Spatial coding-based informed source separation

Antoine Liutkus, Alexey Ozerov, Roland Badeau, Gael Richard

SPATIAL CODING-BASED INFORMED SOURCE SEPARATION

Antoine Liutkus[1] *Alexey Ozerov*[2] *Roland Badeau*[1] *Gaël Richard*[1]

[1] Institut Telecom, Telecom ParisTech, CNRS LTCI, France.
[2] Technicolor Research & Innovation, France

ABSTRACT

Recent advances in multimedia technology opened the path for individual manipulation of the different audio objects within a multichannel mix, for both sampling and karaoke applications. This requires the transmission of these objects as an additional information. Informed Source Separation (ISS) is an adequate framework for this problem. Its main idea is not to transmit the objects themselves, but rather the parameters required to recover them using the mixtures and separation algorithms. In recent studies, the connection was made between ISS and source coding and the concept of coding-based ISS (CISS) was introduced. CISS differs from classical source coding in its use of the mixtures, which permits to reduce the bitrates required to convey audio objects compared to source coding alone with the same model. In this study, we extend existing work on CISS to the case of multichannel mixtures and demonstrate a considerable increase of performance over classical ISS.

1. INTRODUCTION

Emerging technologies have created new ways to interact with musical contents, the so-called *active listening* scenarios, which include separate manipulation, muting or respatialization of the constituent sound objects, or sources, playing *within* a musical track. Special cases of interest include karaoke or immersion of the listener into surround rendering. To this purpose, it is mandatory to transmit not only the *mixture* as in the usual case, but also its separate constituent audio objects. It was early acknowledged [2] that a solution is to consider the whole set of objects as one multichannel signal and to make use of spatial cues to recover it from the downmix. This idea led to the Spatial Audio Object Coding (SAOC) standard. Independently, researchers from the source separation community reported [9] that source separation could be used to recover constituent sources from a mixture in this context. The difference between the classical *blind* scenario and this particular *informed* configuration is that the sources are known at some *encoding* stage, during which a side information can be computed and transmitted along with the mixtures to be used for separation at a *decoding* stage, when the sources are no longer available. PARVAIX et al. introduced the term Informed Source Separation (ISS) for this strategy. The common idea of all methods exploiting ISS [9, 6] is not to transmit the sources, but rather parameters that permit to recover them using the mixtures.

The main issue with these methods is that their performance is bounded by the best estimates that can be provided by the considered separation method. Hence, whatever the bitrate spent on providing better parameters for the separation algorithms, the quality of the estimates does not improve consequently. This phenomenon is reminiscent of *parametric coding* of waveforms and stems from the intrinsic limitations of the model used to encode the signals of interest. Recently, OZEROV et al. [8] demonstrated that ISS could be significantly improved so as to consistently benefit from additional bitrate as in source coding [3, 11] and introduced Coding-based ISS (CISS) for this purpose. The main idea of CISS is to encode the signals of interest using a probabilistic model as in source coding. Instead of using a distribution which does not make use of the mixture, CISS encodes the sources relying on their *a posteriori* distribution given the mixture, whose entropy is necessarily smaller and which thus leads to reduced bitrates. This idea can be somewhat related to the coding of the residual between the original sources and their estimates as in SAOC. Indeed, these techniques first perform separation of the mixtures and then encode the residuals using a separate source model. CISS can be understood similarly as an estimation of the sources as the mean of their posterior distribution given the mixture, followed by an encoding of the residuals using posterior covariance as signal statistics. Given this parallel, several advantages of CISS over the aforementioned approaches can be highlighted. First, CISS exploits *posterior* dependencies between the sources, instead of independently encoding the residuals. Second, parameters used for parametric source reconstruction and waveform coding of residuals are coupled via posterior distribution and can thus be transmitted more efficiently.

Even if the fundamental idea to encode the signals using posterior distributions can be exploited in many settings, a particular CISS scheme was presented in [8] for single-channel mixtures and Gaussian source model. In this study,

This work is partly funded by the French National Research Agency (ANR) as a part of the DReaM project (ANR-09-CORD-006-03) and partly supported by the European Commission under contract "FP7-ICT-287723-REVERIE."

we generalize this model to multichannel mixtures using a framework inspired by [1] and [6] and we make use of a more specialized Nonnegative Tensor Factorization (NTF) source model as presented in [6, 7], which can efficiently exploit long-term and inter-sources redundancies.

This article is structured as follows. First, we present the Gaussian model we consider for multichannel mixtures in Section 2. Sections 3 and 4 are devoted to model estimation and encoding and to source encoding, given the mixture and the encoded model. Finally, the proposed method is evaluated in Section 5.

2. MODEL

2.1. Notation

All signals considered are regularly sampled time series of length N_n. We will make use of a Time-Frequency (TF) representation for the signals and we choose the Short-Term Fourier Transform (STFT) for this purpose. In this study, waveforms in the temporal domain are written using tilde, e.g. $\tilde{x}(n)$ and their STFT using the corresponding letter without tilde, e.g. $x(\omega, t)$. Bold lowercase indicates a vector while uppercase indicates a matrix or a tensor. N_ω and N_t are, respectively, the number of frequency bins ω and the number of frames t of the STFT.

The *sources*, or audio objects, are defined as M time series $\tilde{s}_m$ and the *mixture* is defined as a set of K time series $\tilde{x}_k$. The mixture is obtained through a processing of the sources. For some given source $\tilde{s}_m$, a *mixing process* produces a set of K signals $\{\tilde{y}_{km}\}_{k=1,\cdots,K}$ called m^{th} source *image*. All images are summed up to produce the k^{th} channel $\tilde{x}_k$ of the mixture.

In the STFT domain, we denote

$$\begin{aligned}\mathbf{s}(\omega,t) &= [s_1(\omega,t),\ldots,s_M(\omega,t)]^\top \\ \mathbf{x}(\omega,t) &= [x_1(\omega,t),\ldots,x_K(\omega,t)]^\top\end{aligned}$$

as the $M \times 1$ and $K \times 1$ column vectors gathering all sources and all channels of the mixture for TF bin (ω, t) with $\cdot^\top$ denoting transposition. The $K \times 1$ image of source m at (ω, t) writes $\mathbf{y}_m(\omega,t) = [y_{1m}(\omega,t),\ldots,y_{Km}(\omega,t)]^\top$.

2.2. Source model

In all the following, we assume that the sources are *independent* and modeled as Locally Stationary Gaussian Processes (LSGP). Basically, this means that within each frame, the sources are stationary and that all the frames of the signals can be considered independent. Under this assumption and provided that the frames are of sufficient length, it can be shown that all TF bins of s_m are independent and normally distributed:

$$s_m(\omega,t) \sim \mathcal{N}_c(0, v_{m,\omega,t}),$$

where $\mathcal{N}_c$ is the circular symmetric complex normal distribution and $v_{m,\omega,t}$ is the variance of source m at TF bin (ω, t). As can be seen, the LSGP model is parameterized by the $M \times N_\omega \times N_t$ tensor $\mathbf{V} = \{v_{m,\omega,t}\}_{m,\omega,t}$ and thus comes with as many parameters as the number of TF bins for the source signals. Since $\mathbf{V}$ has to be transmitted from the coder to the decoder, it is of importance to find an appropriate compression scheme to reduce its weight. Many methods can be used to this end, corresponding to different *source models*. Following the work in [6, 7], we make use of Nonnegative Tensor Factorization (NTF) to decompose $v_{m,\omega,t}$ as:

$$v_{m,\omega,t} = \sum_{r=1}^{R} w_{\omega r} h_{tr} q_{mr}, \tag{1}$$

where $\mathbf{W} = \{w_{\omega r}\}_{\omega,r}$, $\mathbf{H} = \{h_{tr}\}_{t,r}$ and $\mathbf{Q} = \{q_{mr}\}_{m,r}$ are $N_\omega \times R$, $N_t \times R$ and $M \times R$ nonnegative matrices, respectively, and where R is often called the *number of components*. In that case, the *source parameters* θ_s are given by $\theta_s = \{\mathbf{W}, \mathbf{H}, \mathbf{Q}\}$. As demonstrated e.g. in [6, 7], an interesting feature of this particular source model is that it permits to exploit long-term as well as inter-sources redundancies. Other models may be used for $\mathbf{V}$ including image compression schemes as in [6].

The source model being given, two main elements are still missing. First, how the parameters θ_s are estimated and second, how they are quantized so as to yield a transmitted *quantized source model* $\bar{\theta}_s$. Both problems are considered in Section 3.

2.3. Mixing model

Many studies in source separation model the image of a source $\tilde{s}_m$ as produced through *convolutive* mixing, which means that there are K filters $\tilde{a}_{km}$ such that $\tilde{y}_{km}(n) = (\tilde{a}_{km} * \tilde{s}_m)(n)$ where $*$ denotes convolution. Provided the mixing filters are sufficiently short, this expression can be cast into the STFT domain as:

$$y_{km}(\omega,t) \approx a_{km}(\omega)\, s_m(\omega,t), \tag{2}$$

where $a_{km}(\omega)$ is the frequency response of filter $\tilde{a}_{km}$ at frequency bin ω. Let $\mathbf{a}_m(\omega) = [a_{1m}(\omega),\ldots,a_{Km}(\omega)]^\top$ and let $\mathbf{A}(\omega) = [\mathbf{a}_1(\omega),\ldots,\mathbf{a}_M(\omega)]$ be the mixing matrix at frequency bin ω. The mixture is then supposed to be the sum of the images:

$$\mathbf{x}(\omega,t) = \sum_{m=1}^{M} \mathbf{y}_m(\omega,t) + \boldsymbol{\epsilon}(\omega,t), \tag{3}$$

where $\boldsymbol{\epsilon}$ is a complex $K \times 1$ additive white Gaussian term which accounts for both model and mixing noise and which is supposed to be distributed as follows:

$$\boldsymbol{\epsilon}(\omega,t) \sim \mathcal{N}_c\left(0, \text{diag}\boldsymbol{\sigma}^2(\omega)\right), \tag{4}$$

where $\boldsymbol{\sigma}^2(\omega) = \left[\sigma_1^2(\omega), \ldots, \sigma_K^2(\omega)\right]^\top$ and $\mathrm{diag}\boldsymbol{\sigma}^2(\omega)$ is a diagonal matrix whose diagonal coefficients are given by $\boldsymbol{\sigma}^2(\omega)$. Let $\theta_m = \left\{\left\{\mathbf{A}(\omega)\right\}_\omega, \left\{\boldsymbol{\sigma}^2(\omega)\right\}_\omega\right\}$ be the set of all *mixing parameters*. Combining (2), (3), (4) and defining $\mathbf{C}_{ss}(\omega,t) = \mathrm{diag}\left[v_{1,\omega,t}, \ldots, v_{M,\omega,t}\right]$, we get

$$\mathbf{x}(\omega,t) \mid \theta_s\theta_m \sim \mathcal{N}_c\left(0, \mathbf{C}_{xx}(\omega,t)\right),$$

where

$$\mathbf{C}_{xx}(\omega,t) = \mathbf{A}(\omega)\,\mathbf{C}_{ss}(\omega,t)\,\mathbf{A}^H(\omega) + \mathrm{diag}\boldsymbol{\sigma}^2(\omega) \quad (5)$$

is the *prior covariance matrix* of the mixture.

2.4. A posteriori distribution

Now, assume the $K \times 1$ mixture $\mathbf{x}(\omega,t)$ is available as well as the parameters $\theta = \{\theta_s, \theta_m\}$ of the LSGP formalism as defined above. We consider the case where the original M sources are to be recovered. For some TF bin (ω,t), we focus on the distribution $p(\mathbf{s}(\omega,t) \mid \mathbf{x}(\omega,t), \theta)$ of $\mathbf{s}(\omega,t)$ given $\mathbf{x}(\omega,t)$ and θ, which summarizes what is known about $\mathbf{s}(\omega,t)$ after observation of $\mathbf{x}(\omega,t)$ and knowledge of θ. First, the joint distribution of $\mathbf{s}(\omega,t)$ and $\mathbf{x}(\omega,t)$ given θ writes:

$$\mathcal{N}_c\left(0, \begin{bmatrix} \mathbf{C}_{ss}(\omega,t) & \mathbf{C}_{ss}(\omega,t)\,\mathbf{A}^H(\omega) \\ \mathbf{A}(\omega)\,\mathbf{C}_{ss}(\omega,t) & \mathbf{C}_{xx}(\omega,t) \end{bmatrix}\right). \quad (6)$$

Then, the distribution of $\mathbf{s}(\omega,t)$ given $\mathbf{x}(\omega,t)$ and θ is obtained through conditioning of this joint distribution to yield:

$$\mathbf{s}(\omega,t) \mid \mathbf{x}(\omega,t), \theta \sim \mathcal{N}_c\left(\boldsymbol{\mu}_{\text{post}}(\omega,t), \mathbf{C}_{\text{post}}(\omega,t)\right), \quad (7)$$

with

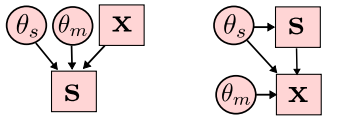
$$\begin{aligned} \mathbf{G}(\omega,t) &= \mathbf{C}_{ss}(\omega,t)\,\mathbf{A}^H(\omega)\,\mathbf{C}_{xx}(\omega,t)^{-1} \\ \boldsymbol{\mu}_{\text{post}}(\omega,t) &= \mathbf{G}(\omega,t)\,\mathbf{x}(\omega,t) \\ \mathbf{C}_{\text{post}}(\omega,t) &= \mathbf{C}_{ss}(\omega,t) - \mathbf{G}(\omega,t)\,\mathbf{A}(\omega)\,\mathbf{C}_{ss}(\omega,t). \end{aligned}$$

3. MODEL ESTIMATION AND ENCODING

3.1. Model estimation

At the encoder, we suppose that both the sources $\mathbf{s}(\omega,t)$ and the mixture $\mathbf{x}(\omega,t)$ are available. $\mathbf{s}(\omega,t)$ is to be encoded using the distribution $p(\mathbf{s} \mid \mathbf{x}, \theta)$ given by (7). However, the model parameters θ that will be transmitted need to be estimated first. In a Bayesian paradigm, they may be chosen as those maximizing (7) when both $\mathbf{s}$ and $\mathbf{x}$ are known, leading to a *discriminative* approach as depicted in Fig. 1 (a). Indeed, such a choice produces θ that maximize the *a posteriori* probability of the signals to be recovered given the mixture and thus leads to the minimal required bitrate for encoding.

Still, such a discriminative model learning is hard to handle using the parameterization of $p(\mathbf{s} \mid \mathbf{x}, \theta)$ given in (7). Another solution is to maximize $p(\mathbf{s}, \mathbf{x} \mid \theta)$ instead, leading to a



(a) Discriminative (b) Generative

Fig. 1. Two different approaches for model learning.

generative approach, as was done in [8, 6] and as is depicted in Fig. 1 (b). Basically, the generative approach permits to estimate the parameters θ which best model the *generation* of the data, instead of focusing on the best parameters for the *estimation of the sources* given the mixture. Even if it leads to suboptimal results in terms of encoding, the generative approach has an appealing advantage of tractability.

First, θ_s was trained through maximization of $p(\mathbf{s} \mid \theta_s)$. In many studies, it was demonstrated that learning θ_s in the Gaussian case is equivalent to the minimization of the Itakura-Saito (IS) divergence between $|\mathbf{S}|^2 = \left\{|s_m(\omega,t)|^2\right\}_{m,\omega,t}$ and $\mathbf{V}$. In the NTF source model (1) considered here, learning can hence be done through decomposition of $|\mathbf{S}|^2$ using the IS divergence as a cost function as in [6].

Second, supposing that $\bar{\theta}_s$, $\mathbf{s}$ and $\mathbf{x}$ are known, θ_m was trained through maximization of $p\left(\mathbf{x} \mid \mathbf{s}, \bar{\theta}_s, \theta_m\right)$. This is achieved through the EM algorithm presented in [1] with the difference that some quantities in the ISS case are kept fixed, e.g. $\mathbf{s}$ and the quantized source model $\bar{\theta}_s$. This algorithm yields the maximum likelihood estimate θ_m for the mixing parameters.

3.2. Model encoding

When θ has been estimated, it is to be quantized in order to form a quantized model $\bar{\theta} = \left\{\bar{\theta}_s, \bar{\theta}_m\right\}$. Using some approximations [6, 8], it can be shown[1] that a quantization of the source model θ_s minimizing the squared error between $\log v$ and its quantized version maximizes the likelihood. For the NTF source model (1), this leads to uniform quantization of $\log \mathbf{W}$, $\log \mathbf{H}$ and $\log \mathbf{Q}$ using step-sizes respectively proportional to $\sqrt{N_\omega}$, $\sqrt{N_t}$ and $\sqrt{M}$. The resulting indices are Huffman encoded as in [7].

4. SOURCE ENCODING

In our previous work on ISS [6], we simply considered the source estimates to be given by their a posteriori mean. This strategy may also be followed here by estimating the sources $\mathbf{s}$ at the decoder as the a posteriori mean given by (7). Indeed, as this distribution is Gaussian, this leads to the Minimum Mean Squared-Error (MMSE) estimate given $\mathbf{x}$ and θ.

Still, as highlighted in [8], this scheme can be significantly improved when one considers the *source encoding* of $\mathbf{s}(\omega,t)$

[1] Due to page limitation, this derivation is left for a longer study.

Algorithm 1 Coding-based ISS for multichannel mixtures.

For all TF bins (ω, t):

1. Compute $\boldsymbol{\mu}_{\text{post}}$ and $\mathbf{C}_{\text{post}}$ as in (7).
2. Compute $\mathbf{C}_{\text{post}} = \mathbf{U}\text{diag}\left[\lambda_1, \ldots, \lambda_M\right]\mathbf{U}^H$.
3. Compute $\mathbf{z} = \mathbf{U}^H\left(\mathbf{s}\left(\omega, t\right) - \boldsymbol{\mu}_{\text{post}}\right)$.
4. Quantize each dimension m of $\mathbf{z}$ using two uniform quantizers of step-size $\frac{\Delta_s}{2}$ for the real and imaginary parts of $\mathbf{z}_m$, to yield quantized $\bar{\mathbf{z}}_m$. Using an arithmetic coder as an entropy coder [11], the effective codeword length (in bits) is given by:
$$-\sum_{m=1}^{M}\log_2\int_{\substack{\text{re}\left(z-\bar{\mathbf{z}}_m\right)\leq\frac{\Delta_s}{2}\\ \text{im}\left(z-\bar{\mathbf{z}}_m\right)\leq\frac{\Delta_s}{2}}}\mathcal{N}_c\left(z\mid 0,\lambda_m\right)dz$$
5. Quantized vector $\bar{\mathbf{s}}\left(\omega, t\right)$ can be reconstructed through
$$\bar{\mathbf{s}}\left(\omega, t\right) = \mathbf{U}\bar{\mathbf{z}} + \boldsymbol{\mu}_{\text{post}}.$$

using distribution (7) instead. This Coding-based ISS scheme has several advantages. First, the quality of the estimates is no longer bounded by Oracle estimators. Second, it is more efficient than usual source coding using only prior distributions, because it makes use of the mixture. Finally, recent advances in source coding [11] can be straightforwardly used instead of having to rely on ad-hoc techniques for the transmission of the residuals as done in [2].

More specifically, the sources $\mathbf{s}\left(\omega, t\right)$ are encoded using model-based constrained entropy quantization based on scalar quantization in the mean-removed Karhunen-Loeve Transform (KLT) as described in [11]. For some particular TF bin (ω, t), let $\mathbf{C}_{\text{post}}$ be the posterior covariance matrix as given in (7) and let $\mathbf{C}_{\text{post}} = \mathbf{U}\text{diag}\left[\lambda_1, \ldots, \lambda_M\right]\mathbf{U}^H$ be its eigenvalue decomposition. $\mathbf{C}_{\text{post}}$ being positive definite, $\lambda_m \in \mathbb{R}_+$. $\mathbf{U}^H\mathbf{s}\left(\omega, t\right)$ is the KLT of $\mathbf{s}\left(\omega, t\right)$ given $\mathbf{x}\left(\omega, t\right)$. Assuming the MSE distortion, uniform quantization is asymptotically optimal for the constrained entropy case [3]. Thus, we consider uniform scalar quantization of $\mathbf{s}\left(\omega, t\right)$ with a fixed source step-size Δ_s in the mean-removed KLT domain, which is summarized in Alg. 1.

Recent advances in source coding [5] may be used to demonstrate on theoretical grounds why it is much more efficient to allocate some bitrate to source coding than to increase the quality of the model as experimentally verified in [8]. If no bitrate is allocated to source quantization, CISS becomes equivalent to classical ISS, i.e. the source estimates coincide with $\boldsymbol{\mu}_{\text{post}}$ as given in (7). This latter scheme is called MMSE-ISS in the following.

5. EXPERIMENTS

The proposed method was evaluated on a set of 14 excerpts sampled at 44.1kHz of professionally produced recordings for which all constituent sources are available. Each excerpt is approximately 30s long and composed of 5 to 10 sources. Two mixing scenarios were considered: linear instantaneous and convolutive mixtures using short Head Related Transfer Function filters of order 200.

The metrics considered for evaluation are the Signal to Distortion Ratio (SDR) of BSSEval [10] between original and estimated sources as well as the Perceptual Similarity Measure (PSM) of PEMO-Q [4]. Both metrics are intended to be related to perceptual quality of the estimates, but SDR is mostly used in the source separation community, while PSM is more common in the coding community.

CISS and MMSE-ISS were run at various levels of quality, corresponding respectively to different choices for the source quantization step-size Δ_s in Alg. 1 for CISS and to different number of components R for MMSE-ISS. As done in [6], all results are compared to those of the oracle MMSE estimate, obtained as the mean given by (7) using $|\mathbf{S}|^2$ instead of $\mathbf{V}$. This permits to compare metrics across different excerpts.

For a given excerpt and a given quality, the estimated audio objects were first compared to the original. Second, the obtained SDR and PSM scores were averaged so as to obtain the corresponding metric for this excerpt and quality. Third, the metrics obtained by the oracle estimate on the same excerpt were subtracted so as to obtain the differential metric δSDR $(\text{excerpt}, \text{bitrate})$ and δPSM $(\text{excerpt}, \text{bitrate})$. Finally, for a given method and each metric, the δ of all excerpts were merged together and the scatter plots $(\text{bitrate}, \delta)$ are displayed in Fig. 2 for both instantaneous and convolutive mixtures.

As can be seen, CISS outperforms MMSE-ISS for both the SDR and PSM metrics. Most noticeably, the performance of CISS is seen not to be bounded by oracle performance but to consistently increase with the bitrate. Still, the proposed method does not yet include a perceptual model, which explains why the SDR score benefits more than PSM from the source coding strategy. Indeed, it is close to a squared-loss criterion. However, CISS permits to easily include perceptual coding through a further perceptual weighting in Alg. 1, expressed directly on the source signals. We are currently investigating this point.

6. CONCLUSION

In this study, we extended recent work on coding-based informed source separation to multichannel mixtures. Such an extension allows recovering the original sources for convolutive mixing processes. Furthermore, the framework we propose is compatible with any compression technique applied on the spectrograms of the mixture for source modeling. In

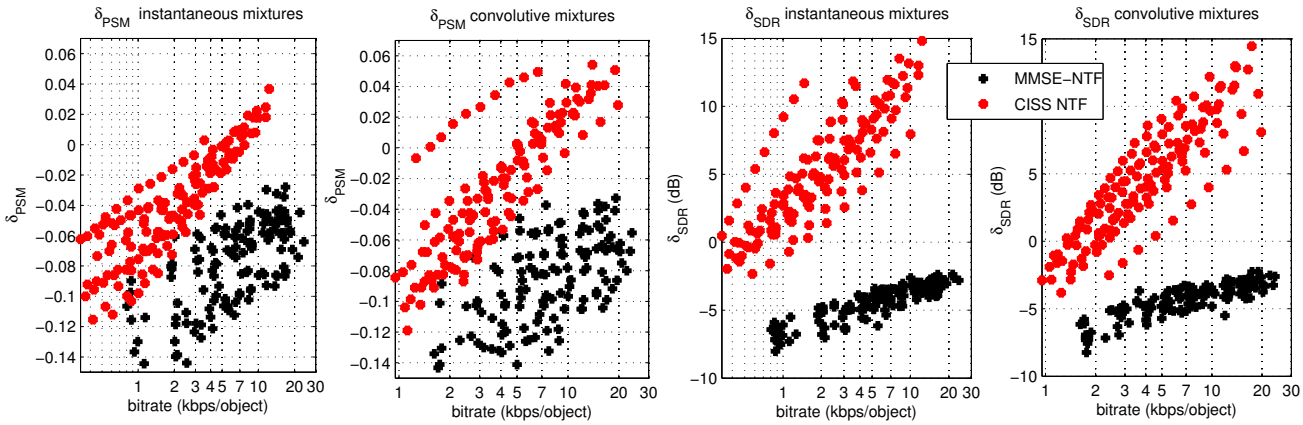


Fig. 2. Rate-SDR and rate-PSM curves for the proposed CISS and MMSE-ISS schemes using NTF as a source model for both instantaneous and convolutive mixtures.

this study, we made use of the recent Nonnegative Tensor Factorization model, which efficiently exploits long-term as well as inter-sources redundancies.

The use of a coding-based strategy to encode the signals permits to consistently increase the quality of separation when more bitrate is available and our experiences have shown that it is often most efficient to use this approach than to spend bitrate in better model parameters, as is predicted by the theory. Current work focuses on better source models and the use of perceptual weighting. Both can easily be included in the proposed framework.

7. REFERENCES

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

[2] 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.

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

[4] 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.

[5] 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)*, pages 243–246, New Paltz, NY, Oct. 2007.

[6] 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.

[7] 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, October 2011.

[8] 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, October 2011.

[9] 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.

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

[11] D.Y. Zhao, J. Samuelsson, and M. Nilsson. On entropy-constrained vector quantization using Gaussian mixture models. *IEEE Transactions on Communications*, 56(12):2094–2104, 2008.