Coding-based Informed Source Separation: Nonnegative Tensor Factorization Approach

Alexey Ozerov, Antoine Liutkus, Roland Badeau, Gael Richard

Coding-based Informed Source Separation: Nonnegative Tensor Factorization Approach

Séparation de Sources Informée de type Codage: Méthode de Factorisation Tensorielle Non-négative

Alexey Ozerov, Antoine Liutkus, Roland Badeau, and Gaël Richard

Abstract

Informed source separation (ISS) aims at reliably recovering sources from a mixture. To this purpose, it relies on the assumption that the original sources are available during an *encoding* stage. Given both sources and mixture, a *side-information* may be computed and transmitted along with the mixture, whereas the original sources are forgotten. During a *decoding* stage, both mixture and side-information are processed to recover the sources. Most ISS techniques proposed so far rely on a source separation strategy and cannot achieve better results than oracle estimators. In this study, we introduce Coding-based ISS (CISS) and draw the connection between ISS and source coding. CISS amounts to encode the sources using not only a model as in source coding but also the observation of the mixture. This strategy has several advantages. First, it can reach any quality, provided sufficient bandwidth is available as in source coding. Second, it makes use of the mixture in order to reduce the bitrate required to transmit the sources, as in classical ISS. Furthermore, we introduce Nonnegative Tensor Factorization as a very efficient model for CISS and report rate-distortion results that strongly outperform the state of the art.

Index Terms

Informed source separation, spatial audio object coding, source coding, constrained entropy quantization, probabilistic model, nonnegative tensor factorization.

Résumé

La séparation de sources informée (ISS) vise à extraire de manière fiable des sources à partir d'un mélange. Pour atteindre cet objectif, elle exploite l'hypothèse que les sources originales sont disponibles lors d'une étape dite d'*encodage*. Connaissant les sources et le mélange, une *information annexe* peut alors être calculée et transmise conjointement avec le mélange, tandis que les sources originales sont oubliées. Dans une seconde étape de *décodage*, le mélange et l'information annexe sont traités conjointement afin d'extraire les sources. La plupart des techniques d'ISS proposées jusqu'à présent s'appuient sur une stratégie de séparation de sources et leurs performances sont donc limitées par celles des estimateurs oracle. Dans cette étude, nous introduisons l'ISS de type codage (CISS) et nous établissons le lien entre ISS et codage de source. L'approche CISS consiste à encoder les sources en utilisant non seulement un modèle comme en codage de source, mais également l'observation du mélange. Cette stratégie présente plusieurs avantages. Premièrement, elle permet d'atteindre n'importe quel niveau de qualité, pourvu qu'une largeur de bande suffisante soit disponible, comme en codage de source. Deuxièmement, elle prend en compte le mélange afin de réduire le débit requis pour transmettre les sources, comme en ISS classique. Enfin, nous montrons que la Factorisation Tensorielle Non-négative s'avère être un modèle très efficace de CISS et nous présentons des résultats en terme de débit-distorsion qui surpassent largement l'état de l'art.

Mots clés

Séparation de sources informée, codage spatial d'objets audio, codage de source, quantification entropique contrainte, modèle probabiliste, factorisation tensorielle non-négative.

Alexey Ozerov is with Technicolor Research & Innovation, 1, avenue de la Belle Fontaine F-35576 Cesson Sévigné, France, e-mail: alexey.ozerov@technicolor.com.

Antoine Liutkus, Roland Badeau, and Gaël Richard are with Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, 37-39, rue Dareau, 75014 Paris, France, email: firstname.lastname@telecom-paristech.fr.

This work was partly supported by the European Commission under contract FP7-ICT-287723 - REVERIE and by the ANR through the DReaM project (ANR-09-CORD-006-03).

I. INTRODUCTION

AUDIO compression has been a very active field of research for several decades due to the tremendous demand for transmitting, or storing, digital audio signals at reduced rates. Audio compression can either be *lossless* (the original signal can be exactly recovered) or *lossy* (the original signal can only be approximately recovered). The latter scheme which reaches much higher compression ratios usually exploits psychoacoustic principles to minimize the perceptual loss. A large variety of methods were developed amongst which some have been standardized. MPEG1-Layer 3 (e.g. mp3) [1] or Advanced Audio Coding (AAC) [2] are probably amongst the most widely popular standardized lossy audio compression schemes. It is generally admitted that most coding schemes either rely on a parameterized signal model (e.g. as in *parametric coding approaches*), or on a direct quantization of the signal (as in *waveform* or *transform coding*), but also in some cases on a combination of both [3].

Concurrently, the domain of source separation (and audio source separation in particular) has also seen a great interest from the community but with little or no interaction with the audio compression sphere [4]. The general problem of source separation can be described as follows: assume J signals (*the sources*) $\mathbf{S}$ have been mixed through I channels to produce I signals (*the mixtures*) $\mathbf{X}$. The goal of source separation is to estimate the sources $\mathbf{S}$ given their mixtures $\mathbf{X}$. Many advances were recently made in the area of audio source separation [5], [6]. However, the problem remains challenging in the undetermined setting ($I < J$), including the single-channel case ($I = 1$), and for convolutive mixtures [7].

It is now quite clear that audio source separation performances strongly depend on the amount of available prior information about the sources and the mixing process one can introduce in the source separation algorithm. In unsupervised source separation, this information can be under the form of a specific source model (as for example the source/filter model used in [8] for singing voice separation or more generally a composite model from a library of models [6]). However, this information can also be provided by a user [9], [10] or by a partial transcription in the case of music signals (see for example [11]). In the extreme case, this information can be the sources themselves. In these cases, we refer to *informed source separation (ISS)*.

Such so-called ISS schemes were recently developed for the case where both the sources and the mixtures are assumed known during an *encoding* stage [12]–[15]. This knowledge enables the computation of any kind of *side-information* that should be small and should help the source separation at the *decoding* stage, where the sources are no longer assumed to be known. The side-information can be either embedded into the mixtures using watermarking methods [14] or just kept aside. Note that the performances of *source separation* and the above-mentioned *conventional ISS* methods, depending on the underlying models and assumptions, are bounded by those of oracle estimators [16].

Indeed, since the majority of conventional ISS methods [13], [14] are source separation-inspired and thus fall into the category of parametric coding approaches, their best (the minimal) achievable distortion is incompressible, i.e., it is bounded below[1]. In order to outperform the oracle estimators, some hybrid approaches have been developed, which involve waveform source coding. In [15], some sources are encoded using a source coding method and the remaining sources are recovered by a conventional ISS method. However, such a straightforward hybridization does not allow to overcome the above-mentioned drawbacks that are still valid for individual sources.

With regard to the above description, it is quite clear that ISS shares many similarities with the recently introduced Spatial Audio Object Coding (SAOC) (see [17]–[19] and [20] for the ISO/MPEG SAOC standard version). Developed as a multichannel audio compression scheme, SAOC also aims at recovering so called *sound objects* at the decoding side from a transmitted downmix signal and side information about the audio objects. In the literature, different kinds of side information were also considered in the framework of Spatial Audio Coding (SAC), such as the inter and intra-channel correlation [21], spatial coherence cues [22], source localization parameters [23], or a sinusoids plus noise model of the sources [24]. In SAOC [20], high quality remixing is guaranteed by also encoding and transmitting residual signals resulting from an imperfect object extraction at the encoding side (therefore jointly exploiting waveform coding and parametric coding principles). However, this scheme has a major drawback which limits its potential. Indeed, in SAOC the "separation step" (sound object extraction) is independent of the "residual compression step" while this could be done jointly.

The purpose of this paper is then:

[1]This remark does not concern [12], where the distortion can be always decreased by increasing the size of the corresponding molecular dictionary, which would lead, however, to an excessive rate needed to transmit such a dictionary.

1) to further develop and to present in an even more general manner the novel concept of *Coding-based ISS (CISS)* recently introduced in [25], [26] and to highlight its main theoretic advantage against the approaches followed in both conventional ISS [13], [14] and SAOC [20];
2) to extend the previous "proof of concept" model used in [25] by integrating a more elaborate model based on Non-Negative Tensor Factorization (NTF). [2] We also discuss how the proposed approach relates to other relevant state of the art methods such as non-negative matrix factorization (NMF) or NTF-based coding methods [27], [28], but to the best of our knowledge this is the first attempt of using NTF models with waveform coding principles.
3) and to show that the proposed scheme allows for a smooth transition between low rate object-based parametric coding and high-rate waveform coding relying on the same object-based model, thus exploiting long-term redundancy.

It is also important to underline that although our model is presented in the ISS framework, it is directly applicable to traditional audio coding or multichannel audio coding (that is without assuming the mixture to be known at the decoder side).

The paper is organized as follows: Section II introduces the general concept of CISS and thoroughly discusses its relation to the state of the art. Then, its particular variant based on NTF (CISS-NTF) is described in details and analyzed in section III in the case of single-channel mixtures with the Mean Squared Error (MSE) criterion for optimisation.

Experimental results are presented in section IV and the conclusions and perspectives are drawn in the final section.

II. Coding-based Informed Source Separation

The general probabilistic framework introduced herein for ISS is called coding-based ISS (CISS). This approach consists in quantizing the sources, as in waveform source coding, while using the *a posteriori* source distribution, given the mixture and some generative probabilistic source model, as in source separation. The quantization can be performed by optimizing the MSE or some perceptually-motivated distortion driven by a perceptual model. In this section the framework is presented in a very general manner, i.e., it is not limited to a particular problem dimensionality (e.g., multichannel or single-channel mixtures), mixing type (e.g., linear instantaneous or convolutive mixture), source model or perceptual model. A particular instance of the framework will be described in the following section III and evaluated in section IV.

Fig. 1 and 2 give very high-level presentations of the state of the art approaches, notably the conventional IIS approaches [13], [14] and the SAOC [17]–[20], where all audio objects are enhanced.[3] In the conventional ISS approaches (Fig. 1), at the encoding stage, a source model parameterized by $\hat{\theta}$ is estimated, given the sources $\mathbf{S}$ and the mixtures $\mathbf{X}$. It is then encoded and transmitted as a side-information yielding its quantized version $\bar{\theta}$. At the decoding stage, the model parameter $\bar{\theta}$ is reconstructed, and the sources $\hat{\mathbf{S}}$ are reconstructed in turn, given $\bar{\theta}$ and the mixture $\mathbf{X}$ (e.g., by Wiener filtering, as in [13]). However, as mentioned in the introduction, the best achievable distortion of such parametric coding approaches is inherently limited.

At a very high level view, the parametric coding part of SAOC approaches (Fig. 2) follows exactly the same scheme as the conventional IIS (Fig. 1), except that the parametric model, called *SAOC parameters*, is different. To achieve a higher quality at the expense of a higher transmission rate, the residuals of the parametric SAOC reconstruction can be encoded using a perceptual waveform coder. However, as we see in Fig. 2, the parametric and waveform coding steps are performed independently and using different models. This is suboptimal since there is no evidence that the residual encoding should be independent of the parametric source encoding.

Fig. 3 gives a high-level representation of the proposed CISS approach. At the encoding stage, the model parameter $\hat{\theta}$ specifying the posterior distribution $p(\mathbf{S}|\mathbf{X},\hat{\theta})$ from a particular family of distributions is estimated, given the sources $\mathbf{S}$ and the mixtures $\mathbf{X}$. A perceptual model $\mathbf{\Omega}$ can be optionally computed as well. $\hat{\theta}$ and $\mathbf{\Omega}$ are then jointly encoded and transmitted as a side-information yielding their quantized versions $\bar{\theta}$ and $\bar{\mathbf{\Omega}}$. This encoding can optionally use the knowledge of the mixtures $\mathbf{X}$. Finally, using the posterior $p(\mathbf{S}|\mathbf{X},\bar{\theta})$ and a perceptual distortion

[2] While an NTF model for CISS was already considered in a short study [26] in the multichannel case, here we consider the single-channel case and conduct a more thorough evaluation. Moreover, we provide some theoretical support to the results that were used in [25], [26].

[3] Within this paper, if the contrary is not stated, we always consider SAOC with enhanced audio objects as in [19], [20].

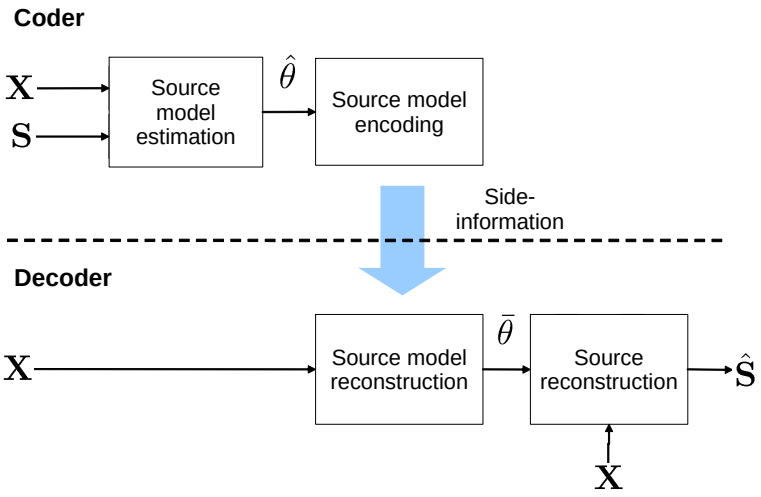

Fig. 1. High level presentation of the conventional ISS [13], [14].

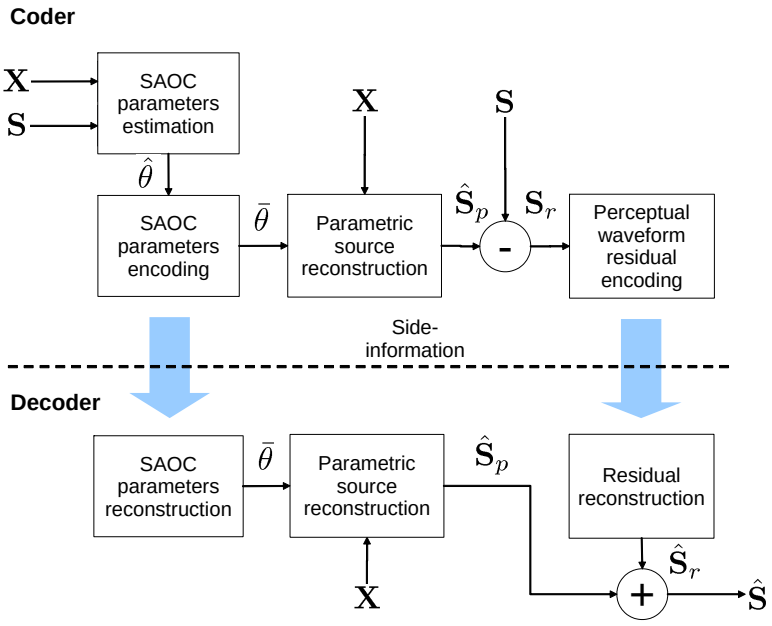

Fig. 2. High level presentation of SAOC (all objects are enhanced) [17]–[20].

measure driven by $\bar{\mathbf{\Omega}}$ the sources $\mathbf{S}$ are encoded and transmitted as a side-information. More precisely, we use to that aim probabilistic model-based quantization under high-rate theory assumptions (see, e.g., [29], [30]). At the decoding stage, the quantized parameters $\bar{\theta}$ and $\bar{\mathbf{\Omega}}$, and then the quantized sources $\hat{\mathbf{S}}$ are reconstructed.

Thus, in contrast to the conventional ISS methods, the CISS framework allows the distortion being unbounded below as in waveform source coding (see Fig. 3 vs. Fig. 1). Moreover, in contrast to SAOC, CISS permits, as we will see below, to use more advanced source models that better exploit the redundancy of audio signals, and to use the knowledge of the mixture and model parameters to encode the residuals (see Fig. 3 vs. Fig. 2).

In this work we propose a particular instance of the general CISS framework, referred herein as *CISS-NTF*, that is based on an (object-based) probabilistic NTF source model. Moreover, CISS-NTF is designed for the single-channel case and for the MSE distortion criterion. Investigation of distortions driven by more advanced perceptual models (e.g., those considered in [31]–[33]) is left for a further study.

The major differences of the proposed CISS-NTF approach compared to the state of the art can then be highlighted

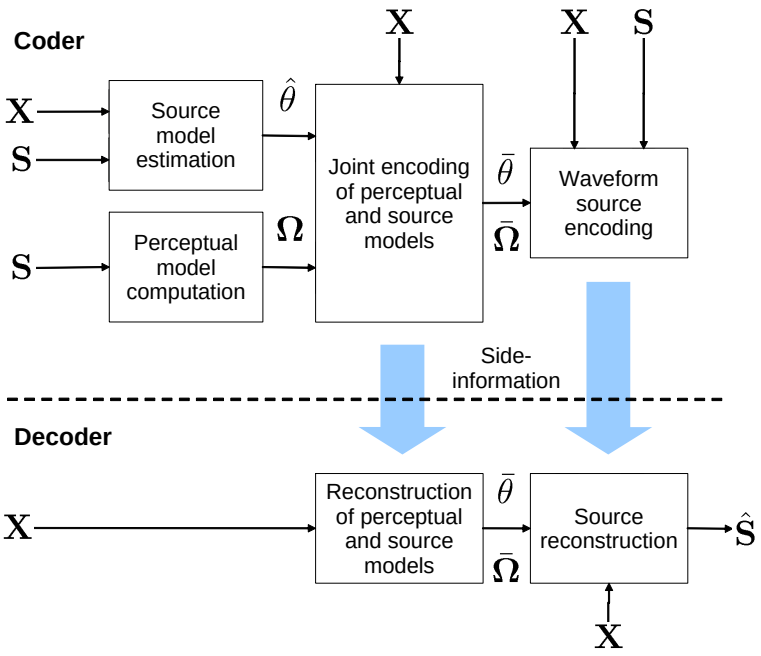

Fig. 3. High level presentation of CISS (proposed).

as follows:

- In contrast to conventional ISS methods [12]–[14], it is based on waveform coding, thus leading to much superior quality for moderate and high rates, as it was already mentioned for CISS in general.
- In contrast to SAOC [20], based on some local parameters, it exploits advanced source models, i.e., NTF. First, this allows using long-term redundancy of audio signals for coding. Second, the parameters used for parametric coding (as in the earlier version of SAOC [18]) and those used for waveform coding (as in [20]) are all computed from the NTF source model and the mixture. Thus, these parameters are coupled (or jointly encoded), while in SAOC [20] they are encoded separately. Moreover, the proposed method exploits posterior correlations between sources (given the mixture), while in SAOC the residuals of the enhanced audio objects are encoded independently.
- In the NMF / NTF-based methods [27], [28] the signal short-time Fourier transform (STFT) (a redundant signal representation) amplitudes are encoded by approximating them with an NMF / NTF decomposition. In [27] the STFT phase is then entropy encoded and the rate (between phase and amplitude encoding) is allocated empirically. Also, the rate between different NMF / NTF model parameters is empirically allocated.
  Besides the fact that we consider a different coding problem, the proposed approach has the following possible advantages over [27], [28]. First, we consider a probabilistic NTF applied to the modified discrete cosine transform (MDCT) or STFT of the sources. As such, we do not split amplitude and phase, but encode them jointly within the corresponding time-frequency representation, while minimizing a target distortion under the constrained entropy. Thus, we consider our approach as a waveform coding-based within the NTF framework. Second, our probabilistic NTF formulation and quantization under high-rate theory assumptions, allows us deriving (under some approximations) analytical expressions for rate allocation between different NTF model parameters which allows avoiding time-consuming empirical parameter optimization. Third, MDCT being a critically sampled signal representation, we show its great advantage over redundant STFT within this application. To our best knowledge NMF / NTF models were not so far applied to MDCT signal representations for compression purposes.

III. Single-channel CISS-NTF with MSE

In this section, we investigate the proposed approach in the case of single-channel mixtures ($I = 1$) using the NTF source model and MSE distortion criterion.

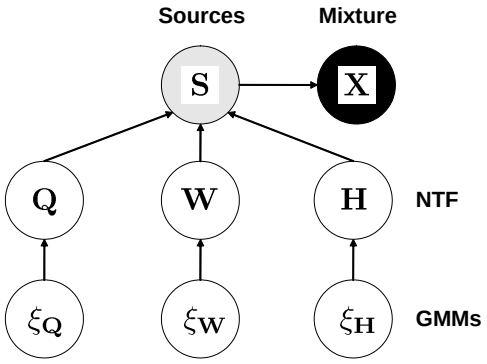

Fig. 4. High-level graphical representation of CISS-NTF probabilistic hierarchical modeling. Shadings of nodes: variables observed at both coder and decoder sides (black), variables observed at the coder side, quantized and transmitted (gray), and parameters estimated at the coder side, quantized and transmitted (white).

All signals are represented in a real-valued or complex-valued (here, respectively, MDCT or STFT) time-frequency domain. In the time-frequency domain the mixing equation writes

$$x_{fn} = \sum\nolimits_{j=1}^{J} s_{jfn} + b_{fn}, \tag{1}$$

where $j = 1, \ldots, J$, $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote, respectively, the source index, the frequency index and the time-frame index; and x_{fn}, s_{jfn} and b_{fn} denote, respectively, the time-frequency coefficients of the mixture, of the sources and of an additive noise. Depending on the particular configuration this additive noise can represent any combination of the following distortions:

1) a background or recording noise, if $\mathbf{X} = \{x_{fn}\}_{f,n}$ is an unquantized mixture of sources $\mathbf{S}_j = \{s_{jfn}\}_{f,n}$ $(j = 1, \ldots, J)$,
2) a quantization noise if $\mathbf{X}$ is a quantized version of its clean version $\mathbf{X}^{\text{clean}}$, i.e., $b_{fn} = x_{fn} - x_{fn}^{\text{clean}}$ (e.g., as in SAOC [18], [19]),
3) additional sources $\{\mathbf{S}_j\}_{j=J+1}^{J^*}$ if one is only interested to encode J sources among J^* $(J < J^*)$ sources in the mixture.

Fig. 4 gives a high-level graphical representation of CISS-NTF probabilistic hierarchical modeling described in details below. It includes mixture, sources, NTF parameters (Sec. III-A) and Gaussian mixture models (GMM) used to encode these parameters (Sec. III-D2c).

A. NTF source model

As a source model we use the NTF model described in [10] and [14]. Its main idea is to assume that the spectrograms of the sources can be considered as the activation over time of some *spectral templates*. To avoid a pre-defined choice for the number of spectral templates for each source, a refinement of the model is to consider a common *pool* of spectral templates jointly approximating all spectrograms of the sources. Such a strategy permits to reduce the number of parameters of the model and to share the same templates for several sources, which may be of interest when there is some kind of redundancy among sources.

Formally, the NTF model can be described as follows. First, the source and noise time-frequency coefficients s_{jfn} and b_{fn} are assumed mutually independent, i.e., over j, f and n, and distributed as follows:

$$s_{jfn} \sim \mathcal{N}_{r/c}(0, v_{jfn}), \qquad b_{fn} \sim \mathcal{N}_{r/c}(0, \sigma_{b,fn}^2), \tag{2}$$

where the distribution $\mathcal{N}_{r/c}(\cdot, \cdot)$ is the standard Gaussian distribution if s_{jfn} is real-valued, or the circular complex Gaussian distribution if it is complex-valued. The source variances v_{jfn} are structured as

$$v_{jfn} = \sum\nolimits_{k=1}^{K} q_{jk} w_{fk} h_{nk}, \tag{3}$$

with $q_{jk}, w_{fk}, h_{nk} \geq 0$ and the noise variances $\sigma^2_{b,fn}$ are assumed to be known. The noise variances can be either constant and fixed ($\sigma^2_{b,fn} = \sigma^2_b$) to represent a background noise or have a structure similar to that of the source variances to represent a nonstationary noise.

We here assume the noise variances to be constant and fixed. This model can be parameterized as follows

$$\theta = \left\{\mathbf{Q}, \mathbf{W}, \mathbf{H}, \sigma^2_b\right\}, \tag{4}$$

with $\mathbf{Q} = \{q_{jk}\}_{j,k}$, $\mathbf{W} = \{w_{fk}\}_{f,k}$ and $\mathbf{H} = \{h_{nk}\}_{n,k}$ being, respectively, $J \times K$, $F \times K$ and $N \times K$ nonnegative matrices (see Fig. 4).

This model is in fact an object-based approximation of the 3-valence tensor of source power spectra

$$\mathbf{P} \triangleq \{p_{jfn}\}_{j,f,n} \quad (p_{jfn} \triangleq |s_{jfn}|^2) \tag{5}$$

consisting of K objects (rank-1 tensors) that represent individual sounds. Whereas each column of $\mathbf{W}$ stands for one spectral template, its activation over time is given by the corresponding column of $\mathbf{H}$. Finally, the columns of $\mathbf{Q}$ model the possible couplings between the spectral templates, i.e., different sources can share the same templates.

An illustrative example of this NTF modeling is given in Fig. 5, where the first raw shows MDCT power spectrograms p_{jfn} (Eq. (5)) of three sources (drums, guitar and singing voice), the second raw shows their structured approximations v_{jfn} (Eq. (3)), and the third raw includes NTF matrices $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$. First, by investigating matrix $\mathbf{Q}$ one can note that among the $K = 9$ components (in average 3 components per source) 7 components were automatically assigned (dark brown color) to each source, while sharing the 6-th component between drums and voice and sharing the 9-th component between all three sources. This last component can be interpreted as the background noise floor that is common to all three sources. Second, one can note that while this is a good approximation (MDCT power spectrograms and their structured approximations look very similar), it drastically reduces the dimensionality (i.e., the number of parameters to be transmitted). Indeed, for this example, instead of $J \times F \times N = 3 \times 1024 \times 421 = 1293312$ coefficients p_{jfn}, one have only $(J + F + N) \times K = (3 + 1024 + 421) \times 9 = 13032$ entries of NTF matrices, which divides the number of parameters by 100.

B. Prior and posterior distributions

Since the source time-frequency coefficients are modeled as distributed with respect to independent Gaussian distributions, the additive noise is as well assumed Gaussian, and the mixing (1) is linear, the posterior distribution of the sources given the observed mixture is Gaussian, and analytical expression of this distribution is readily obtained. Let $\mathbf{s}_{fn} = [s_{1fn}, \ldots, s_{Jfn}]^T$ be the vector containing the time-frequency coefficients of all sources at bin (f, n). Provided all parameters (4) are available, the prior and posterior distributions of $\mathbf{s}_{fn}$ write, respectively, as [6]

$$p(\mathbf{s}_{fn}|\theta) = N_{r/c}\left(\mathbf{s}_{fn}; \boldsymbol{\mu}^{\mathrm{pr}}_{fn}, \boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}\right), \tag{6}$$

$$p(\mathbf{s}_{fn}|x_{fn};\theta) = N_{r/c}\left(\mathbf{s}_{fn}; \boldsymbol{\mu}^{\mathrm{pst}}_{fn}, \boldsymbol{\Sigma}^{\mathrm{pst}}_{\mathbf{s},fn}\right), \tag{7}$$

where $N_{r/c}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the probability density function (pdf) of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ for either real-valued or complex-valued cases; and

$$\boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn} = \mathrm{diag}\left[\{v_{jfn}\}_j\right], \qquad \boldsymbol{\mu}^{\mathrm{pr}}_{fn} = 0, \tag{8}$$

$$\boldsymbol{\Sigma}^{\mathrm{pst}}_{\mathbf{s},fn} = \left(\mathbf{I}_J - \mathbf{g}_{fn}\mathbf{1}_J\right)\boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}, \tag{9}$$

$$\boldsymbol{\mu}^{\mathrm{pst}}_{fn} = \mathbf{g}_{fn}x_{fn}, \tag{10}$$

$$\mathbf{g}_{fn} = \boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}\mathbf{1}^T_J\left(\mathbf{1}_J\boldsymbol{\Sigma}^{\mathrm{pr}}_{\mathbf{s},fn}\mathbf{1}^T_J + \sigma^2_b\right)^{-1}, \tag{11}$$

with $\mathbf{I}_J$ and $\mathbf{1}_J$ denoting, respectively, the $J \times J$ identity matrix and the J-length row vector of ones.

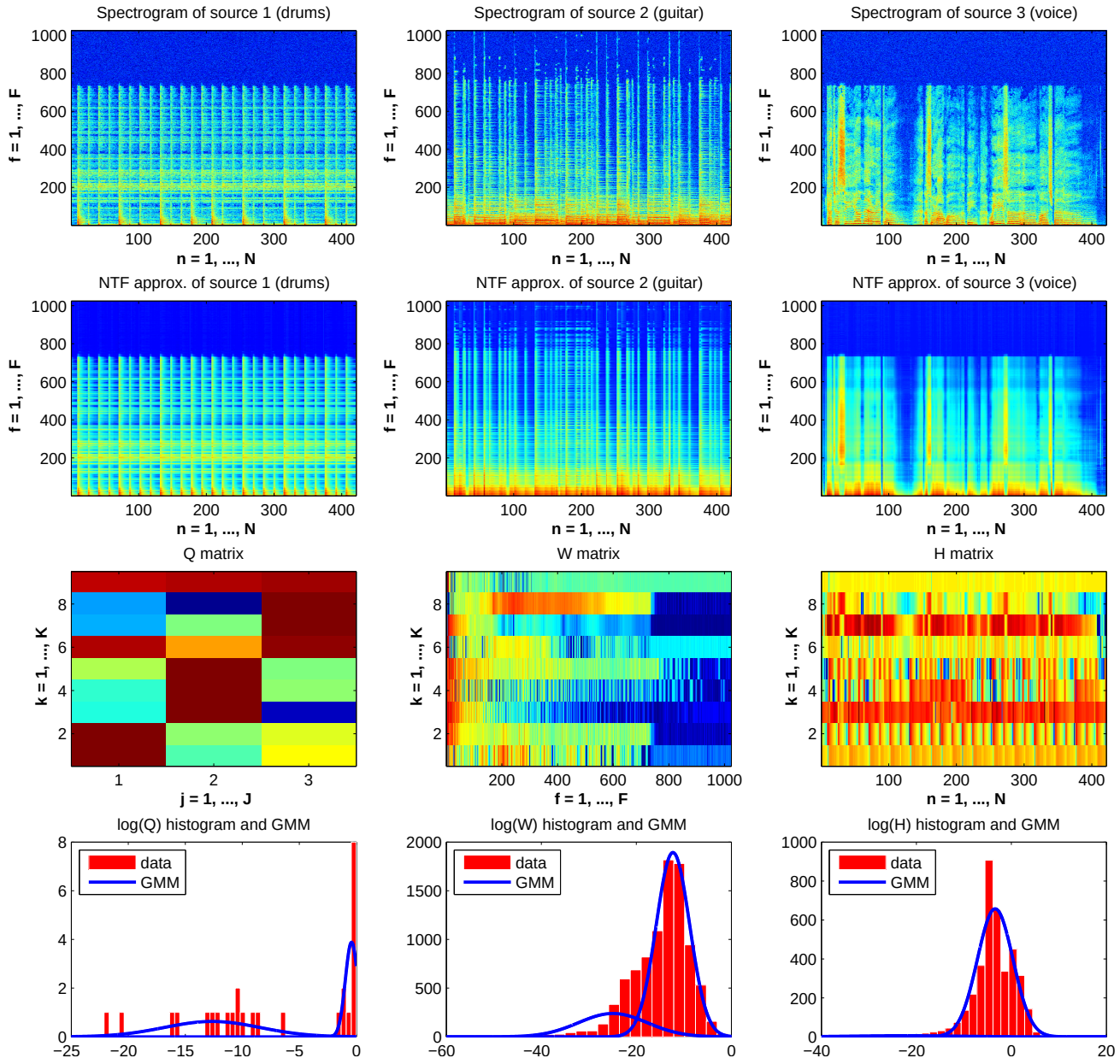

Fig. 5. Source MDCT power spectrograms p_{jfn} (5) (first row), their structured approximations v_{jfn} (3) (second row), NTF matrices (third row), histograms of NTF log-coefficients (*bars*) and two-state GMMs (*solid line*) modeling them (fourth row). In this example $J = 3$, $F = 1024$, $N = 421$ and $K = 9$.

C. Source encoding and reconstruction

Given the Gaussian model outlined above, source coding would amount to encode each source vector $\mathbf{s}_{fn}$ according to its prior distribution (6). The main idea of CISS is to employ exactly the same techniques as in source coding, but to use instead its *posterior* distribution (7).

In the Gaussian case, such an encoding is readily performed through constrained entropy quantization relying on scalar quantization in the mean-removed Karhunen-Loeve transform (KLT) domain, as described in [30]. We summarize below its main steps.

Let $\mathbf{\Sigma}_{\mathbf{s},fn}^{\mathrm{pst}} = \mathbf{U}_{fn}\mathbf{\Lambda}_{fn}\mathbf{U}_{fn}^H$ be the eigenvalue decomposition of the covariance matrix, where $\mathbf{U}_{fn}$ is an orthogonal matrix ($\mathbf{U}_{fn}^H\mathbf{U}_{fn} = \mathbf{I}_J$) and $\mathbf{\Lambda}_{fn} = \mathrm{diag}\{\lambda_{1fn}, \ldots, \lambda_{Jfn}\}$ is a diagonal matrix of eigenvalues. The linear transform $\mathbf{U}_{fn}^H$ decorrelating $\mathbf{s}_{fn}$ is the KLT. Assuming the MSE distortion, uniform quantization is asymptotically optimal for the constrained entropy case [34]. Thus, we consider here scalar uniform quantization with a fixed step size Δ

in the mean-removed KLT domain, which can be summarized as follows:

1) Remove the mean and apply the KLT

$$\mathbf{y}_{fn} = \mathbf{U}_{fn}^H(\mathbf{s}_{fn} - \boldsymbol{\mu}_{fn}^{\text{pst}}). \tag{12}$$

2) In the real-valued case, quantize each dimension of $\mathbf{y}_{fn} = [y_{1fn}, \ldots, y_{Jfn}]^T$ with a uniform scalar quantizer $Q_\Delta : y_{jfn} \to \hat{y}_{jfn}$ having a constant step size Δ. In the complex-valued case, the same quantization is applied independently to real and imaginary parts of y_{jfn}. Using an arithmetic coder as an entropy coder [30], the effective codeword length (in bits) is given by

$$L(\mathbf{s}_{fn}|x_{fn};\theta) = -\sum_{j=1}^{J} \log_2 \int_{y-\hat{y}_{jfn}\in\mathcal{A}(\Delta)} N_{r/c}(y;0,\lambda_{jfn})dy. \tag{13}$$

where $\mathcal{A}(\Delta) \triangleq [-\Delta/2, \Delta/2]$ in the real valued case, and $\mathcal{A}(\Delta) \triangleq \{z \in \mathbb{C} | \max(|\Re z|, |\Im z|) \leq \Delta/2\}$ in the complex-valued case.

3) Reconstruct the quantized source vector $\hat{\mathbf{s}}_{fn}$

$$\hat{\mathbf{s}}_{fn} = \mathbf{U}_{fn}\hat{\mathbf{y}}_{fn} + \boldsymbol{\mu}_{fn}^{\text{pst}}. \tag{14}$$

D. Model estimation and encoding

In this section we first detail the strategy for the estimation and quantization of the NTF parameters θ (see Fig. 3). Our derivations mostly follow those from [29]. However, they are applied here to the NTF model instead of the autoregressive model considered in [29]. As highlighted above, the optimal approach would consider posterior distribution (7) for the model estimation and encoding. However, the derivation of the corresponding estimation strategy is overly complex and did not permit us to obtain a simple solution. To simplify this analysis we then assume that the sources were quantized using the prior distribution (6) instead of the posterior one (7). This choice leads us to an optimization strategy based upon some standard algorithms, and we leave the more optimal case of the posterior optimization for further study. Note however that if the posterior distributions are not used in the analysis of model estimation and encoding they are indeed used below for the residual sources encoding.

1) Model estimation: Under high-rate theory assumptions and given the model parameter θ, the total rate (in bits) required to encode the sources $\mathbf{S} = \{s_{jfn}\}_{j,f,n}$ is [29]

$$R(\mathbf{S}|\theta) = -\log_2 p(\mathbf{S}|\theta) - \frac{\mathcal{M}(J,F,N)}{2} \log_2 \frac{D}{C_s}, \tag{15}$$

where $D = C_s\Delta^2$ is the mean distortion (per real-valued dimension), defined as $D \triangleq \mathbb{E}[|\hat{s}_{jfn} - s_{jfn}|^2]$ in the real-valued case and as $D \triangleq (1/2)\mathbb{E}[|\hat{s}_{jfn} - s_{jfn}|^2]$ in the complex-valued case; $C_s = 1/12$ is the coefficient of scalar quantization, and $\mathcal{M}(J,F,N)$ denotes the total number of real-valued coefficients in $\mathbf{S}$, i.e., $\mathcal{M}(J,F,N) \triangleq JFN$ in the real-valued case and $\mathcal{M}(J,F,N) \triangleq 2JFN$ in the complex-valued case. Thus, the model parameter θ should be estimated in the maximum likelihood (ML) sense, as follows

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{S}|\theta) \tag{16}$$

that, in the case of the NTF model, can be shown equivalent to [14], [35]

$$\widehat{\mathbf{Q}}, \widehat{\mathbf{W}}, \widehat{\mathbf{H}} = \arg\min_{\mathbf{Q},\mathbf{W},\mathbf{H}} \sum_{jfn} d_{IS}\left(p_{jfn} \middle| \sum_{k=1}^{K} q_{jk}w_{fk}h_{nk}\right), \tag{17}$$

where p_{jfn} is defined by (5) and $d_{IS}(x|y) = x/y - \log(x/y) - 1$ is the Itakura-Saito (IS) divergence. The optimization of criterion (17) can be achieved by iterating the following multiplicative updates [10], [14]:

$$q_{jk} \leftarrow q_{jk} \left(\frac{\sum_{f,n} w_{fk} h_{nk} p_{jfn} v_{jfn}^{-2}}{\sum_{f,n} w_{fk} h_{nk} v_{jfn}^{-1}} \right), \tag{18}$$

$$w_{fk} \leftarrow w_{fk} \left(\frac{\sum_{j,n} h_{nk} q_{jk} p_{jfn} v_{jfn}^{-2}}{\sum_{j,n} h_{nk} q_{jk} v_{jfn}^{-1}} \right), \tag{19}$$

$$h_{nk} \leftarrow h_{nk} \left(\frac{\sum_{j,f} w_{fk} q_{jk} p_{jfn} v_{jfn}^{-2}}{\sum_{j,f} w_{fk} q_{jk} v_{jfn}^{-1}} \right). \tag{20}$$

2) Model quantization and encoding:

a) Criterion for quantization: Assuming the model parameter quantized and transmitted (Fig. 3), the total rate required to encode the sources becomes [29]

$$R(\mathbf{S}) = \psi(\bar{\theta}, \hat{\theta}, \mathbf{S}) + R(\mathbf{S}|\hat{\theta}), \tag{21}$$

where

$$\psi(\bar{\theta}, \hat{\theta}, \mathbf{S}) \triangleq R(\bar{\theta}) + \log_2 \left(p(\mathbf{S}|\hat{\theta}) / p(\mathbf{S}|\bar{\theta}) \right) \tag{22}$$

is the *index of resolvability* [29] involving the rate required to encode the model $R(\bar{\theta})$ and a term representing the loss in the rate for source encoding due to the usage of the quantized $\bar{\theta}$ model instead of the ideal ML model $\hat{\theta}$. Relying on some realistic approximations (see below) this term can be shown independent of $\mathbf{S}$, and, denoted by $\Psi(\bar{\theta}, \hat{\theta}) \triangleq \log \left(\frac{p(\mathbf{S}|\hat{\theta})}{p(\mathbf{S}|\bar{\theta})} \right)$, it can be expressed as

$$\begin{aligned}
\Psi(\hat{\theta}, \bar{\theta}) &= \frac{1}{2} \sum_{j,f,n} \left(\frac{p_{jfn}}{\bar{v}_{jfn}} - \frac{p_{jfn}}{\hat{v}_{jfn}} - \log \frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} \right) && (23) \\
&= \frac{1}{2} \sum_{j,f,n} \left(\frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - \log \frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - 1 \right) + \\
&\quad \frac{1}{2} \sum_{j,f,n} \left(\frac{p_{jfn} - \hat{v}_{jfn}}{\hat{v}_{jfn}} \frac{\hat{v}_{jfn} - \bar{v}_{jfn}}{\bar{v}_{jfn}} \right) && (24) \\
&\approx \frac{1}{2} \sum_{j,f,n} \left(\frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - \log \frac{\hat{v}_{jfn}}{\bar{v}_{jfn}} - 1 \right) && (25) \\
&\approx \frac{1}{4} \sum_{j,f,n} \left(\log \hat{v}_{jfn} - \log \bar{v}_{jfn} \right)^2, && (26)
\end{aligned}$$

where approximation (25) follows from a reasonable assumption that the relative error of modeling $(p_{jfn} - \hat{v}_{jfn})/\hat{v}_{jfn}$ and that of quantization $(\hat{v}_{jfn} - \bar{v}_{jfn})/\bar{v}_{jfn}$ are uncorrelated [29] and at least one of these errors is zero-mean.

The last approximation (26) is obtained using the following second order Taylor expansion $u \approx 1 + \log(u) + \frac{1}{2}\log(u)^2$ in the neighborhood of $u = 1$ (with $u = \hat{v}_{jfn}/\bar{v}_{jfn}$), as in [29], [36]. Note that we find again the IS divergence in the expression (25), and the last approximation (26) indicates that the NTF model variances $\hat{v}_{jfn}$, structured as in (3), should be quantized by minimizing the MSE of their logarithms.

This result is quite similar to what was done in [14], where the log-spectrograms were compressed using the JPEG image coder. However, while [14] does not justify this particular choice, we provide here a theoretical explanation of its appropriateness.

b) NTF parameters quantization: Although the criterion (26) is quite simple, it does not give yet any precise idea of how to quantize individual NTF model parameters, i.e., matrices $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$. Using (3), the criterion (26) can be rewritten as

$$\Psi(\hat{\theta}, \bar{\theta}) \approx \frac{1}{4} \sum_{j,f,n} \left(\log \sum_{k=1}^{K} \hat{q}_{jk} \hat{w}_{fk} \hat{h}_{nk} - \log \sum_{k=1}^{K} \bar{q}_{jk} \bar{w}_{fk} \bar{h}_{nk} \right)^2. \tag{27}$$

We see that there are quite complicated dependencies between elements of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ in this criterion. To simplify this expression we consider the following criterion

$$\Phi(\hat{\theta},\bar{\theta}) = \frac{1}{4}\sum_{j,f,n}\sum_{k}\left(\log\hat{q}_{jk}\hat{w}_{fk}\hat{h}_{nk} - \log\bar{q}_{jk}\bar{w}_{fk}\bar{h}_{nk}\right)^2 \tag{28}$$

that is in fact an upper bound of (27), i.e.,

$$\Psi(\hat{\theta},\bar{\theta}) \leq \Phi(\hat{\theta},\bar{\theta}), \tag{29}$$

which can be shown by applying Lemma A.1 from Appendix A with $c = 1$ and $f(u) = \log(u)^2$. Note however that this upper bound is not very tight, as it can be seen from the proof of Lemma A.1.

Now, assuming that the entries of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ are quantized independently the cross-terms in (28) will be canceled in average (if $K \times \min(J,F,N)$ is big enough), due to the fact that the quantization noise of say $\mathbf{Q}$ will be decorrelated with that of say $\mathbf{W}$. Thus, (28) can be rewritten

$$\begin{aligned}\Phi(\hat{\theta},\bar{\theta}) &= \frac{1}{4}\sum_{j,f,n}\sum_{k}\left[(\log\hat{q}_{jk} - \log\bar{q}_{jk})^2 + (\log\hat{w}_{fk} - \log\bar{w}_{fk})^2 + \left(\log\hat{h}_{nk} - \log\bar{h}_{nk}\right)^2\right] \\ &= \frac{JFN}{4}\sum_{k}\left[\frac{1}{J}\sum_{j}(\log\hat{q}_{jk} - \log\bar{q}_{jk})^2 + \frac{1}{F}\sum_{f}(\log\hat{w}_{fk} - \log\bar{w}_{fk})^2 + \frac{1}{N}\sum_{n}\left(\log\hat{h}_{nk} - \log\bar{h}_{nk}\right)^2\right]. \end{aligned} \tag{30}$$

Under all approximations above, we conclude that, if we choose to independently quantize NTF coefficients under an entropy constraint, we should use scalar quantizers of their logarithms. Thus, we opt for a logarithmic compressor, followed by a scalar quantizer and an exponential expander. It is interesting to note that Nikunen *et al.* [27], [28] use μ-law compressor and expander to quantize NTF / NMF coefficients, and the μ-law compressor also acts as logarithmic for high values. Note finally that our NTF model has a different goal than the one presented in [27], [28]. The NTF considered in [27], [28] models both the source and the perception, while our goal is to model source distribution only, and we propose addressing perceptual aspects separately (see Fig. 3). Thus, given different modeling goals the ways the NTF parameters are quantized may be different as well.

We see that squared log-differences of different NTF parameters appear with different weights in the summation of (30). Thus, in order to have the MSE over all parameters, the parameters, up to the same uniform quantization, should be divided by the square roots of these weights, or, equivalently, they should be quantized with different step-sizes $\Delta_{\mathbf{Q}}$, $\Delta_{\mathbf{W}}$ and $\Delta_{\mathbf{H}}$ (respectively, to quantize logarithms of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$) computed as follows

$$\Delta_{\mathbf{Q}} = \sqrt{J/(J+F+N)}\cdot\Delta_\theta, \tag{31}$$

$$\Delta_{\mathbf{W}} = \sqrt{F/(J+F+N)}\cdot\Delta_\theta, \tag{32}$$

$$\Delta_{\mathbf{H}} = \sqrt{N/(J+F+N)}\cdot\Delta_\theta, \tag{33}$$

where Δ_θ is some global model quantization step-size governing the rate-distortion trade-off. We see that within our framework (that is based on high-rate theory) we are able to find an analytical solution for the allocation of the rate between different NTF parameters, while in [27], [28] such an allocation was established experimentally. Thus, our approach has the following advantages over [27], [28]. First, it permits to considerably reduce the number of parameters to be optimized experimentally. Second, we show that the rate allocation between NTF parameters depends on the NTF dimensions J, F and N, and, as a consequence it depends, e.g., on the length of the signal to be encoded and on the number of sources. Thus, we show that even if an experimental optimization of this rate allocation is followed, it should be performed again every time one of these parameters (e.g., signal length) changes.

c) NTF parameters encoding by GMMs: In order to quantize each of the three NTF matrices we model the distribution of its log-coefficients by a two-state Gaussian mixture model (GMM) (see the fourth raw of Fig. 5). GMMs are denoted $\xi_{\mathbf{Q}}$, $\xi_{\mathbf{W}}$ and $\xi_{\mathbf{H}}$ (see Fig. 4) and optimized in the ML sense for each matrix, thus their parameters must be transmitted resulting in a very small extra rate (there are only 15 parameters, i.e., 5 parameters per matrix: two means, two variances and one weight). As an alternative the Huffman coding can be used as well, as it is done in [27], [28]. There are pros and cons for using Huffman coding. From the one hand, it is optimal. From the other hand, it requires transmitting a codebook to the decoder, which can be more costly, as compared to transmitting just the five parameters of a GMM.

E. Operational rate-distortion function and parameter optimization

Now we write a so-called *operational rate-distortion function (RDF)* [37] that is accurate for high rates and gives a practical relation between rate and distortion for our CISS coding scheme. Considering (15), but now with posterior $p(\mathbf{S}|\mathbf{X},\theta)$ instead of prior $p(\mathbf{S}|\theta)$, and adding to it the rate required to encode the model parameter $R(\bar{\theta})$, one can show that the total rate (in bits) R_{tot} relates to the mean distortion (per dimension) as

$$R_{\text{tot}} = -\frac{\mathcal{M}(J,F,N)}{2}\log_2\frac{D}{C_s} + \eta(\mathbf{S},\mathbf{X},\bar{\theta}), \tag{34}$$

with

$$\eta(\mathbf{S},\mathbf{X},\bar{\theta}) \triangleq R(\bar{\theta}) - \log_2 p(\mathbf{S}|\mathbf{X},\bar{\theta}), \tag{35}$$

that is independent [4] of the rate R_{tot} and distortion D. Thus, in order to optimize operational RDF (34) for any high rate, one needs to minimize (35).

The only free parameters we need to optimize experimentally are the model quantization step-size Δ_θ (determining the model rate $R(\bar{\theta})$) and the number of NTF components K. We optimize these parameters so as to minimize $\eta(\mathbf{S},\mathbf{X},\bar{\theta})$ from (35). These parameters can be either optimized globally for a set of signals, or they can be re-optimized for each signal to be encoded. In the last case the parameters must be quantized and transmitted to the decoder.

IV. Experiments

In this section we evaluate the proposed single-channel CISS-NTF method for both STFT and MDCT representations. This evaluation includes the optimization of different parameters and the comparison with relevant state of the art methods.

A. State of the art methods

As for conventional ISS, we consider two state of the art methods proposed in [14], [38]. Both methods are based on a parametric reconstruction of the sources via Wiener filtering in the STFT domain, while the source spectrograms (the variances used to compute Wiener filter) are encoded differently. In the first method, referred to as *Wiener-JPEG*, the images of source log-spectrograms are encoded by the JPEG lossy coder. In the second method, referred to as *Wiener-NTF*, source spectrograms are approximated by exactly the same NTF model as the one considered here. Parvaix *et al.* [12] introduced another conventional ISS method that is suitable for single channel mixtures. This method is based on binary masking of sources in the MDCT domain, while it is known [16] that oracle bounds of binary masking-based methods are lower than those of Wiener filter-based methods [14], [38]. Thus, we here consider only Wiener filter-based methods for comparison.

B. Testing methodology

1) Data: We considered seven single-channel mixtures of several musical sources such as singing voice, bass, guitar, piano, distorted guitars, etc ... The number of sources J varies from 3 to 6, and the duration of each mixture is about 20 seconds. All signals are sampled at either 48kHz or 44.1kHz. For each mixture the sources were obtained by summing up stereo source images from the QUASI database [5] and by restricting them to a desired time duration. Sources from the same artist were never included into different mixtures.

2) Parameters: MDCT and STFT were computed with frames of 2048 samples and 50 % overlap for STFT. Note however that due to STFT redundancy, as compared to MDCT, this representation includes twice as many real-valued coefficients $\mathcal{M}(J,F,N)$ to be encoded.

[4] We know from [29] that, under high-rate theory assumptions, the optimal model rate $R(\bar{\theta})$ is constant, thus independent on the total rate R_{tot}.

[5] http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

3) Evaluation metrics: Since ISS is an emerging research area lying in between source separation and lossy audio coding, we used evaluation metrics coming from these two fields. Notably, we used signal-to-distortion ratio (SDR) [39] usually used to evaluate source separation algorithms and perceptual similarity measure (PSM) of PEMO-Q [40] usually used to evaluate perceptual quality of lossy audio coding schemes. We used the implementation provided by [41] for this purpose.

In most experiments presented below we do not consider directly SDR and PSM but rather the improvements of these measures, denoted as $\delta\mathrm{SDR}$ and $\delta\mathrm{PSM}$, over the corresponding measures computed with the oracle Wiener filtering source estimates [6] in the STFT domain. These oracle performances are shown in Fig. 8 for each mixture from test dataset.

C. Simulations

1) High-rate optimal parameters: As it is explained in section III-E, for high rates, the optimal model quantization step size Δ_θ and the optimal number of NTF components K must be constant, i.e., independent of the total rate. To find these optimal parameters in the case of the STFT representation we have computed $\eta(\mathbf{S}, \mathbf{X}, \bar{\theta})$ from (35) for each mixture for different combinations of model quantization step sizes $\Delta_\theta = [1.8, 0.5, 0.13, 0.04, 0.01]$ and numbers of NTF components per source $K/J = [2, 3, 4, 5, 10, 15, 20, 30]$, and we averaged the result over all mixtures. We observed that the average $\eta(\mathbf{S}, \mathbf{X}, \bar{\theta})$ reaches its minimum for $\Delta_\theta = 0.1$ and $K/J = 4$, which are thus in average the optimal parameters for high rates. These results, i.e. in average 4 NTF components per source, are in fact consistent with what was found in [42], where a similar modeling was considered for conventional source separation.

2) CISS-NTF with STFT and different ways of optimizing the parameters: The parameters Δ_θ and $K/J = 4$, that have been found optimal in the previous section, are only optimal for high rates and in average. Thus, first, it could be that for some low rates (that can be attractive in practice) the optimal parameters are different. Second, it could be that the optimal parameters, especially the optimal number of NTF components per source K/J, varies from one mixture to another. Indeed, intuitively it seems that a mixture composed of "simple" sources (e.g., triangle) should require less NTF components than a mixture composed of "complex" sources (e.g., organ). The goal of the following experiments is to clarify these points by first evaluating the proposed CISS-NTF for different parameters and over a range of rates, and then by investigating and comparing the optimal parameters for low/high rates and for different mixtures.

We first consider CISS-NTF in the STFT domain, and address the MDCT domain later. This is because the state of the art approaches were designed for STFT domain, and we would like to investigate the possible advantage of CISS-NTF over the state of the art besides the change of the signal representation considered. We have evaluated the CISS-NTF over the same different parameters Δ_θ and K/J as in the previous section, and over a wide range of rates by using 10 logarithmically-spaced values for the source quantization step size as $\Delta = \mathrm{logspace}\,(-0.15, 2.5, 10)$. The source quantization step size $\Delta = +\infty$ has also been tested and corresponds to simply omitting the "waveform source encoding" block in CISS (Fig. 3), so that it essentially becomes a conventional ISS approach (Fig. 1). However, this scheme is still different from Wiener-NTF approach of [14], [38], since in our approach NTF parameters are quantized in log-domain with any step size Δ_θ, while in [14], [38] it was proposed to quantize NTF parameters in the linear domain with a fixed small step size.

The simulations described above gave us many (rate, $\delta\mathrm{SDR}$) pairs, for which we have also computed $\delta\mathrm{PSM}$. Then, for each small range of rates we have chosen (under certain constraints, as described below) the pairs corresponding to the highest $\delta\mathrm{SDR}$. The resulting points in (rate, $\delta\mathrm{SDR}$) and (rate, $\delta\mathrm{PSM}$) planes were then smoothed using LOESS to produce the rate/performance curves. We have computed the following curves:

- **[Opt-HR-avg]** the same parameters (i.e., Δ_θ and K/J) for all rates and all mixtures optimized for high-rates (i.e., exactly as in section IV-C1),
- **[Opt-LR-avg]** the same parameters for all rates and all mixtures optimized for law-rates (0.5-2 kbps per source),
- **[Opt-HR-mix]** parameters constant over rates, but optimized for each particular mixture for high-rates,
- **[Opt-LR-mix]** parameters constant over rates, but optimized for each particular mixture for low-rates,

[6]The oracle Wiener filtering source estimates are computed by equation (10), where the structured prior source variances v_{jfn} in (8) are replaced by the true source power spectrograms $p_{jfn} = |s_{jfn}|^2$.

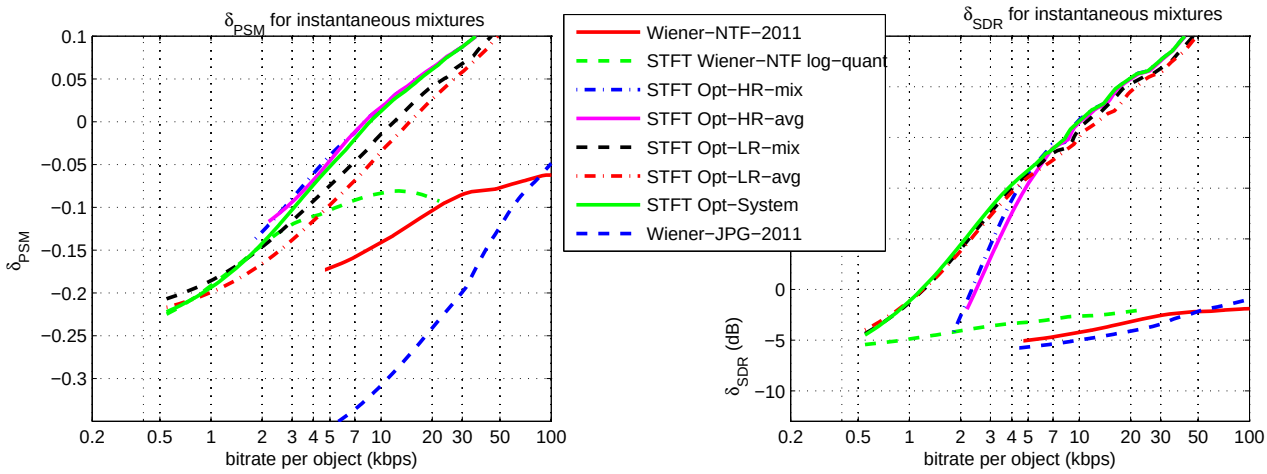

Fig. 6. CISS-NTF with STFT and different ways of optimizing parameters, compared to state of the art. The δ indicates the relative score of the technique considered with respect to the oracle performance.

- **[Opt-System]** parameters systematically optimized to a particular rate and a particular mixture,

and we have plotted them in Fig. 6. This figure includes as well the results of Wiener-NTF [14], [38] state of the art method and the results of the so called *Wiener-NTF-log-quant* method that is similar to Wiener-NTF, but using newly-proposed log-domain NTF parameters quantization, i.e., with $\Delta = +\infty$.

One can note from Fig. 6 that Wiener-NTF-log-quant outperforms Wiener-NTF for the SDR metric for all rates. That shows the advantage of the proposed log-quantization of NTF parameters over the state of the art [14], [38]. Moreover, waveform source quantization of CISS brings further a great advantage over Wiener-NTF, outperforming it by a large margin for all rates. Also, it outperforms the oracle Wiener results (zero levels of δSDR and δPSM measures) starting from 1-2 kbps per source for SDR, and starting from 7-10 kbps per source for PSM. Note also that the performances of CISS-NTF obtained with parameters optimized for each mixture and/or each particular rate are not much better than the performances with fixed parameters (optimized in average for low or high rates). This is a very good news for a practical coder implementation. Indeed, that means that one does not need to adjust Δ_θ and K/J to each particular mixture, and can just keep them fixed. Finally, it should be noted that for PSM, high-rate optimized parameters (in terms of SDR) are better for low-rates than low-rate optimized parameters (in terms of SDR). This observation indicates a possible use of the distribution preserving quantization (DPQ) [43] to better model perceptual quality.

3) CISS-NTF with MDCT and STFT vs. the state-of-the-art: We have performed for CISS-NTF with MDCT exactly the same simulations as for CISS-NTF with STFT. The qualitative behavior of the results with different ways of optimizing the parameters was exactly the same as for CISS-NTF with STFT, as reported in the previous section. Thus, for these results we can draw exactly the same conclusions as in the previous section for STFT, and we here show in Fig. 7 the results with average parameters optimized for high/low-rates (**[Opt-HR-avg]** and **[Opt-LR-avg]**) for both STFT and MDCT. We have also added the results of the two state of the art methods: Wiener-NTF and JPEG-NTF [14], [38]. We see that CISS-NTF with MDCT outperforms CISS-NTF with STFT for very low rates. This improvement is mostly due to the fact that the MDCT representation is critically sampled, i.e., includes as many coefficients as the time-domain signal, while the STFT is redundant.

In any case, using CISS-NTF with STFT is attractive in the multichannel case, and this is what we have done in [26]. Indeed, most of probabilistic multichannel models in source separation involving convolutive mixing [5], [6] are specified in the STFT domain.

4) Summary of results: Fig. 8 gives a summary of the results for each mixture obtained by the oracle Wiener filtering, two state of the art methods (Wiener-NTF and JPEG-NTF), and the proposed CISS-NTF with MDCT. The results are now presented in terms of SDR and PSM absolute values (not their increments δ as before) and for a mean bitrate of 5 kbps per source, which is attractive for practical applications. We observe on this figure that the proposed method largely outperforms state of the art, while it uses a smaller bitrate.

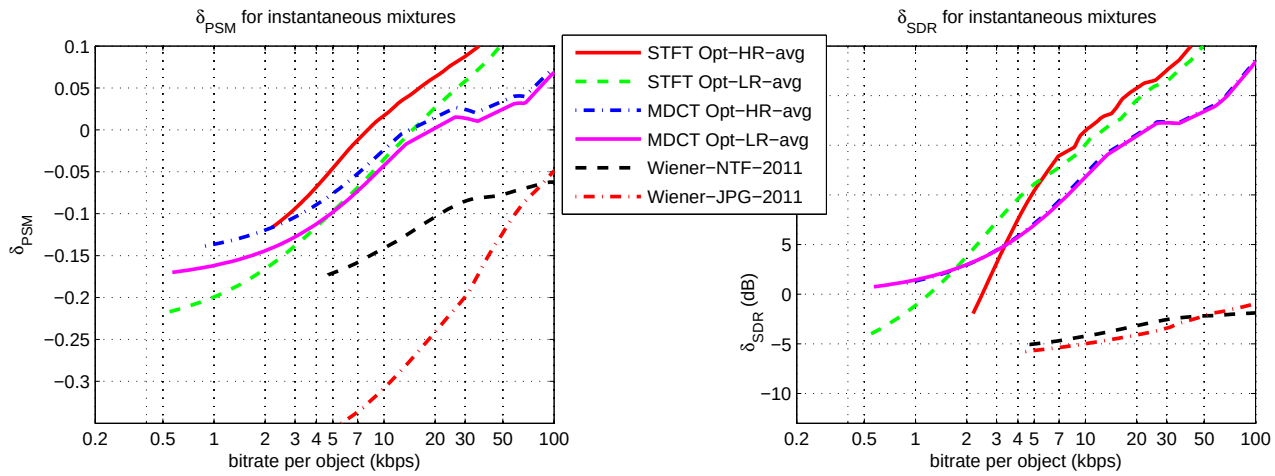

Fig. 7. CISS-NTF with MDCT and STFT vs. the state-of-the-art. Evaluation was performed using both STFT and MDCT transforms. We see that performances of MDCT are better than those of STFT for low-bitrates.

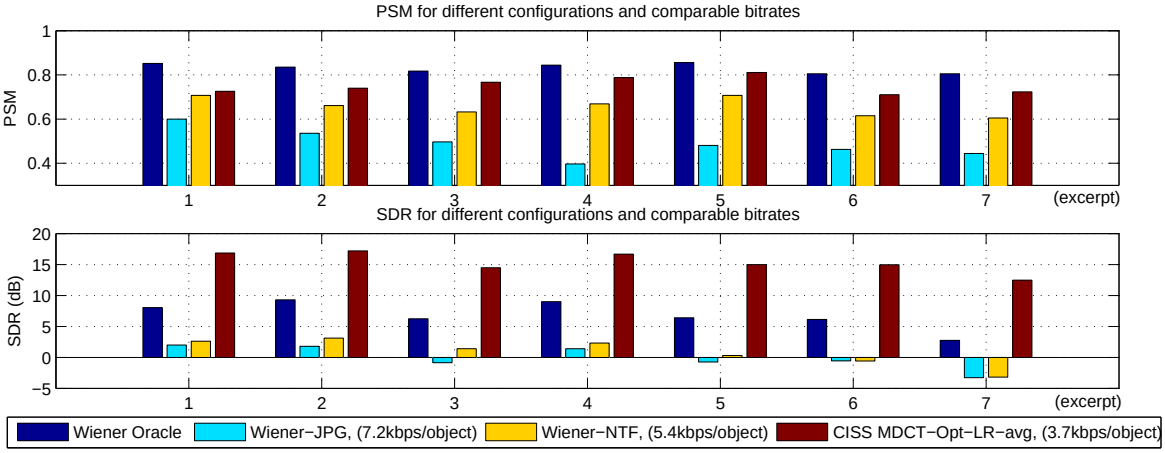

Fig. 8. Summary of results for all 7 excerpts of the database. For each excerpt, the SDR and PSM scores of Oracle source separation is compared to those of state of the art and of the proposed method. CISS-NTF largely outperforms all other techniques, for a smaller bitrate.

V. Conclusion

We have introduced CISS, a general probabilistic framework for ISS and SAOC. We have further detailed and evaluated in the single-channel mixture case its particular instance called CISS-NTF based on a probabilistic NTF source representation. This approach relates at the same time to different state of the art areas, notably ISS [13], [14], SAOC [17]–[19] and NTF / NMF model-based audio compression [27], [28]. We have discussed possible advantages of CISS in general and of its particular instance, CISS-NTF, over all these state of the art approaches. In summary, without going into details, the main advantages of CISS and CISS-NTF are:

1) waveform quantization based on a structural probabilistic source model (NTF) allowing modeling long-term redundancy in audio signals;
2) in contrast to the conventional ISS and SAOC methods, the parameters used for parametric and waveform coding are jointly encoded within this probabilistic model;
3) the proposed probabilistic formulation allows using NTF / NMF models specified over critically sampled signal representations such as the MDCT, which are known more efficient for compression;
4) in contrast to the conventional ISS methods, even if it was not yet implemented, the proposed CISS allows using advanced perceptual models for enhanced perceived quality.

Our extensive experimental evaluation has shown a great advantage of the proposed CISS-NTF approach over the state of the art conventional ISS methods.

This work opens the doors for various further investigations. First, given that most music recordings nowadays are at least stereo, CISS-NTF should be extended to the multichannel case in order to improve its efficiency due to the spatial source diversity. Some work using STFT signal representation being already done in this direction [26], some questions remain still open. Notably, how to cope with STFT redundancy that is undesirable within compression applications either by reducing STFT overlap or by resorting to critically sampled transforms such as MDCT (see discussion in Sec IV-C3). Second, perceptual modeling should be integrated within CISS-NTF and it should be compared with SAOC, when an optimized encoder for this emerging standard is available. The sensitivity matrix approach [33] combined with the newly introduced distribution preserving quantization (DPQ) [43] (see also discussion in Sec IV-C2) seem to be good candidates for modeling perception within this Gaussian model-based approach. Third, remember that in order to simplify the optimization we have chosen here a *generative* model estimation approach optimizing the prior distribution (6) instead of a *discriminative* model estimation optimizing the posterior (7), which is more optimal (see Sec. III-D). Thus, new model estimation algorithms should be proposed to implement the discriminative approach. Finally, the NTF source model can be replaced by possibly better structured probabilistic models to improve coding efficiency. In fact, any model from those implementable by a general source separation framework presented in [6] can be used in principle.

More generally, besides ISS and SAOC applications, and in line with [27], [28], the proposed NTF-based approach (with some modifications) could be applied for regular and multichannel audio coding. Moreover, our approach is related to the context-based adaptive entropy coding schemes used for audio and video compression [44], [45]. However, our approach seems to be "more locally adaptive", since each frame is encoded by its own arithmetic coder having a distribution derived from local signal statistics. In other words, each frame has its own context. Thus, it would be interesting to extend such kind of advanced statistical model-based approaches for image or video compression.

APPENDIX A
ONE LEMMA

Lemma A.1. *Let $K \in \mathbb{N}$ and $c \in \mathbb{R}_+^*$. Let $f : \mathbb{R}_+^* \to \mathbb{R}$ a continuous function, that is strictly decreasing on $]0, c[$ and strictly increasing on $]c, +\infty[$.*

Then $\forall\ \hat{x}_1 \dots \hat{x}_K,\ \bar{x}_1 \dots \bar{x}_K \in \mathbb{R}_+^*$,

$$f\left(\frac{\sum_{k=1}^K \hat{x}_k}{\sum_{k=1}^K \bar{x}_k}\right) \leq \sum_{k=1}^K f\left(\frac{\hat{x}_k}{\bar{x}_k}\right). \tag{36}$$

Proof: We assume that $\forall k \in \{1 \dots K\}$, $u_k = \frac{\hat{x}_k}{\bar{x}_k}$, $\lambda_k = \frac{\bar{x}_k}{\sum_{k'=1}^K \bar{x}_{k'}}$ and $u = \sum_{k=1}^K \lambda_k u_k = \frac{\sum_{k=1}^K \hat{x}_k}{\sum_{k=1}^K \bar{x}_k}$. With these notations we need to prove that $f(u) \leq \sum_{k=1}^K f(u_k)$.

Since f is continuous, strictly decreasing on $]0, c[$ and strictly increasing on $]c, +\infty[$, it is clear that it reaches its maximum on any interval of the form $[a, b]$ (with $0 < a < b < +\infty$), and this maximum is reached either in a or in b.

We then define $a = \min(u_1 \dots u_K)$ and $b = \max(u_1 \dots u_K)$. Since $\forall k$, $\sum_{k=1}^K \lambda_k = 1$, it is clear that $u \in [a, b]$. Thus, we conclude that $f(u) \leq \max(f(a), f(b)) \leq \sum_{k=1}^K f(u_k)$. ■

REFERENCES

[1] MPEG-1 Audio, Layer III, "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s – Part 3: Audio," *ISO/IEC 11172-3:1993*, 1993.

[2] MPEG-2 Advanced Audio Coding, AAC, "Information technology – Generic coding of moving pictures and associated audio information – Part 3: Audio," *ISO/IEC 13818-3:1998*, 1998.

[3] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, Apr. 2009, pp. 1 –4.

[4] P. Comon and C. Jutten, *Handbook of blind source separation: independent component analysis and applications*. Academic Press, 2010.

[5] E. Vincent, M. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010, ch. 7, pp. 162–185.

[6] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[7] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[8] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

[9] P. Smaragdis and G. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, Oct. 2009, pp. 69 –72.

[10] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, May 2011, pp. 257–260.

[11] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

[12] M. Parvaix, L. Girin, and J.-M. Brossier, "A watermarking-based method for informed source separation of audio signals with a single sensor," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1464–1475, 2010.

[13] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1721 – 1733, 2011.

[14] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[15] 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 *20th International Congress on Acoustics (ICA 2010)*, Sydney, Australia, Aug. 2010.

[16] E. Vincent, R. Gribonval, and M. Pumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933 – 1950, Aug. 2007.

[17] 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)*, Beijing, China, Jul. 2007, pp. 1894–1897.

[18] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," in *124th Audio Engineering Society Convention (AES 2008)*, Amsterdam, Netherlands, May 2008.

[19] 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, Sep. 2010.

[20] O. Hellmuth, H. Purnhagen, J. Koppens, J. Herre, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh, "MPEG spatial audio object coding - The ISO/MPEG standard for efficient coding of interactive audio scenes," in *129th Audio Engineering Society Convention (AES 2010)*, 2010.

[21] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "High-fidelity multichannel audio coding with Karhunen-Loève transform," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 365–380, Jul. 2003.

[22] C. Faller, "Parametric multichannel audio coding: Synthesis of coherence cues," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 299–310, Jan. 2006.

[23] B. Cheng, C. Ritz, and I. Burnett, "Encoding independent sources in spatially squeezed surround audio coding," in *8th Pacific Rim Conference on Multimedia (PCM'07)*, Hong Kong, China, Dec. 2007, pp. 804–813.

[24] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "A multichannel sinusoidal model applied to spot microphone signals for immersive audio," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 8, pp. 1483–1497, Nov. 2009.

[25] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed source separation: source coding meets source separation," in *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'11)*, New Paltz, New York, USA, Oct. 2011, pp. 257–260.

[26] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial coding-based informed source separation," in *EUSIPCO, 20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012, to appear.

[27] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," in *128th Audio Engineering Society Convention (AES 2010)*, London, UK, May 2010.

[28] 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'11)*, New Paltz, New York, USA, Oct. 2011, pp. 33–36.

[29] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'07)*, New Paltz, New York, USA, Oct. 2007, pp. 243–246.

[30] D. Y. Zhao, J. Samuelsson, and M. Nilsson, "On entropy-constrained vector quantization using Gaussian mixture models," *IEEE Trans. Commun.*, vol. 56, no. 12, pp. 2094–2104, Dec. 2008.

[31] T. Dau, D. Pušchel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system: I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, pp. 3615–3622, 1996.

[32] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1292–1304, 2005.

[33] J. H. Plasberg and W. B. Kleijn, "The sensitivity matrix: Using advanced auditory models in speech and audio processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 310–319, 2007.

[34] R. M. Gray, *Source coding theory*. Kluwer Academic Press, 1990.

[35] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[36] A. Buzo, A. Gray, R. Gray, and J. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 5, pp. 562–574, 1980.

[37] A. Ozerov and W. B. Kleijn, "Asymptotically optimal model estimation for quantization," *IEEE Trans. Commun.*, vol. 59, no. 4, pp. 1031–1042, Apr. 2011.
[38] 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.
[39] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[40] R. Huber and B. Kollmeier, "PEMO-Q Ű a new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 1902 – 1911, 2006.
[41] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2046–2057, 2011.
[42] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.
[43] M. Li, J. Klejsa, and W. B. Kleijn, "Distribution preserving quantization with dithering and transformation," *IEEE Signal Process. Lett.*, vol. 17, no. 12, pp. 1014–1017, 2010.
[44] K. Lakhdhar and R. Lefebvre, "Context-based adaptive arithmetic encoding of EAVQ indices," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1473 – 1481, Jul. 2012.
[45] D. Marpe, H. Schwarz, and T.Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 620–644, Jul. 2003.