QUANTIZATION-AWARE PARAMETER ESTIMATION FOR AUDIO UPMIXING

Christian Rohlfing[1] *Antoine Liutkus*[2] *Julian M. Becker*[1]

[1]Institut für Nachrichtentechnik, RWTH Aachen University, Germany
[2]Inria, speech processing team, Villers-lès-Nancy, France

ABSTRACT

Upmixing consists in extracting audio objects out of their downmix, given some parameters computed beforehand at a coding stage. It is an important task in audio processing with many applications in the entertainment industry. One particularly successful approach for this purpose is to compress the audio objects through nonnegative matrix factorization (NMF) parameters at the coder, to be used for separating the downmix at the decoder. In this paper, we focus on such NMF methods for audio compression, which operate at very low parameter bitrates. In existing methods, parameter estimation and quantization are conducted independently. Here, we propose two extensions: first, we jointly estimate and quantize the parameters at the coder to ensure good reconstruction at the decoder. Second, we propose a parameter refinement method operated at the decoder, that benefits from priors induced by quantization to yield better performance. We show that our contributions outperform existing baseline methods.

***Index Terms*—source separation, upmixing, NMF, quantization, audio object coding**

I. INTRODUCTION

Audio upmixing is the signal processing task that aims at generating a multichannel signal based on a downmix [1], [2]. It is an important topic in audio processing, because it allows the decomposition of downmixes into several audio objects, enabling numerous applications in the entertainment industry, such as adaptive rendering on loudspeakers arrays, karaoke or active listening.

From an audio coding perspective, upmixing can be considered a classical idea originating from the topic of Spatial Audio Coding (SAC [3], [4]). In this context, a good strategy appears as transmitting multichannel signals through both a downmix — meaning less channels to encode — and a few upmixing parameters enabling good reconstruction of all channels at the decoder. This upmixing strategy occurs as a core building block of recent spatial audio coding standards such as SAOC [5] or MPEG-H [6].

Independently from the audio coding community, upmixing was also early understood in the signal processing literature as a particular case of *source separation*, which aims at decomposing signals into additive components [7]. The main difference with SAC is that audio separation does not usually assume that the components to recover are known at any encoding stage, but only the mixture and some general assumptions such as stereophonic [8] or spectral diversity [9], [10], [11].

Bridging the audio coding and the source separation communities, PARVAIX et al. proposed in [12], [13] to consider the extreme case where source separation could be achieved by parameters learned from the original audio objects to recover. This scenario was coined in as Informed Source Separation (ISS) and gave rise to an important research effort in the following years [14], [15], [16]. The conceptual similarities between ISS and SAOC are very strong, but it took some years to realize that two communities were in fact addressing the same exact problem with different methodologies. The gap was filled when the theory of source coding was applied in the particular setting of ISS to yield the Coding-based ISS (CISS) framework [17], [18], [19].

Although very good systems are available at bitrates around 5 kbps/object with CISS, designing very low bitrate ISS/SAOC working at bitrates close to or under 1 kbps/object remains a challenge. As demonstrated in [14], existing systems fall short at providing efficient solutions in this regime. In this respect, the baseline method [20], [21], [22] exploits a Nonnegative Tensor Factorization model (NTF [23], [24]) that proves effective in concisely encoding the power spectral densities of the sources, to be used for Wiener filtering at the decoder with a bitrate of around 1 kbps/object. A particularity of these systems is that the computations at the decoder are very simple and can be achieved in real time. However, it was shown in recent studies [25], [26] that whenever the decoder has some available computing resources, the bitrate could be dramatically reduced again. The idea here is to proceed to classical blind separation at the decoder, using only very crude binary versions of the optimal parameters as an initialization provided by the coder. The resulting system is capable of reaching bitrates as low as 0.5 kbps/object, with reasonable performance.

In this study, we go further in this direction of enhanced compression for audio upmixing/ISS. The main idea of this paper is to consider the quantization of the parameters already in the design of their estimation strategy. In practice, this is achieved by including further constraints for the NTF. Such constraints already have a rich history and were used to yield e.g. sparse [27] or smooth decompositions [28], [29]. Here, our contributions are twofold. First, we propose a *self-quantizing* constraint for the NTF that leads to parameters that both account for the signals while being maximally quantized already. This provides consistent increases of performance compared to posterior quantization applied after the learning, as done in most ISS studies [20], [21], [18]. As in [25], our second contribution exploits the computing capabilities of the decoder to fit the parameters again, on the mixture only, before applying upmixing. Instead of simply considering an initialization as in [25], we exploit quantization again and propose another new constraint called *quantized-matching.* Its idea is to fit NTF parameters at the decoder so that they best describe the mixture, additionally making sure that *their quantized version* matches the quantized parameters transmitted by the coder. As we advocate, this is more efficient than a simple initialization for which the parameters are free to strongly deviate from what we know was a good value at the coder. The result is an improvement in upmixing quality at no cost in bitrate.

This paper is structured as follows. In Section II, we present the baseline ISS systems [21], [25] on which we improve. In Section III, we present the self-quantizing and quantization-matching cost functions we are proposing for NTF. Finally, we evaluate the impact of using them for upmixing in Section IV.

II. PARAMETRIC AUDIO UPMIXING

II-A. Notations and general architecture

In this study, the complex Time-Frequency Representations (TFR) of the J sources and of the mixture are denoted $s_j(f,t)$ and $x(f,t) = \sum_j s_j(f,t)$, respectively. The sources are taken as

Cauchy-harmonizable processes [30], which is a generalization of the classical Gaussian case [31]:

$$s_j\left(f,t\right) \sim \mathcal{C}_c\left(P_j\left(f,t\right)\right),$$

where $\mathcal{C}_c$ is called the complex isotropic Cauchy distribution [32], [33]. $P_j\left(f,t\right)$ is called here the Magnitude Spectral Density (MSD) of source j at Time-Frequency (TF) bin $\left(f,t\right)$. It is a nonnegative quantity accounting for the scale of source j at that TF bin.

Under this model and given all the MSDs P_j, it can be shown that each source j can be estimated using its posterior expectation given the mixture through 1-Wiener filtering [30] as:

$$\widehat{s}_j\left(f,t\right) \leftarrow \mathbb{E}\left[s_j\left(f,t\right) \mid x, P_j\right] = \frac{P_j\left(f,t\right)}{\sum_{j'} P_{j'}\left(f,t\right)} x\left(f,t\right) \ . \quad (1)$$

Just like their Gaussian Power Spectral Density (PSD) counterparts, the MSDs are theoretical objects never observed in practice. Given s_j and x, we define the 1-spectrograms V_j and V_x as empirical estimates, that can roughly be understood as equal in average to P_j and P_x, respectively [34]:

$$V_j\left(f,t\right) \triangleq \left|s_j\left(f,t\right)\right| \approx P_j\left(f,t\right) \quad \text{and}$$
$$V_x\left(f,t\right) \triangleq \left|x\left(f,t\right)\right| \approx P_x\left(f,t\right),$$

where $\triangleq$ stands for a definition. Setting an nonnegative tensor factorization (NTF) parameter on the source MSDs, we write:

$$P_j\left(f,t \mid \Theta\right) \triangleq \sum_{k=1}^{K} W\left(f,k\right) H\left(t,k\right) Q\left(j,k\right), \quad (2)$$

where W, H, and Q are $F \times K$, $T \times K$ and $J \times K$ nonnegative matrices, respectively, all gathered under the general notation $\Theta = \{W, H, Q\}$. Depending on the strategy used for estimation, Θ can bear subscripts or additional decorations as in $\bar{\Theta}$. In any case, these all pertain to NTF parameters.

Then, given some NTF parameters Θ, the 1-Wiener filter (1) becomes:

$$\widehat{s}_j\left(f,t\right) \leftarrow \mathbb{E}\left[s_j\left(f,t\right) \mid x, \Theta\right] = \frac{P_j\left(f,t \mid \Theta\right)}{\sum_{j'} P_{j'}\left(f,t \mid \Theta\right)} x\left(f,t\right), \quad (3)$$

where the MSD used is the one given in (2).

The baseline ISS method based on NTF [21] consists of two stages: In the *encoder*, the sources s_j are perfectly known and used for computing compact parameters Θ_s. These parameters are then transmitted to the *decoder* in a quantized form $\bar{\Theta}_s$. At the decoder-side only the downmix x is available. The sources are estimated as $\mathbb{E}\left[s_j\left(f,t\right) \mid x, \bar{\Theta}_s\right]$. A recent variation over this scheme was introduced in [25] and is used here as a second baseline. The NTF parameters $\bar{\Theta}_s$ received at the decoder are refined to recover from the very coarse quantization of Θ_s by exploiting the mixture only.

II-B. Parameters estimation at coder

Both baseline algorithms [21], [25] use NTF learning on the source spectrogram V_j to get the source NTF parameters Θ_s. Here, we use NTF with multiplicative update rules minimizing the β-divergence between the spectrograms V_j and their approximation (3):

$$d_\beta\left(V_j \mid \Theta_s\right) \triangleq \sum_{f,t} d_\beta\left(V_j\left(f,t\right) \mid P_j\left(f,t \mid \Theta_s\right)\right) \ . \quad (4)$$

The β-divergence includes e.g. Itakura-Saito distance ($\beta = 0$), Kullback-Leibler divergence ($\beta = 1$) and Euclidean distance ($\beta = 2$). The gradient of the reconstruction cost $d_\beta\left(V_j \mid \Theta_s\right)$ with respect to one NTF source parameter, e.g. W_s, can be expressed as follows

$$\nabla_{W_s} d_\beta\left(V_j \mid \Theta_s\right) = \nabla^{+}_{W_s} d_\beta\left(V_j \mid \Theta_s\right) - \nabla^{-}_{W_s} d_\beta\left(V_j \mid \Theta_s\right) \quad (5)$$

with $\nabla^{+}_{W_s} d_\beta\left(V_j \mid \Theta_s\right)$ and $\nabla^{-}_{W_s} d_\beta\left(V_j \mid \Theta_s\right)$ both nonnegative terms. The corresponding multiplicative update rule for W_s then depends on these positive and negative gradient terms:

$$W_s \leftarrow W_s \cdot \frac{\nabla^{-}_{W_s} d_\beta\left(V_j \mid \Theta_s\right)}{\nabla^{+}_{W_s} d_\beta\left(V_j \mid \Theta_s\right)} \ . \quad (6)$$

The same derivation can be conducted for the other NTF parameters H_s and Q_s. The update rules for all parameters are given in detail e.g. in [21]. Note that the overall NTF performance strongly depends on the choices of the initial parameters.

The quantization of the source parameters is conducted after parameter estimation by NTF in the logarithmic domain as proposed in e.g. [18], [21]:

$$\bar{W}_s \triangleq \exp\left(q\left(\log W_s\right)\right) \quad (7)$$

with all operations performed element-wise. We use scalar quantization $q\left(\cdot\right)$ on each element of the NTF parameters independently (see Section III-A). The quantized versions all parameters, $\bar{H}_s$ and $\bar{Q}_s$, are obtained the same way.

II-C. Parameters (re-)estimation and upmixing at decoder

The NTF parameters Θ_s describing the sources are quantized and transmitted to the decoder as $\bar{\Theta}_s$. The decoder of the first baseline algorithm [21] simply estimates the sources as $\mathbb{E}\left[s_j\left(f,t\right) \mid x, \bar{\Theta}_s\right]$ in (3). The second baseline scheme [25] proposes an additional so called *mix NTF* step occurring at the decoder with only the mix spectrogram V_x as observation. In this setup, the transmitted parameters $\bar{\Theta}_s$ are not used for Wiener filtering directly but only for initialization of the mix NTF instead. The rationale of the method is that parameters initialized close to their true values should converge to the right solution while correctly describing the mix.

This procedure yields new NTF parameters Θ_x at the decoder (not to be confused with the source parameters Θ_s) that are used to recover the sources as $\mathbb{E}\left[s_j\left(f,t\right) \mid x, \Theta_x\right]$. Note that this allows for very coarse quantization of $\bar{\Theta}_s$ leading to very low bitrates. Refinements include allocating additional bitrate to the difference between Θ_s and Θ_x or to information guiding the mixture NTF as in [26]. These refinements are not considered in the present study.

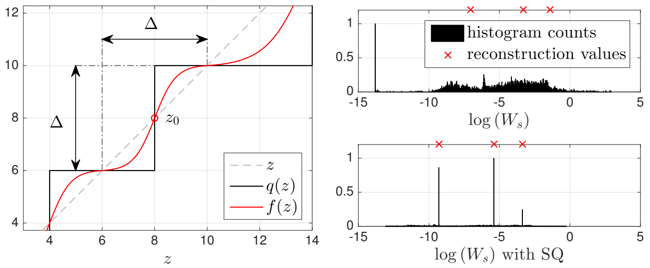
III. QUANTIZATION-AWARE ESTIMATION

In Section III-A, we propose a derivable approximation to the scalar quantization curve. In Section III-B, we propose a constraint on the source NTF allowing simultaneous parameter estimation and quantization. In Section III-C, we propose a constraint on the mix NTF which prevents the estimated parameters from deteriorating from the quantized initialization parameters.

III-A. Quantization and a derivable approximation

For decreasing the parameter bitrate, quantization of the NTF parameters is needed. As previously done in NTF-based ISS algorithms e.q. [20], [21], [22], [25], we use scalar quantization of each entry of the NTF parameter matrices. Scalar quantization maps continuous scalar values z to a discrete set of N values, here denoted as reconstruction values $q\left(z\right)$. The reconstruction values are obtained in this paper by the Lloyd-Max algorithm (refer e.g. to [35]) minimizing the squared error between the continuous values z and the reconstruction values $q\left(z\right)$. This procedure results in N non–uniformly spaced reconstruction values. The distance between two reconstruction values is denoted with the step size Δ (which differs for different pairs of reconstruction values) as shown for an exemplary quantization curve $q\left(z\right)$ in Fig. 1a.

For incorporating quantization in the NTF update rules, we need an approximation of $q(z)$ by a derivable function. Here we use a logistic function $f_0\left(z\right)$ to approximate $q\left(z\right)$ *between two*



(a) Derivable approximation $f(\cdot)$ of scalar quantization $q(\cdot)$.

(b) Histograms of $\log(W_s)$ for coder NTF without and with SQ.

Fig. 1: Hard and soft quantization curves as well as histograms (normalized to interval $[0,1]$) for the proposed SQ constraint.

reconstruction values (interval is marked in Fig. 1a with dotted-dashed lines). This yields the *soft quantization curve* $f(z)$

$$f(z) = z_0 + \frac{\Delta}{d}\left[\underbrace{\frac{1}{1+\exp\left(-\lambda\frac{2}{\Delta}(z-z_0)\right)}}_{=f_0(z)} - \frac{1}{2}\right] \tag{8}$$

with midpoint of the logistic function centered between two reconstruction values $z_0 = q(z) + \frac{\Delta}{2}\operatorname{sgn}(z - q(z))$ and steepness parameter λ. The factor $d = \frac{1}{1+\exp(-\lambda)} - \frac{1}{1+\exp(\lambda)}$ scales $f_0(z)$ to ensure continuity at the corner points (lower and upper reconstruction value). The derivative of $f(z)$ can be expressed as

$$\frac{\partial f(z)}{\partial z} = \frac{2\lambda}{d} f_0(z)\left[1 - f_0(z)\right] \ . \tag{9}$$

Fig. 1a shows the scalar quantization curve $q(z)$ as well as our proposed approximation function $f(z)$ with midpoint z_0, steepness parameter $\lambda = 5$ and step size $\Delta = 4$.

III-B. Self-quantization at coder

As explained in Section II-B, the two baseline systems learn NTF parameters Θ_s describing the source spectrograms V_j and quantize the parameters as a posterior step to yield $\bar{\Theta}_s$. In this section, we propose a novel NTF constraint which accounts for both good signal approximation and quantization of the parameters at the same time. We enforce the parameters Θ_s to be close to their quantized version $\bar{\Theta}_s$ at NTF run-time. The overall cost function for the NTF with the proposed *self-quantizing* constraint (SQ) consists of the signal reconstruction term $d_\beta(V_j \mid \Theta_s)$ as given in Eq. (4) as well as the β-divergence between the parameters and their quantized versions $\bar{W}_s$, $\bar{H}_s$ which we assume constant for each NTF iteration

$$\min\ d_\beta(V_j \mid \Theta_s) + \gamma_{\text{sq}}\left[d_\beta\left(\bar{W}_s \mid W_s\right) + d_\beta\left(\bar{H}_s \mid H_s\right)\right] \ . \tag{10}$$

The self-quantizing constraint on W_s and H_s is weighted with scalar factor $\gamma_{\text{sq}} \geq 0$ and added to the signal reconstruction term $d_\beta(V_j \mid \Theta_s)$. Setting $\gamma_{\text{sq}} = 0$ implies an NTF which only accounts for the signal V_j.

As shown in Section II-B in Eq. (5) for the reconstruction term, the gradient of the constraint cost has to be split up into positive and negative terms to yield the corresponding multiplicative update rules accounting for the constraint. These terms corresponding to the SQ constraint for W_s are given in the first column of Table I. The gradient is equivalent for H_s. Q_s has only few elements compared to W_s and H_s, so we quantize Q_s with high resolution after the NTF and thus do not consider it for the self-quantizing constraint.

To show the impact of SQ, Fig. 1b depicts histograms of $\log W_s$ (as we quantize in the logarithmic domain) for two scenarios: The upper plot in Fig. 1b shows the distribution of $\log(W_s)$ for an unconstrained source NTF in the coder as well as quantization reconstruction levels (marked with crosses) of the posterior quantization step. The lower plot shows the distribution for another NTF on the same sources with SQ activated. SQ shifts the parameter values in direction of the reconstruction values as the corresponding distribution is closer to the quantized version compared to unconstrained NTF.

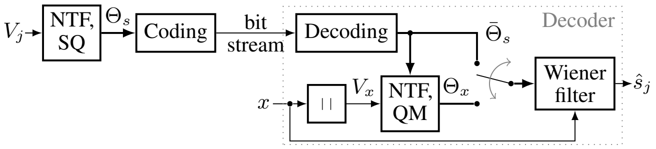


Fig. 2: Block diagram of proposed coder consisting of the decoder.

III-C. Quantized-matching at decoder

The ISS scheme proposed in [25] uses a second NTF on the mixture at the decoder refining the quantized source parameters $\bar{\Theta}_s$. Taking $\bar{\Theta}_s$ as initialization, the decoder NTF may still deviate from the optimal source parameters Θ_s as only few quantization levels for $\bar{\Theta}_s$ are usually used. However, $\bar{\Theta}_s$ is the (coarsely) quantized version of the interference-free parameters Θ_s which the mix NTF should try to get back to. To prevent the mix NTF from deviating too much, we propose to put constraints on the mix parameters Θ_x in the quantization domain: The quantized version of Θ_x shall match $\bar{\Theta}_s$ as much as possible. Constraining Θ_x itself would result in an unnecessary quantization of Θ_x: We want to learn Θ_x with full resolution that gets close to $\bar{\Theta}_s$ only when it is quantized.

To be able to derive NTF update rules for this *quantized-matching* constraint (QM) with regard to one parameter, e.g. W_x, we replace the (hard) quantization curve $q(\cdot)$ as used in (7) by the soft quantization curve $f(\cdot)$ proposed in Section III-A to yield what we call *soft-quantized parameters*

$$\tilde{W}_x \triangleq \exp\left(f(\log W_x)\right) \ . \tag{11}$$

The novel mix NTF cost function then consists of the signal estimation term $d_\beta(V_x \mid \Theta_x)$ as well as the quantized-matching constraint which favors soft quantized parameters being close to the quantized source parameters:

$$\min\ d_\beta(V_x \mid \Theta_x) + \gamma_{\text{qm}}\left[d_\beta\left(\bar{W}_s \mid \tilde{W}_x\right) + d_\beta\left(\bar{H}_s \mid \tilde{H}_x\right)\right] \tag{12}$$

with $\bar{W}_s, \bar{H}_s$ being constant. Again, the constraint is weighted with a scalar factor $\gamma_{\text{qm}} \geq 0$ as done already for Eq. (10). The corresponding positive and negative gradient terms are summarized in the third column of Table I. The terms for QM are dependent on the positive and negative terms for the gradient of the soft-quantized parameter (11) which are given in the second column of Table I. The later terms are derived from (9). Once again, since Q_s is very small, we send it in full resolution and simply set $Q_x = Q_s$.

III-D. Complete system

The complete system is shown in Fig. 2. The basic structure resembles the structure of the ISS method described in [25]. Here we augment the NTFs with our proposed constraints: The self-quantized (SQ) constraint is used for obtaining the source NTF parameters Θ_s in the encoder. The quantized parameters $\bar{\Theta}_s$ are used for initialization of the NTF describing the mixture with parameters Θ_x in the decoder. This mixture model is estimated with quantized-matching (QM) constraint to ensure that quantized Θ_x matches $\bar{\Theta}_s$. The estimated sources $\hat{s}_j$ are obtained by 1-Wiener-filtering mixture x. For comparison with reference method [21], Wiener-filtering can be performed with $\bar{\Theta}_s$ directly instead of Θ_x.

Self-quantizing constraint (coder)	Soft-quantized parameter	Quantized-matching constraint (decoder)
$\nabla^{+}_{W_s} d_\beta \left(\bar{W}_s \mid W_s\right) = W_s^{\beta-1}$	$\nabla^{+}_{W_x} \tilde{W}_x = \frac{2\lambda}{d} f_0 \left(\log W_x\right) \cdot \frac{\tilde{W}_x}{W_x}$	$\nabla^{+}_{W_x} d_\beta \left(\bar{W}_s \mid \tilde{W}_x\right) = \nabla^{+}_{W_x} \tilde{W}_x \cdot \tilde{W}_x^{\beta-1} + \nabla^{-}_{W_x} \tilde{W}_x \cdot \bar{W}_s \cdot \tilde{W}_x^{\beta-2}$
$\nabla^{-}_{W_s} d_\beta \left(\bar{W}_s \mid W_s\right) = \bar{W}_s \cdot W_s^{\beta-2}$	$\nabla^{-}_{W_x} \tilde{W}_x = \frac{2\lambda}{d} f_0^2 \left(\log W_x\right) \cdot \frac{\tilde{W}_x}{W_x}$	$\nabla^{-}_{W_x} d_\beta \left(\bar{W}_s \mid \tilde{W}_x\right) = \nabla^{-}_{W_x} \tilde{W}_x \cdot \tilde{W}_x^{\beta-1} + \nabla^{+}_{W_x} \tilde{W}_x \cdot \bar{W}_s \cdot \tilde{W}_x^{\beta-2}$

Table I: Positive and negative gradient parts for the SQ (Section III-B) and QM (Section III-C) constraints to be used for multiplicative update rules as in Eq. (6). $a \cdot b$ and $\frac{a}{b}$ stand for element-wise multiplication and division of matrices a and b of the same dimension.

IV. EVALUATION

IV-A. Data-set and metrics

For evaluation of the proposed self-quantizing (SQ) and quantized-matching (QM) constraints, we took 10 mixtures consisting of 4 – 7 sources (e.g. vocals, guitar, drums, effects) of the QUASI database[1]. Each mix is sampled at $44100\,\mathrm{kHz}$ and is $30\,\mathrm{s}$ long. Separation performance is given by the signal-to-distortion ratio (SDR, in dB) between original and estimated sources. After taking the mean over the sources, the resulting SDR value is set in reference to the SDR obtained by an oracle estimator [36] which estimates optimal Wiener filter masks. The resulting measure is denoted with $\delta\mathrm{SDR}$. The parameter bitrate R is obtained with GZIP on $\bar{\Theta}_s$ as done in e.g. [21] and measured in kbps per source. Additionally, we measured the reconstruction quality of either the quantized source parameter $\bar{\Theta}_s$ or the refined mix parameters Θ_x by evaluating $d_\beta \left(V_j \mid \Theta_s\right)$ or $d_\beta \left(V_j \mid \Theta_x\right)$.

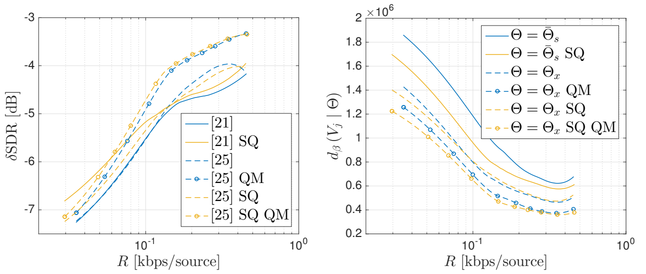
Regarding the Time-Frequency transform, we chose the STFT window size to $93\,\mathrm{ms}$ with $50\,\%$ overlap. The spectral dimension of the spectrograms is filtered with a Mel-filterbank with $F = 400$ Mel-filters [11]. We evaluated the proposed constraints with different numbers of NTF components per source $K/J \in \{1, 2 \ldots 10\}$ with J number of sources and $\beta = 1$ (Kullback–Leibler divergence). We use an SVD-based method [37] to initialize the source NTF. The SQ and QM constraints are either deactivated or activated with scalar weights $\gamma_{\mathrm{sq}}, \gamma_{\mathrm{qm}} \in \left\{10^{-1}, 1, 10, 10^2\right\}$. For QM, the soft quantization steepness is set to $\lambda = 5$. The source parameters are quantized with different numbers of quantization bins $N \in \{2, 3, 4\}$. Note that K/J and N have both strong influence on the overall parameter bitrate R.

For each mixture, each combination of parameters (K/J, N, γ_{sq} and γ_{qm}) results in a $(R, \delta\mathrm{SDR})$-point. These points are then optimized to yield the Pareto front per mixture and finally smoothed using the locally weighted scatter plot smoothing method [38], obtaining rate/quality curves. The same procedure is done for all (R, d_β) points.

IV-B. Experiments and discussion

Fig. 3a shows $(R, \delta\mathrm{SDR})$ curves for the two baseline algorithms and the proposed constraints: Performances of the (unconstrained) reference methods [21] and [25] (solid and dashed blue curves) are compared to the methods with activated SQ and/or QM constraints. As [21] uses solely Wiener filter in the decoder, activating QM is not possible. Several noticeable facts come out of this evaluation:

(i) SQ doesn't prove very useful as soon as mix NTF [25] is enabled. However, it strongly improves performance of [21] for lower bitrates ($> 0.5\,\mathrm{dB}$) and even permits bitrates that are slightly lower. Proceeding jointly to parameter estimation and quantization hence really proves effective in bringing increased performance at no additional cost in bitrate. **(ii)** For extremely low-bitrates, under $0.1\,$kbps/source, SQ permits the very computationally efficient method [21] to outperform the more demanding [25] that requires computations at the decoder. **(iii)** QM significantly improves the performance of mix NTF, at all bitrates. This is a very interesting result because it means this constraint succeeds in bringing Θ_x much closer to Θ_s. For bitrates above $0.1\,$kbps/source, incorporating QM brings a remarkable $1\,\mathrm{dB}$ improvement on performance. Additional SQ gives slightly increased performance again.



(a) $(R, \delta\mathrm{SDR})$ curves.

(b) $\left(R, d_\beta\right)$ curves. $d_\beta \left(V_j \mid \Theta\right)$ calculated with either $\bar{\Theta}_s$ or Θ_x.

Fig. 3: Rate-quality curves for the two baseline methods [21] and [25] in comparison to the proposed SQ and QM constraints. [21] uses quantized source parameters $\bar{\Theta}_s$, [25] refined mix parameters Θ_x for synthesis.

This first evaluation concerned the final quality of the separation result. In a second evaluation, we focused on the reconstruction quality $d_\beta \left(V_j \mid \Theta\right)$ given in Eq. (4) of the source spectrogram V_j using various sets of parameters either $\Theta = \bar{\Theta}_s$ (used in [21]) or $\Theta = \Theta_x$ (used in [25]) obtained by activating or not SQ and QM. Fig. 3b shows the corresponding $\left(R, d_\beta\right)$ curves. Comparing the two solid curves (performance of $\bar{\Theta}_s$ with deactivated and activated SQ) it becomes clear that the quantized source NTF model $\bar{\Theta}_s$ with SQ yields much better reconstruction than using a posterior quantization step ($10\,\%$ smaller d_β-value). This is an important result as it is independent of the ISS setup and usable in any applications where quantized NTF parameters are needed. Then, refining $\bar{\Theta}_s$ by the mix NTF [25] as well as activating QM decreases the β-divergence each time even further, taking now $\Theta = \Theta_x$. This second evaluation suggests that the proposed methods do behave as expected concerning the cost functions that are being minimized. The discrepancies between Figures 3a and 3b thus indicate that it may be appropriate to focus on better cost-functions for fitting the parameters than the Kullback–Leibler divergence used here.

V. CONCLUSION

We proposed two novel constraints for Nonnegative Tensor Factorization (NTF) in an Informed Source Separation (ISS) setup. First, the self-quantizing constraint (SQ) in the ISS encoder leads simultaneously to good signal approximation and quantized parameters. This constraint may be useful whenever NTF is used for signal compression, not only for upmixing. Second, the quantized-matching constraint (QM) in the decoder NTF prevents deviations from the (optimal) source NTF parameters in the quantization domain. Both constraints were evaluated and outperformed reference methods.

Future work could include iterating between encoder and decoder, thus refining the source model given the mix model and vice versa. The influence of weights and parameters on the nonconvex SQ and QM cost functions could be studied as well as the usage of an SQ-constrained NTF in a blind source separation setup.

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

VI. REFERENCES

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

[2] J. Nikunen, *Object-based Modeling of Audio for Coding and Source Separation*, vol. 1276, Tampere University of Technology, Tampere, Finland, 2015, [Available online].

[3] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, "Spatial Audio Coding: Next-Generation Efficient and Compatible Coding of Multichannel Audio," in *Audio Engineering Society Convention 117*, Oct. 2004.

[4] F. Rumsey, *Spatial audio*, CRC Press, 2012.

[5] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, et al., "Spatial audio object coding (SAOC)-the upcoming MPEG standard on parametric object based audio coding," in *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[6] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio – the new standard for universal spatial/3D audio coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, 2015.

[7] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*, Academic Press, 2010.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[9] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2003, pp. 177–180.

[10] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[11] M. Spiertz, *Underdetermined Blind Source Separation for Audio Signals*, vol. 10 of *Aachen Series on Multimedia and Communications Engineering*, Shaker Verlag, Aachen, July 2012, [Available online].

[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 Transactions on Audio, Speech, and Language Processing*, 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 Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1721 –1733, Aug. 2011.

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

[15] S. Zhang, L. Girin, and A. Liutkus, "Informed source separation from compressed mixtures using spatial Wiener filter and quantization noise estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013.

[16] N. Sturmel, L. Daudet, and L. Girin, "Phase-based informed source separation for active listening of music," in *15th International Conference on Digital Audio Effects (DAFx 2012)*, York, United Kingdom, Sept. 2012, p. n/c.

[17] 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)*, New Paltz, New York, USA, Oct. 2011.

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

[19] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial coding-based informed source separation," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2407–2411.

[20] 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.

[21] 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.

[22] 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)*, May 2013.

[23] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley Publishing, Sept. 2009.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sept. 2011.

[25] C. Rohlfing, J. M. Becker, and M. Wien, "NMF-based informed source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 474–478.

[26] C. Rohlfing and J. M. Becker, "Generalized constraints for NMF with application to informed source separation," in *2016 Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*, Aug. 2016, pp. 597–601.

[27] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: Beyond lp-norms," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 858–862.

[28] M. N. Schmidt and H. Laurberg, "Non-negative matrix factorization with Gaussian process priors," *Computational Intelligence and Neuroscience*, vol. ID 361705, 2008.

[29] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[30] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015.

[31] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1830 –1840, Sept. 2010.

[32] G. Samoradnitsky and M. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, vol. 1, CRC Press, 1994.

[33] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy Nonnegative Matrix Factorization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, United States, Oct. 2015.

[34] A. Liutkus, T. Olubanjo, E. Moore, and M. Ghovanloo, "Source Separation for Target Enhancement of Food Intake Acoustics from Noisy Recordings," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, United States, Oct. 2015.

[35] J.-R. Ohm, *Multimedia Signal Coding and Transmission*, Signals and Communication Technology. Springer-Verlag Berlin Heidelberg, 2015.

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

[37] J. M. Becker, M. Menzel, and C. Rohlfing, "Complex SVD initialization for NMF source separation on audio spectrograms," in *Fortschritte der Akustik DAGA '15*, Nürnberg, Germany, Mar. 2015.

[38] W. Cleveland and S. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.