HAL
open science

# Low bitrate informed source separation of realistic mixtures

Antoine Liutkus, Roland Badeau, Gael Richard

## To cite this version:

Antoine Liutkus, Roland Badeau, Gael Richard. Low bitrate informed source separation of realistic mixtures. ICASSP, 2013, Vancouver, Canada. pp.66–70, 10.1109/ICASSP.2013.6637610 . hal-00945299

**HAL Id: hal-00945299**

**https://inria.hal.science/hal-00945299v1**

Submitted on 13 Mar 2014

**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.

# LOW BITRATE INFORMED SOURCE SEPARATION OF REALISTIC MIXTURES

*Antoine Liutkus* *Roland Badeau* *Gaël Richard*

Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, France

**ABSTRACT**

Demixing consists in recovering the sounds that compose a multichannel mix. Important applications include karaoke or respatialization. Several approaches to this problem have been proposed in a coding/decoding framework, which are denoted either as spatial audio object coding or informed source separation. They assume that the constituent sounds are available at an encoding stage and used to compute a side-information transmitted to the end-user. At a decoding stage, only the mixtures and the side information are used to recover the sources. Here, we propose an advanced model, which encompasses many practical scenarios and permits to reach bitrates as low as 0.5kbps/source. First, the sources may be mono or multichannel. Second, the mixing process is not assumed to be linear-instantaneous or convolutive as is usual, but rather diffuse, permitting professional mixes to be processed. Third, the signals to be recovered may either be the original sources or their images.

***Index Terms—*** audio upmixing, Wiener filtering, spatial audio object coding, informed source separation

## 1. INTRODUCTION

The ability to recover the constituent audio signals from their multichannel mixtures is at the core of many applications of audio signal processing. Among them, we can mention karaoke, which consists in muting one of the instruments, usually the voice signal. Another important application is respatialization, which consists in dynamically modifying the spatial positions of the different audio signals *within* the mixtures. This processing is important in recent entertainment applications such as videogames or 3D-movies, where positions of the sources vary constantly over time.

A first naive way to permit such applications is to separately encode all the constituent sounds at the coder and transmit them as such to the decoder. In that case, the desired mixtures are automatically constructed at the decoder using the available separated sounds. This strategy faces three major drawbacks. First, it requires a lot of bitrate. Indeed, the separate encoding of all the audio sources requires bitrates of at least, say, 32kbps/source to be of reasonable quality, leading to high bitrates if the number of sources is important. Second, this strategy does not permit to benefit from a professional mixing. Indeed, correctly mixing audio signals is a difficult art mastered by sound engineers, which requires expert knowledge to be of professional quality and which cannot easily be imitated by an automated process. Finally, transmitting separated signals is often not considered as a viable option by copyrights owners, who are very reluctant to broadcast the separated tracks from famous songs.

Hence, much research has focused on how to efficiently encode the constituent sources present in an audio mixture in a coding/decoding framework. At the coder, both the constituent tracks and the mixtures are known and a side-information is produced. At the decoder, only the mixtures and the side-information are processed to recover the sources. In the literature, this problem has been addressed independently by two distinct communities. First, techniques from spatial audio coding (SAC [7, 1]) have been proposed to handle this particular scenario where the signals to recover at the decoder are separated audio *objects*, rather than a spatialized version of the transmitted *downmix* as in SAC. The resulting methods are referred to as Spatial Audio Object Coding (SAOC, [8, 3, 4]) and are currently in the process of standardization by the MPEG group. Independently from SAOC, researchers from the source separation community have addressed the same exact problem using sophisticated source separation techniques [17, 14, 12, 15]. The resulting methods are commonly referred to as *Informed Source Separation* (ISS) in this community. Interestingly enough, bridges between source separation and audio coding have recently emerged [19] and theoretical analysis of ISS in terms of source coding have been proposed in [16, 13].

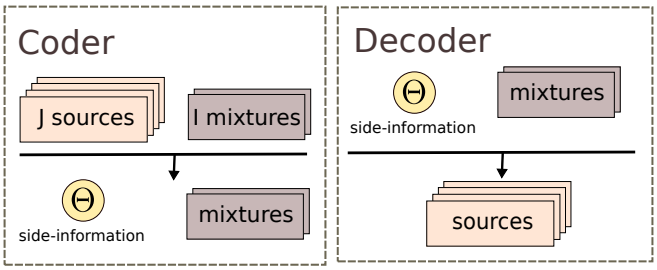


**Fig. 1**. High-level ISS/SAOC scheme

Both SAOC and ISS share the same general framework depicted in figure 1 : the signals to be recovered at the decoder are only observed through some kind of *downmix*, which is to be separated. The operations performed to this end vary from one technique to the other, but the common strategy of all those techniques is to assume that the sources can be efficiently recovered through a *filtering* of the mixtures at the decoder. While some methods [17] perform a local inversion of the mixing process in the time-frequency domain, others make use of an optimal filtering strategy [4, 3, 15, 12]. The mixing process to be inverted is either modeled as linear instantaneous [8, 3, 18, 17, 15] or convolutive [14, 12]. In any case, the side-information which is transmitted from the coder to the decoder usually consists in the optimal parameters to use for the filtering.

Existing informed separation methods exhibit several common drawbacks. First, they are restrictive with respect to the mixing process considered. Most of them rely on the assumptions that the observed mixtures are either linear instantaneous or convolutive and that the mixing parameters are known. These assumptions prevent the processing of real professional mixtures, which may exhibit some non-linearities or some spatial spreading of the sources, due

This work is supported by the Quaero Program, funded by OSEO, French State agency for innovation.

to the use of advanced audio effects. A second restriction of most methods is that they either focus on the recovery of the original mono sources [17, 12, 6] (hence permitting high-quality respatialization) or on the mere separation of source images [10, 14], which is sufficient for karaoke applications. To the best of our knowledge, there is no available method that can handle both cases in a principled way. Finally, there is no available demixing method that can efficiently cope with multichannel sources : all techniques assume that the sources are mono signals, which may not be true in the case of music production (think of the output of a synthesizer).

In this paper, we propose a general Gaussian framework which aims at addressing all these issues. To this purpose, we model the sources as locally stationary Gaussian processes and the mixing process as diffuse — or full-rank — as recently introduced in [2]. In this framework, the sources can be observed at the coder either as mono or multichannel, and can be recovered at the decoder either as observed at the coder or as images, i.e. as they appear within the mixtures. The parameters of this model can be encoded very efficiently through a Nonnegative Tensor Factorization [14, 15] or through image compression algorithms [14]. Resulting bitrates can get as low as 0.5kbps/source, for good separation quality.

This paper is organized as follows. First, we present the notations and models in section 2.1. Then, we detail all the algorithms required by the corresponding coding/decoding framework in section 3. Finally, we provide some experimental results in section 4 which demonstrate the efficiency of the approach.

## 2. NOTATIONS AND MODEL

### 2.1. Notations

At the coder, the observed signals are the waveforms of the *sources* $\tilde{\mathbf{s}}$ and the *mixtures* $\tilde{\mathbf{x}}$. We assume that there are $J$ sources and $I$ mixtures to consider. A typical case is $I = 2$ for stereo mixtures. All waveforms considered are assumed to be of the same length $L$.

Some of the sources may be observed as mono sources. Let $\mathcal{S}_p \subset \mathbb{N}_J$ be the indices of those sources [1] and let $\mathcal{S}_d = \mathbb{N}_J \setminus \mathcal{S}_p$ be the indices of the sources which are observed as multichannel signals. We assume that all multichannel sources have $I$ channels, just like the mixtures (e.g. stereo sources for stereo mixtures) . Let $\tilde{\mathbf{s}}(\cdot,\cdot,j)$ denote the observed waveforms for source $j$. Its dimension is $L \times 1$ if $j \in \mathcal{S}_p$ or $L \times I$ if $j \in \mathcal{S}_d$. Similarly, let $\tilde{\mathbf{x}}(t,i)$ be the observed value of the mixture $i$ at time $t$.

Whereas an observed source $\tilde{\mathbf{s}}(\cdot,\cdot,j)$ may be mono or multichannel, its *image* $\tilde{\mathbf{y}}(\cdot,\cdot,j)$ within the observed mixtures necessarily has $I$ channels. The image of a source is defined as how it ends up *within* the mixtures. For example, if a stereophonic mixture is build from three monophonic sources such as voice, bass and piano, it is not the monophonic sum of these sources which is observed. Rather, they are spatialized so as to produce their corresponding *images*. The observed mixtures are simply defined as the sum of those images :

$$\forall (t,i), \tilde{\mathbf{x}}(t,i) = \sum_{j=1}^{J} \tilde{\mathbf{y}}(t,i,j).$$

Those notations being given, we will not process the signals in the time domain, but rather in a Time-Frequency representation. In this paper, we will consider the Short-Term Fourier Transform (STFT), which consists in splitting each signal considered in small overlapping frames before applying a Fourier transform on each of them. The resulting STFTs will be denoted without the tilde notation. Hence :

- $\mathbf{s}(f,n,\cdot,j)$ denotes the — complex — observed values of the STFTs of source $j$ at Time-Frequency (TF) bin $(f,n)$. If $j \in \mathcal{S}_p$, it is a complex single $(1 \times 1)$ value, because that source is monophonic. If $j \in \mathcal{S}_d$, it is a $I \times 1$ vector, because that source is mutichannel.
- $\mathbf{x}(f,n,\cdot) = [\mathbf{x}(f,n,1),\ldots,\mathbf{x}(f,n,I)]^\top$ is the $I \times 1$ vector gathering the $I$ coefficients of the STFTs of the mixtures at TF bin $(f,n)$.
- $\mathbf{y}(f,n,\cdot,j) = [\mathbf{y}(f,n,1,j),\ldots,\mathbf{y}(f,n,I,j)]^\top$ is the $I \times 1$ vector gathering the $I$ coefficients of the STFTs of the image of source $j$ into the mixtures at TF bin $(f,n)$.

Waveforms of the corresponding signals can be recovered efficiently through overlap-add procedures. All STFTs considered are assumed to have the same number $F$ of frequency bins and the same number $N$ of frames.

### 2.2. Models

#### 2.2.1. Sources model

In the STFT domain and for each source $j$, the observed signal $\boldsymbol{s}(f,n,\cdot,j)$ is supposed to be the realization of an underlying stochastic process $s(f,n,\cdot,j)$. In this study, we will simply assume that all TF bins $(f,n)$ are independent and Gaussian. Such an assumption is classical in the source separation community. As we have demonstrated in [11], this assumption amounts to consider that all the frames are independent and that within each frame, the signals are stationary and Gaussian, which is often a good approximation for audio signals.

Skipping the details that will be presented in a longer study, the model amounts to assuming that :

$$\forall j \in \mathcal{S}_p \quad \boldsymbol{s}(f,n,j) \quad \sim \mathcal{N}_c(0, P(f,n,j)) \tag{1}$$

$$\forall j \in \mathcal{S}_d \quad \boldsymbol{s}(f,n,\cdot,j) \quad \sim \mathcal{N}_c\left(0, P(f,n,j)\, R_j^{obs}(f)\right), \tag{2}$$

where :

- $P(f,n,j) \geq 0$ is the Power Spectral Density (PSD) of source $j$ at bin $(f,n)$. Loosely speaking, it gives the power of that source at bin $(f,n)$. It is a nonnegative quantity.
- $\mathcal{N}_c(\mathbf{z} \mid 0, \sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|\mathbf{z}|^2}{\sigma^2}\right)$ is the complex centered Gaussian distribution of variance $\sigma^2$.
- For multichannel sources, $R_j^{obs}(f)$ is a $I \times I$ positive definite *observation spatial covariance matrix*.

Equation (1) simply means that the different STFT coefficients for one given mono source $j \in \mathcal{S}_p$ are independent and distributed with respect to a complex and centered Gaussian distribution, whose variance is the PSD of the source at that bin.

In the case of multichannel sources ($j \in \mathcal{S}_d$), we assume through equation (2) that the values of the observed source signal in its different channels at TF bin $(f,n)$ are Gaussian and correlated, with a covariance given by equation (2). This model is reminiscent of the work by DUONG et al. in [2]. Basically, the covariance is given by the $R_j^{obs}(f)$ matrix and scaled according to the PSD $P(f,n,j)$ of the source at that bin.

Since the PSD $P(f,n,j)$ are the main parameters to be transmitted from the coder to the decoder, it is important to reduce the corresponding number of coefficients. As in [14], we propose two

1. $\mathbb{N}_J = [1,\ldots,J]$ is the set comprising the $J$ first strictly positive integers.

techniques to approximate $P(f,n,j)$. The first one is a Nonnegative Tensor Factorization model (NTF) :

$$\hat{P}(f,n,j) = \sum_{k=1}^{K} W_{fk} H_{nk} Q_{jk} \tag{3}$$

where $K \in \mathbb{N}$ is called *the number of components* and where $W$, $H$ and $Q$ are $F \times K$, $N \times K$ and $J \times K$ matrices, respectively. The main feature of this model in our context is to reduce the number of parameters required to encode $P$ from $FNJ$ to $(F+N+J)K$. The second source model we propose is based on image compression techniques such as JPEG [20]. Since $P(f,n,j)$ is a nonnegative quantity, $P$ can be seen as a set of $J$ images $P(\cdot,\cdot,j)$ of dimension $F \times N$ that can be compressed using dedicated techniques. Learning of the model parameters will be detailed in section 3.

#### 2.2.2. *Mixing model*

Following the work by DUONG *et al.* in [2], we adopt the *diffuse* mixing model (also called *full-rank* in the litterature) to account for the relation between the sources and their images. This model generalizes both the linear instantaneous and the convolutive mixing model and notably permits to acount for a *stochastic* dependance between the sources and their images instead of the deterministic relashionship assumed by convolutive or instantaneous mixing. The diffuse model is characterized by :

$$\forall j,\ \boldsymbol{y}(f,n,\cdot,j) \sim \mathcal{N}_c\left(0, \hat{P}(f,n,j) R_j(f)\right), \tag{4}$$

where $R_j(f)$ is the *image spatial covariance matrix* of source $j$ at frequency band $f$. It is a $I \times I$ positive definite matrix that encodes the covariances between the different channels of the *image* of source $j$ at frequency $f$. Whereas convolutive or instantaneous mixing boil down to assuming a rank-1 image spatial covariance matrix, this general formulation is significantly more general and permits to model sources that have a spatial spread, hence the "diffuse" adjective used here. It is possible to consider a spatial covariance matrix $R_j$ which is constant throughout the frequency indices [5].

## 3. ALGORITHMS

### 3.1. Coder

The first task of coder is to estimate good values for the PSD $P$ of the sources. Here, we consider both the NTF model (3) and Image Compression (IC). To this purpose, we propose to first estimate the real PSD $P(f,n,j)$ of the sources and then to approximate them using either NTF or IC.

Concerning mono sources, their PSD are easily estimated through maximum likelihood by the power spectrograms of the observations :

$$\forall j \in \mathcal{S}_p,\ P(f,n,j) = |\mathbf{s}(f,n,j)|^2 .$$

Concerning multichannel sources, their PSD is not so easily derived. However, if $R_j^{obs}(f)$ is available, $P(f,n,j)$ can be estimated from $\mathbf{s}(f,n,\cdot,j)$. Conversely, $R_j^{obs}(f)$ can be estimated if all $P(f,n,j)$ are available. This suggests an iterative procedure for the estimation of $P(f,n,j)$ which is summarized in algorithm 1. This algorithm converges in a few iterations.

When the PSD $P(f,n,j)$ are all estimated, the parameters for the NTF model are estimated using the classical algorithm 2 minimizing the Itakura-Saito divergence between $P$ and their model $\hat{P}$ (3). Once the NTF parameters have been learned, it can be shown [16] that an efficient way to encode them is to first use a logarithmic compressor and then to quantize uniformly $\log W$, $\log H$ and $\log Q$. Finally, encoding is done through Huffman encoding. The main parameter to control the bitrate in the NTF model is thus the number $K$ of components.

In the Image Compression (IC) model, encoding is achieved simply by applying an image compression algorithm such as JPG [20] on all $\{\log P(\cdot,\cdot,j)\}_j$. The bitrate in the case of IC is thus controlled by the quality parameter of the image compression algorithm considered.

**Algorithm 1** Estimation of PSD $P(f,n,j)$ for multichannel sources.

- Input : STFT $\boldsymbol{s}(f,n,i,j)$ of multichannel source $j \in \mathcal{S}_d$. $F \times N \times I$ tensor
- Initialization : set $P(f,n,j) = \frac{1}{I}\sum_{i=1}^{I} |\boldsymbol{s}(f,n,i,j)|^2$

**Repeat until convergence :**

- for each $f$, $R_j^{obs}(f) \leftarrow \frac{1}{N}\sum_{n=1}^{N} \frac{\mathbf{s}(f,n,\cdot,j)\boldsymbol{s}(f,n,\cdot,j)^H}{P(f,n,j)}$
- for each $(f,n)$,
  $P(f,n,j) \leftarrow \frac{1}{I}\mathbf{s}(f,n,\cdot,j)^H R_j^{obs}(f)^{-1}\mathbf{s}(f,n,\cdot,j)$

**Algorithm 2** Learning the NTF model from the PSD by minimization of the Itakura-Saito divergence between $\hat{P}$ and $P$. Exponentiation is understood element-wise and $a \cdot b$, $\frac{a}{b}$ denote element-wise multiplication and division. 100 iterations are usually sufficient.

- Inputs : PSD $P(f,n,j)$, $F \times N \times J$ tensor and $K \in \mathbb{N}$
- Initialization : set $W$, $H$ and $Q$ as random $F \times K$, $N \times K$ and $J \times K$ nonnegative matrices.

**Repeat :**

- $W \leftarrow W \cdot \frac{\sum_j \left(\hat{P}(\cdot,\cdot,j)^{-2} \cdot P(\cdot,\cdot,j)\right) H \mathrm{diag}(Q_{j\cdot})}{\sum_j \left(\hat{P}(\cdot,\cdot,j)^{-1}\right) H \mathrm{diag}(Q_{j\cdot})}$
- $H \leftarrow H \cdot \frac{\sum_j \left(\hat{P}(\cdot,\cdot,j)^{-2} \cdot P(\cdot,\cdot,j)\right)^\top W \mathrm{diag}(Q_{j\cdot})}{\sum_j \left(\hat{P}(\cdot,\cdot,j)^{-2}\right)^\top W \mathrm{diag}(Q_{j\cdot})}$
- $Q_{j\cdot} \leftarrow \mathrm{diag}\left(\mathrm{diag}(Q_{j\cdot}) \cdot \frac{W^\top \left(\hat{P}(\cdot,\cdot,j)^{-2} \cdot P(\cdot,\cdot,j)\right) H}{W^\top \left(\hat{P}(\cdot,\cdot,j)^{-1}\right) H}\right)$

#### 3.1.1. *Decoder*

At the decoder, the side-information is recovered and decoded. This permits to obtain the PSD model $\hat{P}(f,n,j)$, either through (3) for NTF or through image reconstruction for IC. At this stage, the STFTs $\mathbf{x}$ of the mixtures are known. What is left to learn are the parameters $R_j(f)$ for the mixing model[2]. Those parameters can be learned efficiently through the Expectation-Maximization algorithm , already presented in [2], with the noticeable difference that the PSDs $\hat{P}$ are assumed known and fixed here, which leads in practice to fast convergence. Only a few iterations of algorithm 3 are usually sufficient. The fact that the mixing parameters need not be transmitted is particularly noticeable : they can be automatically estimated at the decoder. Still, they can be computed at the coder and transmitted along in the side-information for computational efficiency at the decoder.

When the mixing parameters $R_j(f)$ have been estimated, the

2. In case of a constant spatial covariance matrix $R_j$, these parameters reduce to $J$ matrices of dimension $I \times I$.

**Algorithm 3** Estimation of the mixing parameters $R_j(f)$ given $\mathbf{x}$ and $\hat{P}$.

– Inputs STFT $\boldsymbol{x}(f,n,i)$ of the mixtures, estimated PSD $\hat{P}(f,n,j)$
– Initialization : define all $R_j(f)$ as diagonal $I \times I$ matrices

**Repeat :**

– *Expectation step* : for each $(f,n,j)$ :

1. $K_{xx} = \sum_{j=1}^{J} \hat{P}(f,n,j) R_j(f)$
2. $G_j = \hat{P}(f,n,j) R_j(f) K_{xx}^{-1}$
3. $\widehat{\boldsymbol{y}}(f,n,\cdot,j) = G_j \boldsymbol{x}(f,n,\cdot)$
4. $\widehat{K}_{yy}(f,n,j) = \widehat{\boldsymbol{y}}(f,n,\cdot,j)\widehat{\boldsymbol{y}}(f,n,\cdot,j)^H + (I_I - G_j P(f,n,j)) R_j(f)$

– *Maximization-step* :

$$\begin{cases} \forall j, R_j \leftarrow \frac{1}{NF}\sum_{n,f} \frac{\widehat{K}_{yy}(f,n,j)}{\hat{P}(f,n,j)} & \text{if } R_j(f) \text{is constant} \\ \forall (f,j), R_j(f) \leftarrow \frac{1}{N}\sum_{n=1}^{N} \frac{\widehat{K}_{yy}(f,n,j)}{\hat{P}(f,n,j)} & \text{otherwise} \end{cases}$$

images can be very simply estimated through minimum mean-squared error minimization by WIENER filtering [2] :

$$\widehat{\boldsymbol{y}}(f,n,\cdot,j) = \hat{P}(f,n,j) R_j(f) \left[ \sum_{j'=1}^{J} \hat{P}(f,n,j') R_{j'}(f) \right]^{-1} \boldsymbol{x}(f,n,\cdot) \quad (5)$$

If the original mono sources are to be recovered instead of the images, they can be estimated through a beamforming strategy as :

$$\hat{\mathbf{s}}(f,n,\cdot,j) = U_j(f)\widehat{\boldsymbol{y}}(f,n,\cdot,j)$$

where $U_j(f)$ is a $1 \times I$ vector if $j \in \mathcal{S}_p$ and a $I \times I$ matrix if $j \in \mathcal{S}_d$. Since the coder is able to compute the estimated images (5), it can also compute the $U_j(f)$ that minimize the mean-squared error between $U_j(f)\widehat{\boldsymbol{y}}(f,n,\cdot,j)$ and $\mathbf{s}(f,n,\cdot,j)$ and send it as additional side-information.

## 4. EVALUATION

We have performed an extensive evaluation of the proposed demixing method on the QUASI database [3], which is composed of 12 full-length songs sampled at 44.1kHz, along with all their constitutive tracks. For each song, several mixtures are available, as obtained by a professional sound engineer. The simplest mixture consists of a mere panning for the sources (*bal-pan* mix), while the most complex involves dynamic compressions and audio effects (*comp-fx* mix). For all of the excerpts, we have considered the first minute only and performed an encoding of the sources using both the NTF and the IC model, and we have tested the performance of the method for both the bal-pan and the comp-fx mixes. The metric we used is the Perceptual Similarity Measure (PSM) from PEMO-Q [9], which provides a measure of the perceptual similarity between the original tracks and their estimates. PSM lies between 0 (mediocre) and 1 (identical). Results can be found in figure 2 and some audio examples can be listened to on the webpage dedicated to this paper [4].

3. www.tsi.telecom-paristech.fr/aao/?p=605
4. www.tsi.telecom-paristech.fr/aao/lien

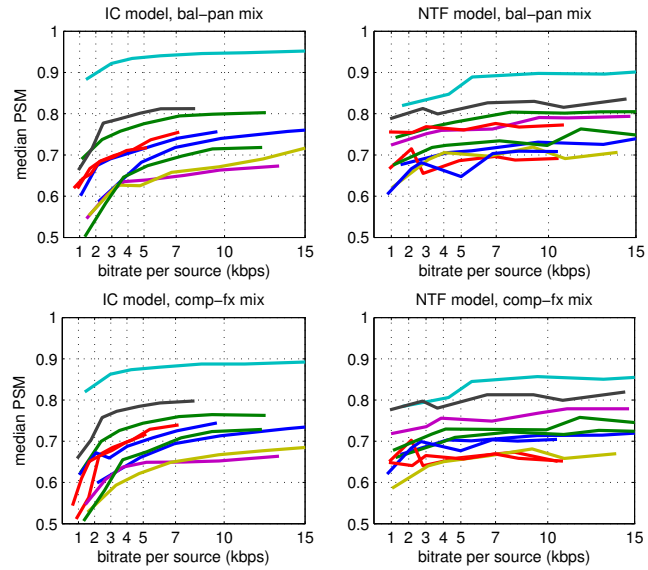


**Fig. 2**. Perceptual similarity between the original images and their estimates for both the NTF and IC models and different mixing conditions. Each line stands for a different excerpt.

As can be seen on figure 2 and listened to online, the proposed technique for ISS permits to reach good performance at very low bitrates. At 1kbps per source, performance is already remarkably good and sufficient for applications that do not require high fidelity. Perceptual similarity gets higher at 5kbps and artifacts get marginal. Performance of the NTF model at very low bitrates are seen as slightly higher than those of IC, at the cost of a higher computational complexity. Finally, both linear instantaneous and professional mixtures are seen to be well supported.

The proposed technique can thus be used for broad audience entertainment applications that require a good quality at low bitrates. For applications that come with a very high-fidelity constraint, parametric ISS as presented here suffers from bounds on achievable performance. This limitation can be overcome using CISS [13, 16], which is based on source coding and is an extension of the ideas presented here.

## 5. CONCLUSION

In this paper, we have proposed a general Gaussian framework for the informed demixing of real-world multichannel mixtures and we have detailed all the corresponding algorithms. The proposed method has several interesting features. First, the source models considered are particularly compact, leading to bitrates as low as $1-10$kbps/source. Second, the powerful diffuse model used to account for the mixing process permits to handle realistic professional mixtures as opposed to the classical linear instantaneous or convolutive models. Third, the mixing parameters are estimated automatically at the decoder and need not be transmitted, leading to lower bitrates. Finally, the observed sources at the decoder can be either mono or multichannel and the signals to recover at the decoder may be either the signals observed at the coder or their images within the mixtures.

## 6. REFERENCES


[1] J. Breebaart, S. van de Par, and A. Kohlrausch. High-quality parametric spatial audio coding at low bit rates. In *AES 116th convention*, Berlin, Germany, May 2004.

[2] N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1830 –1840, sept. 2010.

[3] J. Engdegård, C. Falch, O. Hellmuth, J. Herre, J. Hilpert, A. Hölzer, J. Koppens, H. Mundt, H.O. Oh, H. Purnhagen, B. Resch, L. Terentiev, M.L. Valero, and L. Villemoes. MPEG spatial audio object coding - the ISO/MPEG standard for efficient coding of interactive audio scenes. In *Audio Engineering Society Convention 129*, 11 2010.

[4] C. Falch, L. Terentiev, and J. Herre. Spatial audio object coding with enhanced audio object separation. In *13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010.

[5] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency gaussian models. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 78–81, Mohonk, NY, USA, Oct. 2005.

[6] S. Gorlow and S. Marchand. Informed source separation: Underdetermined source signal recovery from an instantaneous stereo mixture. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 309 –312, October 2011.

[7] J. Herre. From joint stereo to spatial audio coding. In *In Proc. Digital Audio Effects Workshop (DAFx*, Naples, Italy, October 2004.

[8] J. Herre and S. Disch. New concepts in parametric coding of spatial audio: From SAC to SAOC. In *IEEE International Conference on Multimedia and Expo (ICME 2007)*, pages 1894–1897, Beijing, China, July 2007.

[9] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902 –1911, November 2006.

[10] A. Liutkus, R. Badeau, and G. Richard. Informed source separation using latent components. In *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, St Malo, France, 2010.

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

[12] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchang, and G. Richard. Informed source separation : a comparative study. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, August 2012.

[13] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard. Spatial coding-based informed source separation. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, August 2012.

[14] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8):1937 – 1949, 2012.

[15] J. Nikunen, T. Virtanen, and M. Vilermo. Multichannel audio upmixing based on non-negative tensor factorization representation. In *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2011.

[16] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Coding-based informed source separation: Nonnegative tensor factorization approach. *IEEE Trans. on Audio, Speech and Language Processing*, 2012.

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

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

[19] M. Parvaix, L. Girin, L. Daudet, J. Pinel, and C. Baras. Hybrid coding/indexing strategy for informed source separation of linear instantaneous under-determined audio mixtures. In *Proceedings of 20th International Congress on Acoustics*, Sydney, Australia, Aug. 2010.

[20] G.K. Wallace. The JPEG still picture compression standard. *Commun. ACM*, 34:30–44, April 1991.