin. For each TF bin (ω,t), a set of $2^{C(\omega,t)}$ uniform quantizers is defined, whose quantization levels are intertwined, and each quantizer represents a $C(\omega,t)$-bit binary code. Embedding a given binary code on a given IntMDCT

coefficient is done by quantizing this coefficient with the corresponding quantizer (i.e. the quantizer indexed by the code to transmit; see Fig. 4).

At the decoder, recovering the code is done by comparing the transmitted coefficient with the $2^{C(\omega,t)}$ quantizers, and selecting the quantizer with the quantization level closest to the transmitted coefficient. Note that because the capacity values depend on (ω, t), those values must also be transmitted in order to select the right set of quantizers. To this purpose, a fixed-capacity embedding "reservoir" is allocated in the higher frequency region of the spectrum, and the capacity values are actually defined within subbands (see [32] for details).

4.4. Embedding rate and performance

The embedding rate ρ is given by the average total number of embedded bits per second of signal. It is obtained by summing the capacity $C(\omega, t)$ over the embedded region of the TF plan and dividing the result by the signal duration. The performance of the embedding process is determined by the inaudibility constraint[8]. Here, the inaudibility constraint induces an upper bound on the number of quantizers, hence a corresponding upper bound on the capacity $C(\omega, t)$ [31, 32]. More specifically, we constraint the power of the embedding error in the worst case to remain under the masking threshold $M(\omega, t)$ provided by a psycho-acoustic model (PAM) inspired from Perceptual Audio Compression [43, 44]. It can be shown that the optimal capacity is given by [31, 32]:

$$C(\omega, t) = \left\lfloor \frac{1}{2} \log_2 \left(M(\omega, t) \right) + 1 \right\rfloor . \tag{11}$$

where $\lfloor . \rfloor$ denotes the *floor* function. The PAM leads to a masking threshold which is signal-dependant and calculated for each signal frame. This masking threshold is inspired from the MPEG-AAC model [45] and was adapted to the present data hiding problem. In particular, it is possible to control the embedding rate by translating the masking threshold by a scaling factor α (in dB), *i.e.* using the following variant of (11):

[8] In [31], where the MDCT was used, robustness to the 16-bit PCM conversion of the embedded signal was also considered as a constraint. This is no more the case with the IntMDCT, since a 16-bit PCM embedded signal is directly generated.

$$C^{\alpha}(\omega,t) = \left\lfloor \frac{1}{2}\log_2\left(M(\omega,t)\cdot 10^{\frac{\alpha}{10}}\right)+1 \right\rfloor. \quad (12)$$

Similarly to the rate-distortion theory in source coding, signal quality is expected to decrease as the embedding rate increases, and vice-versa. When $\alpha > 0$dB, the masking threshold is raised, allowing for larger values of the quantization error and thus larger capacities and embedding rates, at the price of potentially lower quality. On the contrary, when $\alpha < 0$dB, the masking threshold is lowered, leading to a "safety margin" for the inaudibility of the embedding process, at the price of lower embedding rate. An end-user of the proposed system can thus look for the best trade-off between rate and quality for a given application. It can be shown that the embedding rate ρ^{α} corresponding to C^{α} and the basic rate $\rho = \rho^{0}$ are related by:[9]

$$\rho^{\alpha} \simeq \rho^{0} + \alpha \cdot \frac{\log_2(10)}{10} \cdot F_u. \quad (13)$$

where F_u is the bandwidth of the embedded frequency region. This linear relation enables to easily control the embedding rate by the setting of α.

4.5. Conclusion

With this data hiding technique, maximum embedding rates of about 250kbps can be obtained for many musical signals of different styles such as rock, pop, jazz, funk, metal, electro, bossa, fusion, etc. Note that for stereo signals, this embedding capacity is to be understood *per channel*, and is approximately one third of the 16-bit 44.1kHz PCM bitrate necessary to convey the original signals (705kbps/channel)[10]. Such rates generally correspond to the higher level of the masking curve allowed by the PAM and the limit of masking power can hence be reached. More "comfortable" rates can be set between 150 and 200kbps in each channel, to guarantee transparent quality for the embedded signals [32]. This flexibility is used in the present ISS system to fit the embedding capacity with the size of the side-information. Indeed, we see in section 5 that the required capacities to convey the side-information vary from 30kbps to 250kbps, depending on the method.

[9] Actually, the approximation is an exact equality for α multiple of $10\log_{10}(4)$, and the approximation is very good for other values of α, since the embedding rate results from the averaging on a large number of capacity values.

[10] Note that the data embedding technique is not robust to lossy audio compression such as MP3 or AAC. Some perspectives on ISS for compressed music signals are discussed in the conclusion section.

5. Evaluation

5.1. Data and metrics

In order to perform objective evaluation of source separation in general (and not only of informed source separation), the original sources are needed and quantitative evaluation can hence only be performed for mixtures whose constitutive sources are known beforehand. Fortunately, thanks to the rapidly growing community of musicians working with the Creative Commons licenses[11], such material is now readily available. In this study, experiments were carried out with the internal source separation corpus gathered for the Quaero programme[12], from which 15 different excerpts were chosen of various musical styles along with their constitutive separated tracks. The corpus includes excerpts composed of 5 to 11 separated tracks, which are of many kinds, including acoustic instruments such as piano or guitar, male and female singers, distorted sounds/voices, digital effects, etc. All sampling rates were set to 44100Hz and all signals are approximately 30s long.

Since state of the art methods [26, 25] only allow for linear instantaneous mixtures, we created a set of 7 such mixes from our corpus for comparison with the proposed method. For the rest of the evaluation, all mixing was done either in mono (5 excerpts) or in stereo (10 excerpts) using Digital Audio Workstations (DAW). It includes equalizing, panning and digital effects such as reverberation and compression on some excerpts. The real contributions of the different sources into their respective mixtures have also been obtained as the separate outputs of the DAW mixing tools. This permits to quantitatively evaluate the estimation of the sources contributions f_{km}.

Objective criteria to evaluate the quality of the separation were used as defined in the BSSEVAL toolbox [46]. BSSEVAL is a popular evaluation toolbox that is used for the international Signal Separation Evaluation Campaign (SiSEC [47]) and that produces several metrics assessing separation quality for each source. For each estimated source the metrics produced by BSSEVAL are the Source to Distortion Ratio (SDR), the Source to Artifact Ratio (SAR) and the Source to Interference Ratio (SIR). All these metrics are expressed in dB. Whereas the SDR is a global measure of separation performance, the SAR and SIR respectively measure the amount of separation/reconstruction artifacts and the amount of energy from

[11] www.creativecommons.org
[12] www.quaero.org

the other interfering sources. In any case, higher is better. It was shown in [48] that the metrics from BSSEVAL are indeed representative of perceptual separation quality. In the whole evaluation section, averages of the metrics within a mixture are done by weighting the results for its constitutive sources according to the logarithm of their total energy. When comparing with the state of the art, we also use the PEASS toolkit, which is a complementary toolbox for perceptual evaluation.

We first compare the proposed method with the state of the art in section 5.2. Then, we evaluate the impact of the embedding process on separation in section 5.3 and finally, in section 5.4, we compare the different techniques that we introduced to encode the spectrograms of the mixtures in the side-information, namely dimension reduction (see section 3.4) and image compression (see section 3.5). Sounds from these evaluations can be listened to on our webpage[13].

5.2. Comparison with state of the art

We performed a complete test of the state of the art method [26, 25], called PARVAIX in the following, on a subset of our corpus, composed of linear instantaneous mixtures only. The same bitrates (75kbps) were used in all cases. Results can be found in Fig. 5 (left) for BSSEVAL.

As can be seen in Fig. 5 (left), results are very good with both methods. SDR values systematically range between 5dB and 15dB, largely outperforming underdetermined blind source separation. Furthermore, SIR is seen to lie between 10dB and 30dB for PARVAIX and between 10dB and 20dB with the proposed method, thus clearly confirming that the sources are very well separated in both cases. However, it is noticeable that PARVAIX performs better with respect to the BSSEVAL criteria, which are close to signal to noise ratios. Still, it does not perform as well on perceptive grounds and the difference in performance seen on Fig. 5 (left) may be explained by several facts. First, it is not surprising that PARVAIX obtains very good BSSEVAL performance. Indeed, this method aims at optimizing the output signal to noise ratio of the different sources given the assumption that only two sources are active in each TF bin. Doing so, it leads to musical noise in the separated sources, caused by setting many TF bins to zero, contrary to the Wiener filter used in this study. The fact that musical noise is not well accounted for by BSSEVAL has already been noticed in the literature [49]. Second, it is well known that Wiener

[13] See http://www.telecom-paristech.fr/~liutkus/iss/

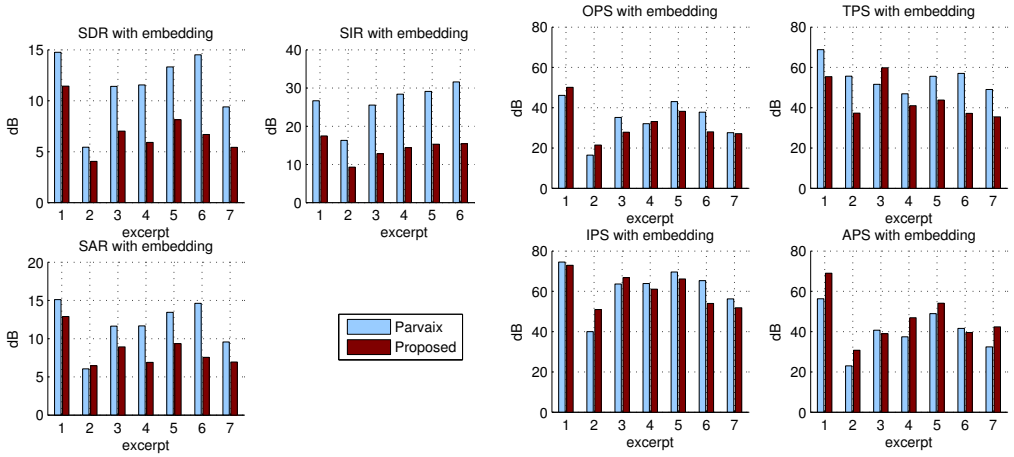


Figure 5: Comparison of the method of Parvaix (red) with proposed method (blue). BSSEval (left) does not account for musical noise whereas PEASS (right) does. OPS, TPS, IPS and APS respectively stand for Overall/Target/Interference/Artifacts Perceptual Score. For all metrics, higher is better.

filtering, which uses the phase of the mixtures, may lead to slight desynchronizations of the signals, that are not perceived perceptually by the human auditory system but that may cause signal to noise ratios to drop dramatically in some cases.

For these reasons, it was demonstrated by Emiya et al. in [49] that further evaluation metrics based on perceptual features may provide better assessment of separation quality. For that purpose, they introduce a toolkit for Perceptual Evaluation of Audio Source Separation (PEASS), available from their website. We therefore evaluated the separated sources with this toolkit[14] and obtained the results given in Fig. 5 (right), which are slightly better for the proposed method, especially when considering artifacts. If the technique proposed by Parvaix handles well sources that have a sparse power spectrogram, it produces poorer estimates of sources with non sparse spectrograms, such as distorted or noisy sounds. On the contrary, the proposed method achieved good performance in every case.

Nevertheless, this evaluation also confirms that objective metrics for the evaluation of source separation is a delicate and open issue, especially for ISS where quality of the estimates is often very good. As recent studies showed that strong connections exist between source coding and ISS [50], evaluation

[14]Very special thanks to A. Ozerov at IRISA (Rennes, France) that assisted us on this point.

for comparing different techniques in ISS may even need to switch from a source separation paradigm as is the case here to perceptual evaluations that are classical in source coding. Still, for the remaining of this study, we use BSSEVAL for evaluation, because *within the same technique*, it gives scores that match perceptive observations. In any case, the reader is encouraged to visit the webpage of this study and listen to the separated tracks himself (see footnote 13).

Anyway, it must be reminded that state of the art is limited to linear instantaneous mixtures only, which is a serious drawback when considering practical applications that involve mixing done by a professional sound engineer. On the contrary, the proposed method proves to be efficient even for mixtures including reverberation of the sources or processed through digital dynamic compressors as demonstrated in the remaining of this evaluation.

5.3. Impact of the embedding process

Since the ISS system presented in section 1 and depicted in Fig. 1 can lead to several kinds of artifacts coming either from data embedding or imperfect Wiener filtering, two evaluations have been performed on the complete dataset. The goal of the first evaluation presented in this section is to assess the impact of the embedding process on separation quality. The second, presented in section 5.4, is to compare the different strategies we proposed for encoding the spectrograms.

To evaluate the impact of embedding, we have performed Oracle separation, i.e. using the original spectrograms of the sources (see 3.3), on the original and embedded mixtures for various bitrates. Since embedding can only decrease performance, we computed the average loss in BSSEVAL criteria induced by embedding. Results are given in Fig. 6, where each line corresponds to one of the 10 stereo signals of our dataset.

Several remarks can be made when considering results in Fig. 6. First, the performance obtained by the Oracle configuration is very good. Perceptually, it is very hard to distinguish the extracted sources from the original. Second, we observe that compared to separation of the unembedded mixtures, the embedding process does not lead to any significant loss in performance up to bitrates of approximately 200kbps for all signals and approximately 250kbps for most signals[15]. Perceptually, it is very hard to

[15] Loud music allows for more embedding

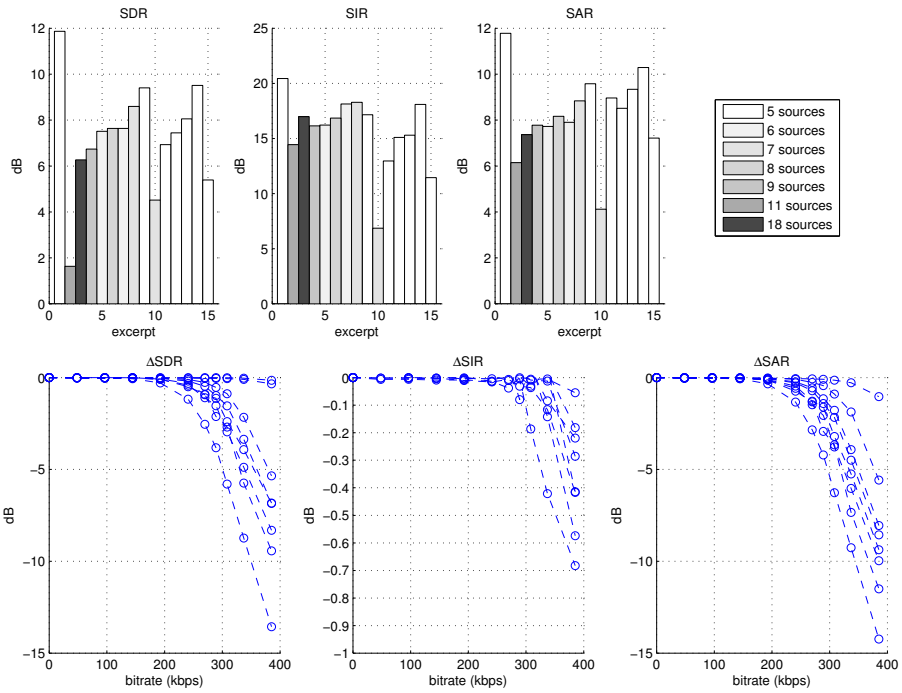


Figure 6: Oracle performance without embedding (above) and loss of performance due to embedding (below) for all excerpts. Higher is better.

notice any difference between the sources extracted through this method and the original source signals, which confirms that the proposed complete model is adequate for the ISS problem. Second, when very high bitrates are required for the embedding of the side-information, Oracle performance drops rapidly after 250kbps. This occurs sometimes with the NTF method on mono mixtures for $R = 150$. These results are very similar to those obtained when assessing the perceptual quality of the embedded mixtures in [32]. This suggests that above 250kbps the mixtures are to degraded for both listening and source separation.

5.4. Comparison of image compression and dimension reduction

The second evaluation performed on the complete dataset concerns comparison of the different methods we suggested in section 3 to encode the spectrograms of the mixtures. The first of these techniques is dimension reduction — called NTF in the following — and has been presented in section 3.4. The number of its parameters Θ_{NTF} is mainly controlled by the number R of components used in the tensor decomposition. The second method uses direct image compression of the spectrograms and it is called IC in the following. It has been presented in section 3.5 and the size of its corresponding side information Θ_{IC} is controlled by the quality parameter of the lossy compression algorithm.

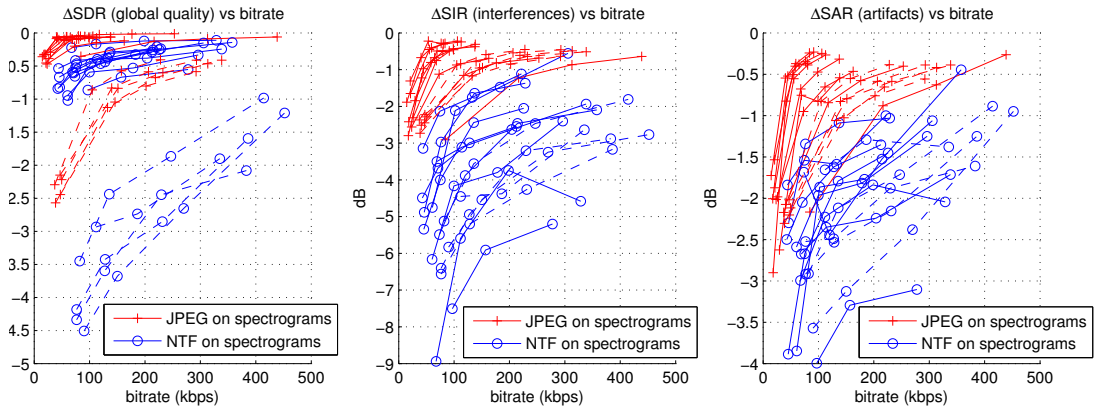


Figure 7: Evaluation with embedding on the 15 excerpts of our corpus and with 4 quality settings for each excerpt and each technique. 0 corresponds to the Oracle performance in each case. Solid and dashed lines are for stereo and mono excerpts respectively.

Both techniques were evaluated with various sizes for their corresponding side information. Since their performance is bounded above by the Oracle technique, we computed the difference between the results given by BSSEVAL on each case with the corresponding Oracle performance (with the same embedded mixtures). Results are given in Fig 7.

Considering Fig 7, we see that performance obtained by IC or NTF is directly controlled by the quality setting considered: the quality parameter of the image compression algorithm for IC and R for NTF[16]. This can be explained by the fact that sophisticated models permit very reliable estimates of the spectrograms whereas small models only permit crude modeling. Still, large models also lead to higher bitrates, and thus to a degradation of the performance as demonstrated in section 5.3. This suggests that a trade-off is to be found between the improvement of spectrograms modeling through higher quality settings for side-information, and the corresponding loss in performance induced by embedding more data. For example, Fig 7 shows that the loss in SDR induced by using IC rather than Oracle is lower than 1dB at 200kbps and Fig. 6 shows that the loss induced by embedding at this rate is negligible. It is not interesting to increase the side information rate to 300kbps because the loss due to embedding in this case is higher than the gain induced by better encoding of the spectrograms.

In any case, we see that both techniques are very close to Oracle performance, which is very satis-

[16] In NTF, we used the same number of 60 iterations for Alg. 1 for all cases in our evaluation, which may not be sufficient for very large R, explaining why high bitrates do not necessarily lead to better performance for NTF.

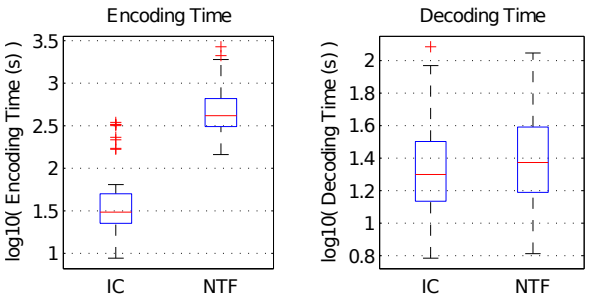


Figure 8: Boxplots of encoding/decoding times over all experiments, depending on the method (IC or NTF).

factory. Furthermore, considering Fig. 8 on which boxplots[17] of the encoding/decoding times[18] over all experiments are displayed, we see that source separation at the decoder can be done extremely rapidly and does not require large computational resources. Whatever the chosen method, the complete system thus permits to successfully model the spectrograms of the sources and embed the corresponding information within the mixtures. The sources recovered through generalized Wiener filtering at the decoder are seen to be very well separated, with almost no interferences and only a very small amount of artifacts. *Active listening* applications such as karaoke or remastering are hence made possible with such a system on realistic mixtures.

Now, comparing the respective performance of IC and NTF, we see that in our evaluation, IC always yields better results than NTF for a given bitrate and a given excerpt. However, this fact may be tempered by noting that the parameters for NTF are not optimally coded in this study, leading to an overestimation of the bitrate necessary to convey Θ_{NTF}. Future work may hence demonstrate better performance for NTF than obtained here, notably through appropriate compression of Θ_{NTF}. Still, considering encoding times given in Fig. 8 we see that IC performs approximately 10 times faster for coding than the NTF method. Even if the image compression method we used is implemented in an extremely mature and optimized library, contrarily to our Matlab implementation of NTF, this difference of encoding time between IC and NTF is mostly explained by the fact that the computations presented in Algorithm 1 are performed many times for NTF, thus leading to heavy computational costs, whereas JPEG encoding of a spectrogram only requires quantization of a cosine transform and can be performed very rapidly [41].

[17] See e.g. `http://en.wikipedia.org/wiki/Box_plot`

[18] Experiments were done on a Quad Core computer with 4Gb RAM.

Computational complexity of NTF is hence inherently much greater than IC. Altogether, spectrograms compression through the IC method leads to lower necessary bitrates, higher separation quality and lower computational costs. It is thus the preferred method to perform ISS[19].

6. Conclusion

Informed source separation consists in providing valuable prior knowledge to a source separation algorithm. This study considered the case where this knowledge has been computed at an *encoding* stage where both the mixtures and the original sources are known, and then inaudibly embedded into the PCM mixtures through an adequate high-capacity data embedding technique. At the *decoding* stage, this side-information is extracted and separation is performed using the side-information.

The statistical framework presented in this study models the sources as locally stationary Gaussian processes that are mixed using linear FIR filters. We have shown that very reliable source separation can be performed in this case when the spectrograms of the sources are known at the decoder. Since these spectrograms are much too large matrices to be possibly embedded into the mixtures, several approaches were proposed to approximate them, including dimension reduction through tensor factorization and image compression.

In this study, we have performed a thorough evaluation of the proposed ISS method. The corpus we considered includes sources that are highly non-sparse in the frequency domain, such as distorted drums or guitars and mixtures that were obtained with professional DAW, including non-linear compressions. These settings correspond to usecases that were not possibly handled by previous methods proposed for ISS. In any case, performance was extremely encouraging and the proposed model based on generalized Wiener filtering of the mixtures is hence adequate for the ISS problem and allows *active listening* applications of musical content such as karaoke or musical remixing.

Interesting features of the proposed statistical method for informed source separation are first that the embedding process at appropriate rate does not perceptually nor quantitatively lead to significant

[19] Other lossy image compression techniques such as JPEG2000 have also been tested, but because of the lack of space, we decided to only report on JPEG, since it is a free, readily available algorithm. Note however that JPEG2000, based on wavelet transforms, exhibited better separation performance than JPEG for very low side-information bitrates: SDR, SAR and SIR were respectively approximately 0.3dB, 1.5dB and 1.5dB better for JPEG2000 than for JPG.

degradation during the source separation step and second that it makes it possible to use any algorithm suitable for image compression as a candidate for side-information. Possible extensions to this work would include synchronized data embedding to allow real-time separation at the decoder and embedding methods that are either robust to compression algorithms such as MP3 or that benefit from dedicated adjunct channels as for example defined in MPEG-SAC/SAOC [51, 52]. Preliminary experiments show that the proposed separation method is quite robust to audio compression of the mixture signals.

References

[1] P. Comon, C. Jutten (Eds.), Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution, Academic Press, 2010.

[2] J.-F. Cardoso, Blind signal separation: statistical principles, Proceedings of the IEEE 90 (1998) 2009–2026.

[3] A. Hyvärinen, J. Karhunen, E. Oja (Eds.), Independent Component Analysis, Wiley and Sons, 2001.

[4] A. Bregman, Auditory Scene Analysis, The perceptual Organization of Sound, MIT Press, 1994.

[5] M. Davy, S. Godsill, Bayesian harmonic models for musical signal analysis, in: Bayesian Statistics, Vol. 7, Oxford University Press, 2002.

[6] T. Virtanen, Algorithm for the separation of harmonic sounds with time-frequency smoothness constraint, in: Proc. of the 6th Conf. on Digital Audio Effects (DAFx-03), London, UK, 2003, pp. 35–40.

[7] J.-L. Durrieu, G. Richard, B. David, An iterative approach to monaural musical mixture de-soloing, in: Proc. IEEE Intl. Conf. Acoust. Speech Signal Processing (ICASSP'09), Washington, DC, USA, 2009, pp. 105–108.

[8] P. Smaragdis, G. J. Mysore, "separation by humming": User guided sound extraction from monophonic mixtures, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2009.

[9] T. Virtanen, A. Mesaros, M. Ryynänen, Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music, in: Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, Brisbane, Australia, 2008.

[10] J. Ganseman, P. Scheunders, G. J. Mysore, J. S. Abel, Source separation by score synthesis, in: Proc. of the Intl. Computer Music Conference (ICMC'2010), New York, NY, 2010.

[11] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking, IEEE Trans. on Signal Processing 52 (7) (2004) 1830–1847.

[12] N. Duong, E. Vincent, R. Gribonval, Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation, in: Proceedings of the 9th international conference on Latent variable analysis and signal separation, LVA/ICA'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 73–80.

[13] M. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, M. Davies, Sparse Representations in Audio and Music: from Coding to Source Separation, Proceedings of the IEEE. 98 (2010) 995–1005.

[14] C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis, Neural Computation 21 (3) (2009) 793–830.

[15] A. Ozerov, E. Vincent, F. Bimbot, A general modular framework for audio source separation, in: Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010), St. Malo, France, 2010, pp. 33–40.

[16] A. Cichocki, R. Zdunek, A. H. Phan, S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, Wiley Publishing, 2009.

[17] O. Dikmen, A. Cemgil, Unsupervised single-channel source separation using Bayesian NMF, in: Proc. of the 2009 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'09), NY, USA, 2009, pp. 93–96.

[18] O. Dikmen, A. T. Cemgil, Gamma Markov random fields for audio source modelling, IEEE Transactions on Audio, Speech, and Language Processing 18 (3) (2010) 589–601.

[19] S. T. Roweis, One microphone source separation, in: Advances in Neural Information Processing Systems, Vol. 13, MIT Press, 2001, pp. 793–799.

[20] L. Benaroya, F. Bimbot, R. Gribonval, Audio source separation with a single sensor, IEEE Trans. on Audio, Speech and Language Processing 14 (1) (2006) 191–199.

[21] A. Cemgil, P. Peeling, O. Dikmen, S. Godsill, Prior structures for Time-Frequency energy distributions, in: Proc. of the 2007 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'07), New Paltz, USA, 2007, pp. 151–154.

[22] A. Liutkus, R. Badeau, G. Richard, Gaussian processes for underdetermined source separation, Signal Processing, IEEE Transactions on 59 (7) (2011) 3155 –3167.

[23] M. Parvaix, L. Girin, J.-M. Brossier, A watermarking-based method for single-channel audio source separation, in: Proc. IEEE Intl. Conf. Acoust. Speech Signal Processing (ICASSP'09), Taipei, Taiwan, 2009, pp. 101–103.

[24] M. Parvaix, L. Girin, J.-M. Brossier, A watermarking-based method for informed source separation of audio signals with a single sensor, IEEE Trans. Audio, Speech, and Language Process. 18 (6) (2010) 1464–1475.

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

[26] M. Parvaix, L. Girin, Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding, Audio, Speech, and Language Processing, IEEE Transactions on 19(6) (6) (2011) 1721 –1733.

[27] A. Liutkus, R. Badeau, G. Richard, Informed source separation using latent components, in: Proc.

9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010), St. Malo, France, 2010, pp. 33–40.

[28] L. Benaroya, L. McDonagh, F. Bimbot, R. Gribonval, Non negative sparse representation for Wiener based source separation with a single sensor, in: Proc. IEEE Intl. Conf. Acoust. Speech Signal Processing (ICASSP'03), Hong-Kong, 2003, pp. 613–616.

[29] A. Ozerov, C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, IEEE Trans. on Audio, Speech and Language Processing 18 (3) (2010) 550–563.

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

[31] J. Pinel, L. Girin, C. Baras, A high-capacity watermarking technique for audio signals based on MDCT-domain quantization, in: Proceedings of the 20th International Congress on Acoustics, Sydney, 2010.

[32] J. Pinel, L. Girin, C. Baras, High-rate data embedding in uncompressed music signals using QIM and IntMDCT, in: Proc. of Digital Audio Effects Workshop (DAFx), Paris, France, 2011.

[33] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, S. Sagayama, Consistent Wiener filtering: Generalized time-frequency masking respecting spectrogram consistency, in: Proc. 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2010), St. Malo, France, 2010, pp. 89–96.

[34] F. Neeser, J. Massey, Proper complex random processes with applications to information theory, Information Theory, IEEE Transactions on 39 (4) (1993) 1293 –1302.

[35] D. Barchiesi, J. Reiss, Automatic target mixing using least-squares optimization of gains and equalization settings, in: Proc. of the 12th Conf. on Digital Audio Effects (DAFx-09), Como, Italy, 2009, pp. 7–14.

[36] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems (NIPS), Vol. 13, The MIT Press, 2001, pp. 556–562.

[37] R. Badeau, N. Bertin, V. Emmanuel, Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization, Trans. Neur. Netw. 21 (2010) 1869–1881.

[38] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, S. Sagayama, Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence, in: Proc. 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010), 2010.

[39] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the beta-divergence, Neural Computation 23 (9).

[40] J. Woods, Multidimensional Signal, Image, and Video Processing and Coding, Academic Press, Inc., Orlando, FL, USA, 2006.

[41] G. Wallace, The JPEG still picture compression standard, Commun. ACM 34 (1991) 30–44.

[42] R. Geiger, J. Herre, J. Koller, K. Brandenburg, IntMDCT - a link between perceptual and lossless audio coding, in: Proc. IEEE Intl. Conf. Acoust. Speech Signal Processing (ICASSP'93), Vol. 2, 2002, p. II.

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

[44] T. Painter, A. Spanias, Perceptual coding of digital audio, Proceedings of the IEEE 88 (2000) 451–515.

[45] 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).

[46] E. Vincent, C. Févotte, R. Gribonval, Performance measurement in blind audio source separation, IEEE Trans. on Audio, Speech and Language Processing 14 (4) (2006) 1462–1469.

[47] E. Vincent, S. Araki, P. Bofill, The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation, in: ICA '09: Proc. of the 8th Int. Conf. on Independent Component Analysis and Signal Separation, Berlin, Heidelberg, 2009, pp. 734–741.

[48] J. Kornycky, B. Günel, A. M. Kondoz, Comparison of subjective and objective evaluation methods for audio source separation, in: Proceedings of Meetings on Acoustics, Vol. 4, ASA, 2008, pp. 1–10.

[49] V. Emiya, E. Vincent, N. Harlander, V. Hohmann, Subjective and objective quality assessment of audio source separation, IEEE Transactions on Audio, Speech, and Language Processing.

[50] A. Ozerov, A. Liutkus, R. Badeau, G. Richard, Informed source separation: source coding meets source separation, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11), Mohonk, NY, 2011, submitted.

[51] J. Herre, K. Kjörling, J. Breebaart, colleagues, MPEG surround - the ISO/MPEG standard for efficient and compatible multi-channel audio coding, in: AES 122nd Convention, Vienna, Austria, 2007.

[52] J. Engdegård, C. Falch, O. Hellmuth, J. Herre, J.Hilpert, A. Hölzer, J. Koppens, H. Mundt, H. Oh, H. Purnhagen, B. Resch, L. T. andM. L. Valero, L. Villemoes, MPEG spatial audio object coding — the ISO/MPEG standard for efficient coding of interactive audio scenes, in: Audio Engineering Society Convention 129, 2010.