Perceptual coding-based informed source separation

Serap Kırbız, Alexey Ozerov, Antoine Liutkus, Laurent Girin

PERCEPTUAL CODING-BASED INFORMED SOURCE SEPARATION

Serap Kırbız[1,2] *Alexey Ozerov*[3] *Antoine Liutkus*[4] *Laurent Girin*[1]

[1] GIPSA-Lab and University of Grenoble, Grenoble, France
[2] MEF University, Istanbul, Turkey
[3] Technicolor, Rennes, France
[4]Inria, CNRS, Loria, Speech Team, UMR 7503 Villers-lès-Nancy, France

ABSTRACT

Informed Source Separation (ISS) techniques enable manipulation of the source signals that compose an audio mixture, based on a coder-decoder configuration. Provided the source signals are known at the encoder, a low-bitrate side-information is sent to the decoder and permits to achieve efficient source separation. Recent research has focused on a Coding-based ISS framework, which has an advantage to encode the desired audio objects, while exploiting their mixture in an information-theoretic framework. Here, we show how the perceptual quality of the separated sources can be improved by inserting perceptual source coding techniques in this framework, achieving a continuum of optimal bitrate-perceptual distortion trade-offs.

***Index Terms*—** Informed source separation, source coding, perceptual models

1. INTRODUCTION

Active listening of music and audio consists in interacting with the different audio objects that compose a mixture signal. This goes from generalized karaoke to respatialization and audio scene content modification. Since only the mixture signal is usually available at the user's side, signal processing techniques must be used to recover the individual audio objects from the mix. However, blind or semi-blind methods [1] are still not efficient enough on complex mixtures to be embedded in large audience products, and solutions based on a coder-decoder configuration have been recently proposed. Provided the source signals are known at the encoder, a low-bitrate side-information is sent to the decoder – in addition to the mix signal itself – and used to achieve efficient source manipulation. In this line, the MPEG Spatial Audio Object Coding (SAOC) technology has been derived from multichannel spatial coding and makes use of spatial cues [2]. In parallel, the coder-decoder configuration has been introduced in the source separation framework, leading to informed source separation (ISS) systems [3]. In a general manner, the side-information consists of models of mixing parameters and source power spectral densities (PSD) which are used to build demixing filters at the decoder [3].

Although efficient, these methods have their performance bounded by the best estimates that can be provided by the separation method and source model. Hence, a large bitrate spent on the separation parameters does not improve consequently the quality of the estimates. Recently, the ISS problem was reformulated in [4] so as to consistently benefit from additional bitrate as in source coding [5, 6]. Coding-based ISS (CISS) was proposed, based on a probabilistic model as in source coding. However, instead of using a *prior* distribution which does not make use of the mixture, CISS encodes the sources relying on their *posterior* distribution given the mixture, whose smaller entropy leads to reduced bitrates. CISS thus provides a flexible scheme that "unifies" source separation and multichannel source coding, and allows for fine and extended tuning of the trade-off between side-information bitrate and separation quality.

Now, CISS can be further improved. In particular, no perceptual considerations have been introduced so far in this framework, though it is well-known that audio coding techniques benefit a lot from the perceptual approach [7]. In the present paper, we propose and experiment a first series of strategies to introduce perceptual models and criteria in the CISS scheme. We report the improvements resulting from those new developments.

This paper is structured as follows. First, we rapidly present the CISS framework in Section 2. Section 3 presents the integration of perceptual models into the CISS framework. Finally, the proposed method is evaluated in Section 4.

2. CODING-BASED ISS

All signals are processed in a Time-Frequency (TF) representation, for instance the Modified Discrete Cosine Transform (MDCT), since it is critically sampled and thus a good candidate for source coding. Let $\mathbf{s}_{\omega t} = [s_{1\omega t} \cdots s_{J\omega t}]^\top$ be the $J \times 1$ column vector gathering all J sources for TF bin (ω, t),

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 Research Council (Grant 278025), the Emergence(s) program from the City of Paris, and LABEX WIFI (Laboratory of Excellence within the French Program "Investments for the Future") under references ANR-10-LABX-24 and ANR-10-IDEX-0001-02 PSL*.

where $\omega \in [1, N_\omega]$ and $t \in [1, N_t]$ respectively denote the frequency and frame index, and $\cdot^\top$ denotes transposition. In the present study, the mixture is assumed to be a linear combination of the sources, with $1 \times J$ mixing vector $\mathbf{A}$:

$$x_{\omega t} = \mathbf{A}\mathbf{s}_{\omega t}, \quad \forall\,(\omega, t)\,. \tag{1}$$

This single-channel case is considered here for simplicity of presentation, but the model can be extended to handle many complex multichannel mixing scenarios [8].

The sources are assumed to be independent and to follow a Local Gaussian Model (LGM) [9]. Under this assumption, all TF bins of $\mathbf{s}_j$ are independent and normally distributed, thus yielding the following *prior distribution*:

$$\mathbf{s}_{\omega t} \sim \mathcal{N}\left(0, \mathbf{C}_{ss,\omega t}\right), \tag{2}$$

where $\mathcal{N}$ denotes the Gaussian distribution and $\mathbf{C}_{ss,\omega t} = \mathrm{diag}\{v_{1\omega t}, \cdots, v_{J\omega t}\}$ is the prior source covariance matrix. The variance $v_{j\omega t}$ of source j at TF bin (ω, t) can be understood as its PSD. The LGM is hence parametrized by the $J \times N_\omega \times N_t$ *sources PSD tensor* $\mathbf{V} = \{V_{j\omega t}\}_{j,\omega,t}$. Assuming that a model $\hat{\mathbf{V}}$ of $\mathbf{V}$ is transmitted to the decoder along with the mixture $\mathbf{x}$ and the mixing coefficients, the (modelled) variance of the mixture at TF bin (ω, t) is

$$\mathbf{C}_{xx,\omega t} = \mathbf{A}\mathbf{C}_{ss,\omega t}\mathbf{A}^\top = \sum\nolimits_{j=1}^{J} A_j^2 \hat{V}_{j\omega t}. \tag{3}$$

It is straightforward to show that the *posterior distribution* of $\mathbf{s}_{\omega t}$ given $x_{\omega t}$, $\hat{\mathbf{V}}$ and $\mathbf{A}$ writes:

$$\mathbf{s}_{\omega t} \mid x_{\omega t}, \hat{\mathbf{V}}, \mathbf{A} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\omega t}, \mathbf{K}_{\omega t}\right), \tag{4}$$

where:

$$\begin{aligned} \boldsymbol{\mu}_{\omega t} &= \mathbf{C}_{ss,\omega t}\mathbf{A}^\top \mathbf{C}_{xx,\omega t}^{-1} x_{\omega t}, \\ \mathbf{K}_{\omega t} &= \mathbf{C}_{ss,\omega t} - \mathbf{C}_{ss,\omega t}\mathbf{A}^\top \mathbf{C}_{xx,\omega t}^{-1}\mathbf{A}\mathbf{C}_{ss,\omega t}. \end{aligned} \tag{5}$$

A first trend of research on ISS [3] has focused on using $\hat{\mathbf{V}}$ and $\mathbf{A}$ at the decoder to build Wiener demixing filters, which exactly corresponds to estimate the sources $\mathbf{s}_{\omega t}$ by their posterior mean $\boldsymbol{\mu}_{\omega t}$ in (5). This leads to a computationally efficient ISS method with good but bounded separation quality. To reach arbitrary distortion on separated sources at the cost of higher bitrate, CISS has been proposed in [4] and has the important advantage of being optimal from an information-theoretical point of view. The main idea of CISS is to apply traditional (lossy) source-coding [5] on $\mathbf{s}_{\omega t}$, with the important difference that the considered distribution is the posterior distribution (4) instead of the classical prior distribution, for instance (2). This difference leads to an important decrease of the source coding bitrate, due to the mutual information between the sources $\mathbf{s}$ and the mixture $\mathbf{x}$ which is available at the decoder. Given this fundamental difference, the practical implementation follows the usual fundamentals of source coding (see [4] for a detailed description). In short, the sources $\mathbf{s}_{\omega t}$ are encoded using model-based (for instance (4)) constrained entropy quantization based on scalar quantization in the mean-removed Karhunen-Loeve Transform (KLT) domain [6]. The user chooses a step-size Δ_s to control the rate-distortion trade-off and Algorithm 1 is applied, while using an arithmetic coder for entropy coding. Finally, the problem of estimating $\hat{\mathbf{V}}$ and transmitting it to the decoder has to be addressed (the cost of transmitting $\mathbf{A}$ is assumed to be negligible). Due to its important number of entries, factorization and compression techniques are required to model and encode $\mathbf{V}$. Previous studies [3, 4, 8] have concentrated on Nonnegative Tensor Factorization (NTF) [10], which drastically reduces the number of parameters by assuming that $\hat{\mathbf{V}}$ is the superposition of a relatively small number K of rank-1 tensors:

$$\hat{V}_{j\omega t} = \sum\nolimits_{k=1}^{K} W_{\omega k} H_{tk} Q_{jk}, \tag{6}$$

where $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{Q}$ are $N_\omega \times K$, $N_t \times K$ and $J \times K$ nonnegative matrices, respectively. Learning these model parameters is achieved through the maximum likelihood (ML) strategy that can be shown equivalent to the following optimization criterion [11]:

$$\{\mathbf{W}^\star, \mathbf{H}^\star, \mathbf{Q}^\star\} = \underset{\{\mathbf{W},\mathbf{H},\mathbf{Q}\}}{\mathrm{argmin}} \sum_{j,\omega,t} d_{IS}\left(V_{j\omega t} \mid \hat{V}_{j\omega t}\right), \tag{7}$$

where d_{IS} is the Itakura-Saito divergence (see, e.g., [11] for a definition). This optimization problem can be tackled with classical NTF multiplicative updating [9], i.e., Algorithm 2 (with $P_{j\omega t} = 1$). Once $\{\mathbf{W}, \mathbf{H}, \mathbf{Q}\}$ are estimated, they are

Algorithm 1 Coding-based ISS for squared error.

For all TF bins (ω, t) (which are omitted here for conciseness):

1. Compute $\boldsymbol{\mu}$ and $\mathbf{K}$ as in (5).
2. Compute the eigenvalue decomposition
$$\mathbf{K} = \mathbf{U}\mathrm{diag}\left\{[\lambda_1, \cdots, \lambda_D]\right\}\mathbf{U}^T,$$
where $\mathbf{U}^T$ is the KLT.
3. Compute $\mathbf{z} = \mathbf{U}^T\left(\mathbf{s} - \boldsymbol{\mu}\right)$.
4. Quantize each dimension of $\mathbf{z}$ using a uniform quantizer of constant step-size Δ_s to yield quantized $\bar{\mathbf{z}}$. Using an arithmetic coder as an entropy coder [6], the effective codeword length (in bits) is given by:
$$-\sum\nolimits_{j=1}^{J} \log_2 \int_{\bar{z}_j - \Delta_s/2}^{\bar{z}_j + \Delta_s/2} N\left(z_j \mid 0, \lambda_j\right) dz_j,$$
where $N\left(\cdot \mid 0, \lambda_j\right)$ is the probability density function of the zero-mean normal distribution with variance λ_j.
5. Quantized source vector is given by $\bar{\mathbf{s}} = \mathbf{U}\bar{\mathbf{z}} + \boldsymbol{\mu}$.

quantized to be transmitted to the decoder. It has been demonstrated in [4] that nearly optimal performance can be achieved by uniform quantization of $\log \mathbf{W}$, $\log \mathbf{H}$ and $\log \mathbf{Q}$. The respective quantization steps are provided in [4]. The resulting indices are entropy encoded, e.g. by a Gaussian Mixture Model-based arithmetic coding as in [4].

3. PERCEPTUAL CISS

It is well known from audio coding that optimizing the Mean Squared Error (MSE) is far to be the best strategy to achieve the best perceptual quality possible at a given bitrate. Instead, audio coding is more efficient when the distortion measure to be minimized between compressed and original data includes some perceptual weighting [12]. Here, we show how this principle can be exploited in the CISS framework to yield an optimal CISS strategy that minimizes both the bitrate *and* the perceptual distortion. Actually, we propose to integrate perceptual models into the CISS scheme either at the NTF model estimation step (Algorithm 2), or at the posterior source coding step (Algorithm 1), or at both steps. The corresponding perceptually motivated CISS-NTF schemes are referred to as CISS-PNTF, PCISS-NTF, and PCISS-PNTF, respectively.

3.1. Perceptual Models and associated Weighting Tensor

Implementing a psychoacoustic model into the CISS framework amounts in defining a perceptual *weighting tensor* $\mathbf{P} = \{P_{j\omega t}\}_{j,\omega,t}$ ($P_{j\omega t} \geq 0$) that gives the perceptual importance of each TF bin (ω, t) of each source j. Depending on the perceptual model considered, the resulting coefficients $P_{j\omega t}$ are computed differently, but all further processing may then be understood as optimizing a perceptually weighted quadratic distortion, which is classical in the audio coding literature. More precisely, while parameter learning presented in section 3.2 below is amenable to a weighted log-likelihood optimization, perceptual source coding is achieved by identifying $\mathbf{P}$ with a sensitivity matrix, as presented in [7, 13].

We first rapidly present the two auditory perceptual models that we tested in this study. The first model, called here the *Par model* was proposed by van de Par *et al.* in [7] and the second model is defined in ITU-R BS.1387 [14].

The Par model [7] is a psychoacoustic masking model based on the peripheral bandpass filtering properties of the human auditory system (HAS). It is based on frequency masking only and does not take into account temporal masking. It evaluates the distortion-to-masker ratio within each auditory filter which can be used for the derivation of a masking threshold and provides a measure for distortion detectability.

In the presence of a source power spectrogram $|s_{j\omega t}|^2$, the weighting tensor is obtained as [13]:

$$P^2_{j\omega t} = \frac{C_s \hat{L}}{N_\omega} \sum_i \frac{|h_\omega|^2 |g^i_\omega|^2}{\frac{1}{N_\omega} \sum_{\omega'} |h_{\omega'}|^2 |g^i_{\omega'}|^2 |s_{j\omega' t}|^2 + C_a}, \quad (8)$$

where h_ω is the transfer function of the outer-middle ear filter, g^i_ω is that of the i-th gamma-tone filter, $\hat{L}$ is the effective duration of the corresponding block, and C_s and C_a are some model calibration constants (see [7, 13] for details).

The ITU-R model recommended in ITU-R BS.1387 is a spectro-temporal model and relies on both spectral and temporal masking [14]. The quantitative significance of an auditory object within a mixture can be measured by its perceived loudness through the HAS. The loudness of one frame is modelled by calculating the excitation using perceptually motivated frequency scale (Bark scale) and critical bandwidth, compressing the excitation, and integrating over frequency [14].

The loudness coefficients $\tilde{P}_{jbt}$ are calculated in each critical band b of ITU-R BS. 1387 model as [14]

$$\tilde{P}_{jbt} = c \left(\frac{E^\tau_b}{\tau_b E_0} \right)^{0.23} \left[\left(1 - \tau_b + \frac{\tau_b \tilde{E}^s_{jbt}}{E^\tau_b} \right)^{0.23} - 1 \right], \quad (9)$$

where τ_b is the threshold index, E^t_b is the excitation threshold, E_0 is a constant and $\tilde{E}^s_{jbt}$ are *excitation patterns* [14].

The weighting coefficients calculated in the critical bands are extended in the linear frequency scale of the STFT such that $P_{j\omega t} = \tilde{P}_{jbt}$, for the frequency components ω lying in the b-th critical band [15]. It can be seen that, loudness is affected by parameters other than sound pressure, including frequency and duration.

3.2. Perceptually weighted model estimation (PNTF)

In this study, instead of the ML approach (7), we propose to estimate the model parameters through a perceptually weighted NTF, which has recently been considered both for multichannel audio coding, and upmixing [10]. Perceptually-weighted NTF amounts to estimate $\{\mathbf{W}, \mathbf{H}, \mathbf{Q}\}$ by minimizing the following cost function instead of (7):

$$\{\mathbf{W}^\star, \mathbf{H}^\star, \mathbf{Q}^\star\} = \underset{\{\mathbf{W},\mathbf{H},\mathbf{Q}\}}{\operatorname{argmin}} \sum_{j\omega t} P_{j\omega t} d_{IS} \left(V_{j\omega t} \mid \hat{V}_{j\omega t} \right). \quad (10)$$

The rationale behind the use of (10) is that all TF bins are not of equal perceptual importance. Hence, focusing on parameters that provide good approximation for perceptually important TF bins may be a good way to optimize performance. The optimization procedure for estimating the parameters through multiplicative updates is given in Algorithm 2.

3.3. Perceptual posterior source coding (PCISS)

Let $\bar{\mathbf{P}} = \left\{ \bar{P}_{j\omega t} \right\}_{j,\omega,t}$ be the weighting tensor coefficients computed as in Section 3.1, its quantized version or its estimation [1], using $\hat{V}_{j\omega t}$ as the sources PSD. To apply this weighting

[1] Note that in contrast to weighed model estimation presented in section 3.2, where the weighting tensor needs only to be computed at the encoder side, for the perceptual posterior source coding the weighting tensor needs to be known on both the encoder and the decoder sides. Thus, the weighting tensor should be either transmitted or estimated somehow. In this work we estimate it by replacing source power spectrograms $|s_{j\omega t}|^2$ (e.g., in (8)) by $|\mu_{j\omega t}|^2$, with $\boldsymbol{\mu}_{\omega t}$ from (5).

Algorithm 2 Weighted-NTF algorithm

- Inputs: $V_{j\omega t} = |s_{j\omega t}|^2$, $P_{j\omega t}$ and $K \in \mathbb{N}$
- Initialize $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{Q}$ randomly.

Iterate several times:

$$Q_{jk} \leftarrow Q_{jk}\left(\frac{\sum_{\omega t} P_{j\omega t} W_{\omega k} H_{tk} V_{j\omega t} \hat{V}_{j\omega t}^{-2}}{\sum_{\omega t} P_{j\omega t} W_{\omega k} H_{tk} \hat{V}_{j\omega t}^{-1}}\right), \quad (11)$$

$$W_{\omega k} \leftarrow W_{\omega k}\left(\frac{\sum_{jt} P_{j\omega t} H_{tk} Q_{jk} V_{j\omega t} \hat{V}_{j\omega t}^{-2}}{\sum_{jt} P_{j\omega t} H_{tk} Q_{jk} \hat{V}_{j\omega t}^{-1}}\right), \quad (12)$$

$$H_{tk} \leftarrow H_{tk}\left(\frac{\sum_{j\omega} P_{j\omega t} W_{\omega k} Q_{jk} V_{j\omega t} \hat{V}_{j\omega t}^{-2}}{\sum_{j\omega} P_{j\omega t} W_{\omega k} Q_{jk} \hat{V}_{j\omega t}^{-1}}\right). \quad (13)$$

matrices for quantization, i.e., to optimize a weighted distortion instead of the MSE, the entries of distribution (4) should be modified as follows

$$\mathbf{s}'_{\omega t} = \text{diag}\{\bar{\mathbf{P}}_{\omega t}\}\mathbf{s}_{\omega t}, \quad (14)$$

$$\boldsymbol{\mu}'_{\omega t} = \text{diag}\{\bar{\mathbf{P}}_{\omega t}\}\boldsymbol{\mu}_{\omega t}, \quad (15)$$

$$\mathbf{K}'_{\omega t} = \text{diag}\{\bar{\mathbf{P}}_{\omega t}\}\mathbf{K}_{\omega t}\text{diag}\{\bar{\mathbf{P}}_{\omega t}\}^{\top}, \quad (16)$$

where $\bar{\mathbf{P}}_{\omega t} = \left[\bar{P}_{1\omega t}, \ldots, \bar{P}_{J\omega t}\right]^{\top}$. After this modification the source vectors should be quantized exactly as described in Algorithm 1. The quantized source vectors will be weighted. Thus, the unweighted coefficients should be obtained by $\bar{\mathbf{s}}_{\omega t} = \text{diag}\{\bar{\mathbf{P}}_{\omega t}\}^{-1}\bar{\mathbf{s}}'_{\omega t}$.

4. EXPERIMENTS

In this section, we evaluate the performance of the proposed perceptually motivated CISS-NTF method for two different auditory models. For this purpose, a set of 14 excerpts sampled at 44.1kHz from the QUASI database [2] is used. Each excerpt is a linear instantaneous mixture of 5 to 10 sources with a duration of approximately 30s long.

The performance is evaluated using the Normalized Signal to Distortion Ratio (NSDR, in dB) and the Normalized Perceptual Similarity Measure (NPSM, between 0 and 1) which measure the improvement of the SDR [16] of BSSEval and PSM of PEMO-Q [17], respectively, over a "do nothing separation", where the mixture itself is considered as a source estimate. NSDR and NPSM are calculated as

$$\text{NSDR}_j(\hat{\tilde{\mathbf{s}}}_j, \tilde{\mathbf{s}}_j, \tilde{\mathbf{x}}) = \text{SDR}_j(\hat{\tilde{\mathbf{s}}}_j, \tilde{\mathbf{s}}_j) - \text{SDR}_j(\tilde{\mathbf{x}}, \tilde{\mathbf{s}}_j), \quad (17)$$

$$\text{NPSM}_j(\hat{\tilde{\mathbf{s}}}_j, \tilde{\mathbf{s}}_j, \tilde{\mathbf{x}}) = \text{PSM}_j(\hat{\tilde{\mathbf{s}}}_j, \tilde{\mathbf{s}}_j) - \text{PSM}_j(\tilde{\mathbf{x}}, \tilde{\mathbf{s}}_j), \quad (18)$$

where $\tilde{\mathbf{s}}_j, \hat{\tilde{\mathbf{s}}}_j$ and $\tilde{\mathbf{x}}$ denote the original source signal, estimated source signal and the mixture signal in time domain, respectively.

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

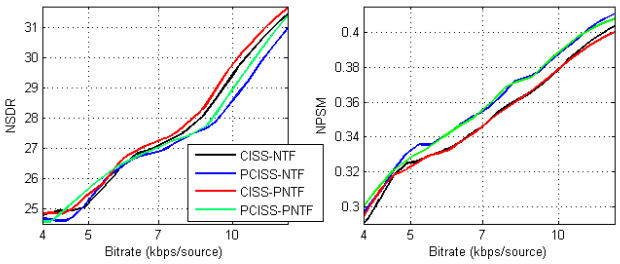

Fig. 1. Performance of the proposed perceptual CISS-NTF scheme.

The NSDR and NPSM values are averaged over different recordings. CISS-NTF, PCISS-NTF, CISS-PNTF and PCISS-PNTF were run at various levels of quality, corresponding respectively to different choices for the source quantization step-size Δ_s in Algorithm 1 for CISS. The number of components is fixed to three for each source, i.e., $K = 3J$. Performance of the proposed perceptual CISS-NTF schemes are reported together with the baseline CISS-NTF [4].

The simulations produced many (rate, NSDR) and (rate, NPSM) pairs. Then, for each small range of rates, the pairs corresponding to the highest NSDR and NPSM are selected. The resulting (rate, NSDR) and (rate, NPSM) planes were then smoothed using the locally weighted scatterplot smoothing (LOESS) method to produce the rate/performance curves. The results can be found in Fig. 1 and audio excerpts may be downloaded on the webpage dedicated to this paper [3].

Although we evaluated the performance using both Par and ITU-R models for extracting the weighting tensor, we reported the results corresponding to the higher NPSM. As such we report the results with the ITU-R model for PNTF and with the Par model for PCISS. If we use the loudness matrix of ITU-R model for weighting the distortion, it forces the louder components to have less distortion, while the distortion in the regions with low energy is increased. Thus, we observe an improvement in terms of NSDR around 0.5 dB for CISS-PNTF and PCISS-PNTF compared to CISS-NTF and PCISS-NTF, respectively. Therefore, applying a perceptual weight on the NTF scheme leads to an improvement in terms of NSDR. At the same time, it keeps the NPSM scores unchanged, which is a bit deceiving. On the other hand, we can see that PCISS-NTF and PCISS-PNTF are outperformed by CISS-NTF and CISS-PNTF in terms of NSDR, especially for bitrates above 6.5 kbps, up to approximately 0.5 dB. This is trivial and predictable since NSDR is optimized with the MSE minimization obtained by the CISS scheme. The important point here is that the NPSM scores are improved consistently by approximately 0.01–0.02 by using PCISS with the Par model.

[3] http://www.gipsa-lab.grenoble-inp.fr/~laurent.girin/demo/PCISS-EUSIPCO2014.zip

5. CONCLUSION

In this study, we have shown how perceptual models can be included in coding-based informed source separation (CISS), so as to yield an information-theoretical optimal way to transmit audio objects when their mixtures are known at the decoder. This work extends the state of the art about CISS that previously only considered squared-error as a distortion measure.

Including perceptual weighting in informed source separation was achieved from two different perspectives. First, we showed how the estimation of the Nonnegative Tensor Factorization source model can benefit from perceptual weighting at no cost in terms of bitrate. Second, we have demonstrated that recent results on source coding using perceptual distortions could be extended to the case of posterior source-coding, i.e., when both the coder and the decoder share some side-information.

This approach is the first Audio Object Coding approach that permits to include perceptual constraints on the audio sources to be transmitted through their downmix. Indeed, even if other approaches from the state of the art do rely on residual perceptual coding [2], the perceptual models they consider are not relative to the audio objects themselves, but rather to the errors performed during estimation, which is different and suboptimal. We have demonstrated through our evaluation that the system we propose permits to recover isolated sources that are of better perceptual quality with a bitrate of only a few kilobits per seconds and per source.

REFERENCES

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

[4] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 8, pp. 1699 – 1712, Aug. 2013.

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

[6] D.Y. 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.

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

[8] A. Liutkus, R. Badeau, and G. Richard, "Low bitrate informed source separation of realistic mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 26-31 May 2013.

[9] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155 –3167, July 2011.

[10] 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.

[11] 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.

[12] A. Ozerov and W. B. Kleijn, "Flexible quantization of audio and speech based on the autoregressive model," in *IEEE Asilomar Conference on Signals, Systems, and Computers (Asilomar CSSC'07)*, Pacific Grove, CA, Nov. 2007.

[13] P. Petkov, "The sensitivity matrix for a spectral auditory model," M.S. thesis, KTH (Royal Institute of Technology), May 2005.

[14] ITU-R Recommendation BS.1387, *Method for Objective Measurements of Perceived Audio Quality*, December 1998.

[15] S. Kırbız and B. Günsel, "Perceptually enhanced blind single-channel music source separation by non-negative matrix factorization," *Digital Signal Processing*, vol. 23, no. 2, pp. 646–658, Mar. 2013.

[16] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[17] 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*, vol. 14, no. 6, pp. 1902 –1911, Nov. 2006.