INFORMED SOURCE SEPARATION: SOURCE CODING MEETS SOURCE SEPARATION

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

[1]INRIA, Centre de Rennes - Bretagne Atlantique, Campus de Beaulieu, F-35042 Rennes Cedex, France
[2]Institut Telecom, Telecom ParisTech, CNRS LTCI, 37-39, rue Dareau, 75014 Paris, France
`alexey.ozerov@inria.fr, firstname.lastname@telecom-paristech.fr`

ABSTRACT

We consider the informed source separation (ISS) problem where, given the sources and the mixtures, any kind of *side-information* can be computed during a so-called *encoding* stage. This side-information is then used to assist source separation, given the mixtures only, at the so-called *decoding* stage. State of the art ISS approaches do not really consider ISS as a coding problem and rely on some purely source separation-inspired strategies, leading to performances that can at best reach those of oracle estimators. On the other hand, classical source coding strategies are not optimal either, since they do not benefit from the mixture availability. We introduce a general probabilistic framework called coding-based ISS (CISS) that consists in quantizing the sources using some posterior source distribution from those usually used in probabilistic model-based source separation. CISS benefits from both source coding, thanks to the source quantization, and source separation, thanks to the use of the posterior distribution that depends on the mixture. Our experiments show that CISS based on a particular model considerably outperforms for all rates both the conventional ISS approach and the source coding approach based on the same model.

Index Terms— Informed source separation, source coding, constrained entropy quantization, probabilistic model.

1. INTRODUCTION

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 [1]. However, the problem remains challenging in the undetermined setting ($I < J$), including the single-channel case ($I = 1$), and for convolutive mixtures. Finally, it is also quite clear now that 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 [2]. Motivated by this observation a new setting called *informed source separation (ISS)* [3, 4, 5, 6] was recently considered, where both the sources and the mixtures are assumed known during a so-called *encoding* stage. This knowledge enables the computation of any kind of *side-information* that should be small and should help the source separation at the so-called *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 [5] or just kept

aside. ISS has numerous applications including, e.g., active remixing, gaming, etc.

Several approaches were proposed for the ISS problem [3, 4, 5], and a common point of these methods is that they all rely on some source model θ transmitted as a side-information. Assuming the sources to be sparse in a given time-frequency (TF) representation, Parvaix *et al.* [4] construct a model θ that for each TF point includes the indices of the sources supposed active in this TF point. A TF molecular dictionary is used as model θ in [3]. Liutkus *et al.* [5] go beyond the sparsity assumption, that is hardly verified for real-world mixtures, and rather consider probabilistic models θ such as local Gaussian models (LGM) [1, 2] with structured or free variances.

Note that the ISS problem stands in between source separation [1, 2] and source coding [7, 8, 9], since the sources are available at the encoding stage, as in source coding, and the mixtures are available at both the encoding and the decoding stages, as in source separation. However, to the best of our knowledge, none of the state of the art ISS methods fully benefits from this double knowledge. Indeed:

1. The performances of *source separation* and most of *conventional ISS* methods, depending on the underlying models and assumptions, are bounded by those of oracle estimators [10]. The best (the minimal) achievable distortion produced by conventional ISS methods [4, 5] is incompressible, i.e., it is bounded below. This remark does not concern [3], 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. Figure 1 gives a simplified interpretation of several model-based methods applied to a mixture of two sources in one TF point, TF indices being omitted (see figure's legend for details about notations). Note from Figure 1 (top, left) that the estimated sources $\hat{\mathbf{s}}$ reconstructed as maximum of the *a posteriori* distribution $p(\mathbf{s}|x, \theta)$ can in general never reach the true source values $\mathbf{s}^*$ whatever the precision of the model θ. At the same time, with an efficient source coding strategy the distortion should always go down with increasing rate [7, 8] (see Fig. 1 (top, right)).

2. *Source coding* methods are usually based on a source *a priori* distribution that can be also described by some probabilistic model θ [8, 9]. As mentioned above, the distortion is unbounded below and can be optimally governed by designing an appropriate quantizer. However, the knowledge of the mixture x is not exploited, which leads to a significant overhead in the rate. Indeed, source coding alone would spend an extra rate for codewords lying far away from the mixing equation hyperplane $x = s_1 + s_2$ (see Fig. 1 (top, right)),

while it is known that the data of interest lie on this hyper-plane or close to it in the case of a noisy mixture (see Fig. 1 (top, left)).

A hybrid approach was proposed in [6], where 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 abovementioned drawbacks that are still valid for individual sources.

In this work we introduce a general probabilistic framework for ISS called *coding-based ISS (CISS)* that allows to overcome the limitations of the state-of-the-art methods mentioned above. This approach consists in quantizing the sources, as in source coding, while using the *a posteriori* source distribution $p(\mathbf{s}|x, \theta)$, as in source separation (see Fig. 1, bottom). That way, CISS allows both the distortion to be unbounded below as in source coding, and a decreased rate as in source separation, thanks to the use of the mixing equation. To derive practical adaptive quantizers relying on the *a posteriori* distribution, we use probabilistic model-based quantization under high-rate theory assumptions (see, e.g., [8, 9]). Finally, it should be noted that the goal of ISS is close to that of the spatial audio object coding (SAOC) [11]. However, to the best of our knowledge, such a probabilistic framework was not yet proposed for the SAOC.

This paper is organized as follows. Section 2 introduces CISS in a very general manner. A particular CISS scheme for single-channel mixtures based on the local Gaussian model is described in details and analyzed in section 3. Experimental results are presented in section 4 and the conclusions are drawn in section 5.

2. CODING-BASED INFORMED SOURCE SEPARATION

Figure 2 gives a high-level representation of the 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 $\mathbf{x}$. $\hat{\theta}$ is then encoded and transmitted as a side-information yielding its quantized version $\bar{\theta}$. This encoding can optionally use the knowledge of the mixtures $\mathbf{x}$. Finally, using the posterior $p(\mathbf{s}|\mathbf{x}, \bar{\theta})$ the sources $\mathbf{s}$ are encoded and transmitted as a side-information. At the decoding stage, the model parameter $\bar{\theta}$ and then the quantized sources $\hat{\mathbf{s}}$ are reconstructed.

Note that both the conventional ISS methods [3, 5] and model-based source coding approaches [8, 9] are just partial cases of this general scheme. Indeed, this scheme reduces to conventional ISS when the sources are not encoded but simply reconstructed from the posterior $p(\mathbf{s}|\mathbf{x}, \bar{\theta})$, e.g., by maximizing it [5], and this scheme reduces to model-based source coding when the posterior $p(\mathbf{s}|\mathbf{x}, \bar{\theta})$ is replaced by some prior distribution $p(\mathbf{s}|\bar{\theta})$.

3. CISS WITH LOCAL GAUSSIAN MODEL

We here investigate the proposed approach in the case of single-channel mixtures ($I = 1$) using the local Gaussian models (LGM), as in [2, 5]. However, the approach is more general and not restricted to this particular case.

All the signals are represented in the modified discrete cosine transform (MDCT) domain, since the MDCT is usually used for coding thanks to its orthogonality and the fact that it defines a critically sampled filterbank. In the MDCT domain the mixing equation
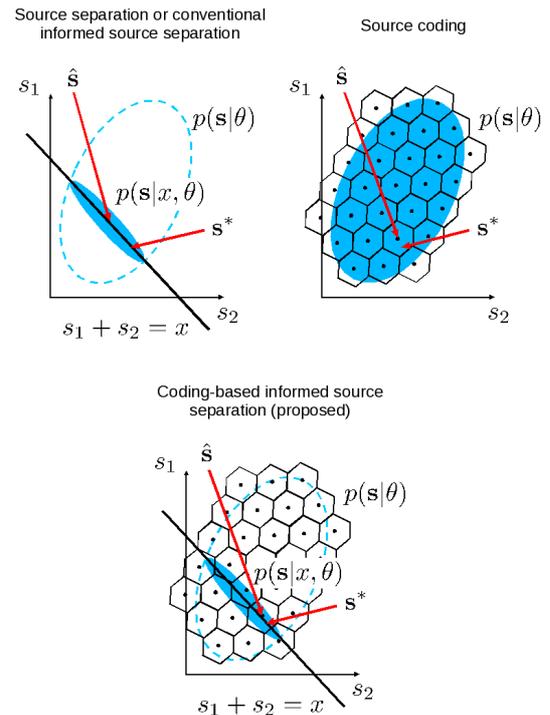

Figure 1: Simplified visualization of the following probabilistic model-based methods applied in one TF point: conventional ISS (top, left), source coding (top, right) and the proposed coding-based ISS (CISS). Notations: x: mixture, $\mathbf{s} = [s_1, s_2]^T$: sources, $p(\mathbf{s}|\theta)$: *a priori* source distribution, $p(\mathbf{s}|x, \theta)$: *a posteriori* source distribution, $\mathbf{s}^*$: true sources, $\hat{\mathbf{s}}$: estimated sources.

writes

$$x_{fn} = \sum_{j=1}^{J} s_{jfn} + b_{fn}, \qquad (1)$$

where $j = 1, \ldots, J$, $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote, respectively, the source index, the MDCT frequency index and the MDCT time-frame index; and x_{fn}, s_{jfn} and b_{fn} denote, respectively, the MDCT coefficients of the mixture, of the sources and of an additive noise representing, e.g., a background or a quantization noise.

3.1. Local Gaussian model

The source and noise coefficients s_{jfn} and b_{fn} are assumed mutually independent, i.e., over j, f and n, and distributed as follows [2, 5]:

$$s_{jfn} \sim \mathcal{N}(0, v_{jfn}), \qquad b_{fn} \sim \mathcal{N}(0, \sigma_b^2), \qquad (2)$$

where the noise variance σ_b^2 is assumed to be known and fixed. This model can be parameterized as $\theta = \{\{v_{jfn}\}_{j,f,n}, \sigma_b^2\}$.

Let $\mathbf{s}_{fn} = [s_{1fn}, \ldots, s_{Jfn}]^T$ be a vector of sources corresponding to the same MDCT coefficient (f, n), its prior and posterior distributions write, respectively, as [2]

$$p(\mathbf{s}_{fn}|\theta) = N\left(\mathbf{s}_{fn}; \boldsymbol{\mu}_{fn}^{\mathrm{pr}}, \boldsymbol{\Sigma}_{\mathbf{s},fn}^{\mathrm{pr}}\right), \qquad (3)$$

$$p(\mathbf{s}_{fn}|x_{fn}; \theta) = N\left(\mathbf{s}_{fn}; \boldsymbol{\mu}_{fn}^{\mathrm{pst}}, \boldsymbol{\Sigma}_{\mathbf{s},fn}^{\mathrm{pst}}\right), \qquad (4)$$

where $N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the probability density function (pdf) of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$;
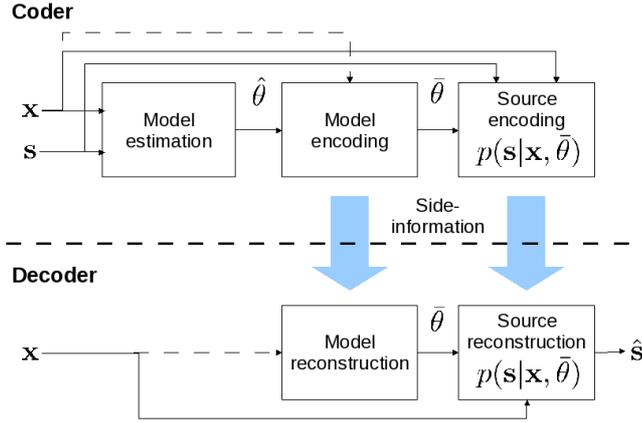
Coder

Figure 2: Coding-based informed source separation scheme.

and

$$\mathbf{\Sigma}_{\mathbf{s},fn}^{\mathrm{pr}} = \mathrm{diag}\left[\{v_{jfn}\}_j\right], \qquad \boldsymbol{\mu}_{fn}^{\mathrm{pr}} = 0, \qquad (5)$$

$$\mathbf{\Sigma}_{\mathbf{s},fn}^{\mathrm{pst}} = \left(\mathbf{I}_J - \mathbf{g}_{fn}\mathbf{1}_J\right)\mathbf{\Sigma}_{\mathbf{s},fn}^{\mathrm{pr}}, \qquad (6)$$

$$\boldsymbol{\mu}_{fn}^{\mathrm{pst}} = \mathbf{g}_{fn}x_{fn}, \qquad (7)$$

$$\mathbf{g}_{fn} = \mathbf{\Sigma}_{\mathbf{s},fn}^{\mathrm{pr}}\mathbf{1}_J^T \left(\mathbf{1}_J\mathbf{\Sigma}_{\mathbf{s},fn}^{\mathrm{pr}}\mathbf{1}_J^T + \sigma_b^2\right)^{-1}, \qquad (8)$$

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

3.2. Source encoding and reconstruction

Each source vector $\mathbf{s}_{fn}$, given its posterior distribution specified by (4) (for CISS) or its prior distribution specified by (3) (for source coding) $f(\mathbf{s}_{fn}|x_{fn};\theta) = N\left(\mathbf{s}_{fn};\boldsymbol{\mu}_{fn},\mathbf{\Sigma}_{\mathbf{s},fn}\right)$, is encoded using model-based constrained entropy quantization relying on scalar quantization in the mean-removed Karhunen-Loeve transform (KLT) domain, as described in [9] and summarized below.

Let $\mathbf{\Sigma}_{\mathbf{s},fn} = \mathbf{U}_{fn}\mathbf{\Lambda}_{fn}\mathbf{U}_{fn}^T$ be the eigenvalue decomposition of the covariance matrix, where $\mathbf{U}_{fn}$ is an orthogonal matrix ($\mathbf{U}_{fn}^T\mathbf{U}_{fn} = \mathbf{I}_J$) and $\mathbf{\Lambda}_{fn} = \mathrm{diag}\{\lambda_{1fn},\ldots,\lambda_{Jfn}\}$ is a diagonal matrix of eigenvalues. The linear transform $\mathbf{U}_{fn}^T$ decorrelating $\mathbf{s}_{fn}$ is the KLT. Assuming the mean squared error (MSE) distortion, uniform quantization is asymptotically optimal for the constrained entropy case [7]. 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

$$\mathbf{y}_{fn} = \mathbf{U}_{fn}^T(\mathbf{s}_{fn} - \boldsymbol{\mu}_{fn}). \qquad (9)$$

2. Quantize each dimension $\mathbf{y}_{fn} = [y_{1fn},\ldots,y_{Jfn}]$ with a uniform scalar quantizer $Q_\Delta : y_{jfn} \to \hat{y}_{jfn}$ having a constant step size Δ. Using an arithmetic coder as an entropy coder [9], the effective codeword length (in bits) is given by

$$L(\mathbf{s}_{fn}|x_{fn};\theta) = -\sum_{j=1}^J \log_2 \int_{\hat{y}_{jfn}-\Delta/2}^{\hat{y}_{jfn}+\Delta/2} N(y;0,\lambda_{jfn})dy.$$

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

$$\hat{\mathbf{s}}_{fn} = \mathbf{U}_{fn}\hat{\mathbf{y}}_{fn} + \boldsymbol{\mu}_{fn}. \qquad (10)$$

3.3. Rate-distortion relations for high rates

Let us consider the source coding (SC) scheme and the CISS scheme described above. It can be shown [8] that under high-rate theory assumptions the total rate R_{tot} (in bits) relates to the mean distortion $D = \mathbb{E}[|\hat{s}_{jfn} - s_{jfn}|^2] = C_s\Delta^2$ (per dimension), respectively, for these two schemes, as follows:

$$R_{\mathrm{tot}}^{\mathrm{SC}} = R(\bar{\theta}) - \frac{JFN}{2}\log_2\frac{D^{\mathrm{SC}}}{C_s} - \log_2 p(\mathbf{s}|\bar{\theta}), \qquad (11)$$

$$R_{\mathrm{tot}}^{\mathrm{CISS}} = R(\bar{\theta}) - \frac{JFN}{2}\log_2\frac{D^{\mathrm{CISS}}}{C_s} - \log_2 p(\mathbf{s}|\mathbf{x},\bar{\theta}) \quad (12)$$

where $C_s = 1/12$ is the coefficient of scalar quantization and $R(\bar{\theta})$ denotes the rate required to encode the model parameter.

3.4. Model estimation and encoding

In this subsection, by analyzing rate-distortion relations (11) and (12), we figure out how the LGM parameters θ should be estimated and how they should be quantized (see Fig. 2). To simplify this analysis in the case of CISS we are using (11) for both source coding and CISS. While it is not exact, it leads to a reasonable and satisfactory approximation. Following derivations from [8], applied here to the LGM instead of the autoregressive model considered in [8], one can show (these derivations are omitted here and will be included in a longer paper on CISS) that

1. The model should be estimated in the maximum likelihood sense, and we simply have $\hat{v}_{jfn} = |s_{jfn}|^2$.

2. Model variances $\hat{v}_{jfn}$ should be quantized so as to minimize the MSE of their logarithms.

These results are quite similar with what was done in [5], where the log-spectrograms were compressed using the JPEG image coder. However, while [5] does not justify this particular choice, we provide here a theoretical explanation of its appropriateness.

4. EXPERIMENTS

We here present a "proof of concept" evaluation of CISS on a single-channel mixture of five synchronized music sources: bass, chorus, drums, guitar and vocals. These signals together with coding results are available from our demo web page at www.irisa.fr/metiss/ozerov/ciss_demo.html.

We compare the following three coding schemes that can be seen as particular instances of the CISS scheme on Figure 2:

1. *Conventional ISS:* All the rate is spent to encode the model parameter $\hat{\theta}$, and the sources are reconstructed via Wiener filtering (7). This scheme is very similar to the JPEG-based scheme presented in [5].

2. *Source Coding:* Sources are encoded using prior distribution $p(\mathbf{s}|\bar{\theta})$ (3) instead of the posterior one $p(\mathbf{s}|\mathbf{x},\bar{\theta})$ (4). Note that this source coding scheme is certainly not an efficient one, and it should not be comparable with the state-of-the-art audio source coders. It is only considered here to demonstrate the advantage of CISS over source coding using the same parametric model θ.

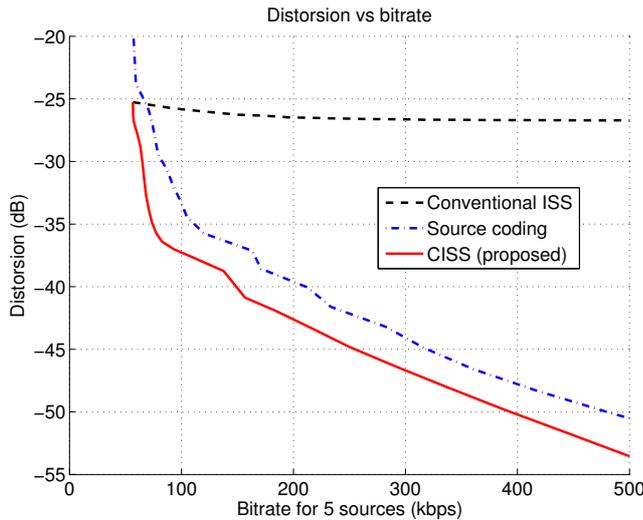3. *CISS:* The scheme of Figure 2, where both the model and the sources are encoded with non-zero rates.

Figure 3: Rate-distortion performance of the conventional ISS scheme (dashed line), the source coding scheme (dash-dot line) and the CISS scheme (solid line).

For all schemes considered, the logarithms of model variances were quantized uniformly and then the resulting images (every quantized variance corresponds to a pixel color in the corresponding image) were encoded using the JPEG lossless coder. Redundancy of the audio sources is thus exploited at this level. In order to improve the efficiency of the JPEG lossless coder, the model variances were thresholded, so that $\hat{v}_{jfn}$ from section 3.4 is chosen as $\hat{v}_{jfn} = \max\left(|s_{jfn}|^2, 10^{-5}\right)$. This leads to better rates, while introducing some interferences from other sources for low rates. Future work will focus on this issue, by designing better models for the sources and including perceptual models.

For the source coding and the CISS schemes, it is known from [8] that under high-rate theory assumptions the optimal rate needed for model encoding is constant and independent of the overall rate. However, since we consider here any rate (low and high), the rate allocation between model and sources was optimized for every distortion specified by source quantization step size Δ. No such optimization is needed for the conventional ISS scheme, since all available rate is spent for model encoding.

Simulation results are shown on Figure 3. Note that for all total rates, the model rate (needed to transmit $\bar{\theta}$) was about 60 kbps. As expected, the distortion of conventional ISS is bounded below. Source coding performs worse than the conventional ISS at low rates (below 70 kilobit per second (kbps) for five sources) and outperforms it for high rates. CISS outperforms both conventional ISS and source coding for all rates with 100 kbps advantage in rate, as compared to source coding, at high rates. Finally, source coding and CISS reach their asymptotic high-rate behaviors predicted by equations (11) and (12) at 400 kbps and 800 kbps respectively. This is not reflected on Fig. 3. At this regime the advantage in rate of CISS, as compared to source coding, is about 250 kbps. Of course, these comparisons apply for the MSE only: a more thorough evaluation will include listening tests in future work, when perceptual models will be considered.

5. CONCLUSION

We have introduced coding-based ISS (CISS), a new general probabilistic framework for informed source separation (ISS), that takes advantages from both source coding and source separation. A preliminary experimental investigation of CISS with a particular source model has shown the advantages of this approach, as compared to both conventional ISS and source coding methods based on the same model. Note also that this probabilistic framework is not restricted to ISS, and can be used to encode any signal **s** conditionally on some other signal **x** correlated with **s**. For example, the approach can be used to encode one or several remixes, given the original recording, or in the context of the parametric stereo coding, where the goal is to encode a stereo recording, given its mono downmix.

Further research will include the following directions. First, more advanced audio-specific structured source models, such as the nonnegative matrix factorization of spectrograms [5] and its extensions [2] should be investigated. Second, new criteria and algorithms for model estimation and encoding that directly optimize the rate-distortion relation (12) should be proposed. Third, CISS should be investigated in the case of multichannel mixtures. Finally, to enhance the perceived sound quality, perceptual models, as in audio coding, should be applied.

6. REFERENCES

[1] 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.

[2] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, submitted.

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

[4] 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*, vol. 19, no. 6, pp. 1721 – 1733, 2011.

[5] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, submitted.

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

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

[8] 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, NY, Oct. 2007, pp. 243–246.

[9] D. Zhao, J. Samuelsson, and M. Nilsson, "On entropy-constrained vector quantization using Gaussian mixture models," *IEEE Transactions on Communications*, vol. 56, no. 12, pp. 2094–2104, 2008.

[10] 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.

[11] 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 AES Convention*, Amsterdam, Netherlands, May 2008.