NMF-BASED INFORMED SOURCE SEPARATION

Christian Rohlfing, Julian M. Becker, and Mathias Wien

Institut für Nachrichtentechnik
RWTH Aachen University, Germany
{rohlfing,becker,wien}@ient.rwth-aachen.de

ABSTRACT

Informed Source Separation (ISS) is a topic unifying the research fields of both source separation and source coding. Its main objective is to recover audio objects out of a mixture with a source separation step assisted by a set of compact parameters extracted with complete knowledge of the sources. ISS can be used for applications such as active listening and remixing of music (e.g. karaoke).

In this paper, we propose a new ISS method which includes a semi-blind source separation (SBSS) step in the ISS decoder to decrease the amount of parameter bit rate. SBSS is conducted by factorizing the mixture in time-frequency domain by nonnegative matrix factorization (NMF). The transmitted parameters consist of a compact NMF initialization as well as residuals calculated in the NMF domain. We show in simulations that using SBSS in the decoder increases the separation quality and that our scheme improves the rate-distortion performance in comparison to a state-of-the art method.

Index Terms— Informed source separation, nonnegative matrix factorization, audio object coding

1. INTRODUCTION

Informed source separation (ISS), initially proposed in [1], is a special case of audio source separation and consists of two stages: In the *encoding stage*, the original audio sources are perfectly known and used to extract a compact set of side-information. In the *decoding stage*, this side-information is used to assist a source separation step trying to extract the sources out of the audio mixture which is assumed to be perfectly known at the decoder. This procedure is also closely related to spatial audio object coding (SAOC) [2].

The ISS algorithm in [3] encodes the source spectrograms in the time-frequency (TF) domain with help of nonnegative tensor factorization (NTF) and uses Wiener-like TF masking as source separation in the decoding step. A comparative study over recently developed ISS methods is given in [4]. The coding scheme proposed in [5] uses a modified NTF for audio upmixing which is similar to coding-based ISS (CISS) [6]. CISS tries to unify the two worlds of source separation and coding even further and uses transform coding of the sources in the TF domain and encodes the necessary parameters with NTF. In [7], perceptual modeling into CISS is introduced. Most of the NTF-based methods [3, 5, 8] use solely Wiener-like TF masking as source separation in the decoding step and transmit quantized versions of the corresponding TF masks whereas [6] additionally encodes quantized source residuals.

In this paper, we include a semi-blind source separation (SBSS) algorithm to the ISS decoder to enhance the separation quality for lower bit rates. The separation step is still based on TF masking where the TF masks are obtained by a SBSS algorithm using nonnegative matrix factorization (NMF). The encoder controls this step to maximize the separation quality by aligning the mixture NMF model and the NMF model obtained by separating each source independently. Therefore we name our method "NMF-ISS".

This paper is structured as follows: Section 2 gives a general overview over the proposed method. The decoder and encoder are described in Section 3 and Section 4. Section 5 shows evaluation results and Section 6 concludes this paper.

2. OVERVIEW

The mono-channel audio mixture $\mathbf{x}$ in time domain consists of M sources $\mathbf{s}_m$ such that $\mathbf{x} = \sum_{m=1}^{M} \mathbf{s}_m$. In the following, all signals are denoted in the TF domain: The time-domain signals are transformed by the short-time Fourier transform (STFT) and a subsequent spectral dimension reduction step which filters the spectral dimension of the absolute-valued STFT output with a Mel-filterbank[1] to speed up the following computation steps [9] and decrease the parameter bit rate. The resulting mixture and source magnitude spectrograms are labelled $\mathbf{X}$ and $\mathbf{S}_m$ respectively.

Instead of transmitting a quantized NMF model describing the sources directly as similarly done in e.g. [3], we propose to transmit a compact initialization for an NMF which estimates the NMF model of the mixture and a residual NMF model to enhance the separation quality.

The flowgraph of the NMF-ISS encoder is depicted in Fig. 1: The encoder contains a complete decoder which performs semi-blind source separation (SBSS) explained in Section 3. The mixture $\mathbf{X}$, which is assumed to be perfectly known at the decoder, is separated with NMF into acoustical events also denoted as components. The estimated sources $\tilde{\mathbf{S}}_m$ are obtained by Wiener-like TF masking. To enhance the SBSS performance, the encoder (described in Section 4) aligns the decoder NMF model describing the mixture and an NMF model which describes the sources: The source magnitude spectrograms $\mathbf{S}_m$ are independently separated by NMF to yield the *source NMF model* which results in an approximation of interference-free NMF components in comparison to the mixture NMF model. This model is then used to calculate a compact *initialization* for the decoder NMF. After executing the decoder, the encoder uses the source NMF model again to compute a *residual NMF model* which takes the remaining differences between the source and the decoder NMF model into account and enhances the overall separation quality even further. Therefore, the side-information transmitted to the decoder consists of the NMF initialization and residual model as well as optimal parameters for the SBSS algorithm.

[1] The Mel-filterbank consists of F triangular filters whose central frequencies are spaced linearly on the mel scale $f = 1127 \log(1 + f_{\mathrm{Hz}}/700)$.

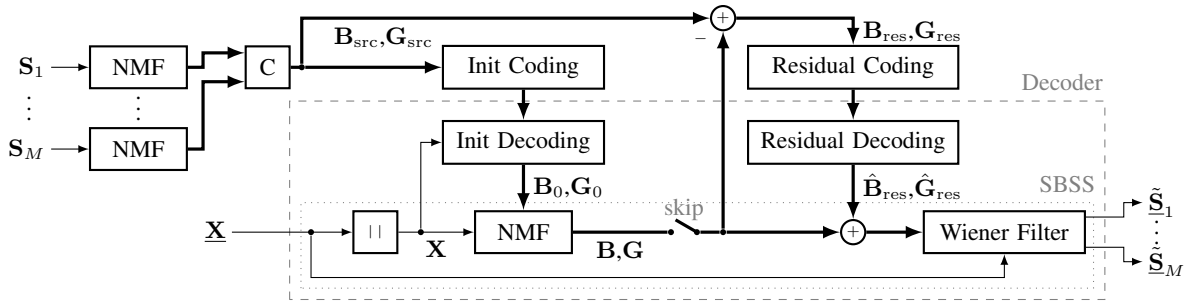


Fig. 1. Block diagram of the proposed NMF-ISS encoder in time-frequency domain. The "C" block concatenates all source NMF matrices in the component dimension to form joint matrices $\mathbf{B}_{\mathrm{src}}$ or $\mathbf{G}_{\mathrm{src}}$. The "Coding" and "Decoding" blocks include quantization and bit coding or the inverse operations respectively. The transmitted side-information consists mainly of a compact NMF model initialization and the residual NMF model to enhance the quality of the semi-blind source separation (SBSS) algorithm.

With this approach, our proposed method becomes flexible: The decoder is able to operate blindly in the case that no side-information is transmitted at all or the encoder could decide to skip the NMF in the decoder such that the residual model would consist of the complete source NMF model as done similarly in [3] (refer to Section 4.3).

3. NMF-BASED SOURCE SEPARATION

The NMF-ISS decoder uses the blind source separation algorithm of [9, 10] which is initialized and refined under knowledge of the sources; therefore we denote the algorithm used here as semi-blind and explain it in the following.

The mixture spectrogram $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ is factorized by NMF into I components

$$\mathbf{X} \approx \mathbf{B}\,\mathbf{G}^{\mathrm{T}}\,, \tag{1}$$

with the spectral basis matrix $\mathbf{B} \in \mathbb{R}_+^{F\times I}$, temporal gain matrix $\mathbf{G} \in \mathbb{R}_+^{T\times I}$, number of spectral (Mel-) bins F and number of time frames T. The number of components I is a user defined parameter. The NMF variant used here, denoted as β-NMF, estimates $\mathbf{B}$ and $\mathbf{G}$ by evaluating multiplicative update rules which are derived by minimizing the β-Divergence between the left and the right hand side of Eq. (1) [9]. Here, the β-Divergence is extended by a widely used constraint which favors temporal continuity of the gain vectors and is weighted with α_{tc} [11]. The constraint is deactivated with $\alpha_{\mathrm{tc}} = 0$.

The estimated complex source spectrograms are reconstructed with Wiener-like TF masking

$$\underline{\tilde{\mathbf{S}}}_m(f,t) = \underline{\mathbf{X}}(f,t) \sum_{\mathbf{r}(i)=m} \mathbf{C}_i(f,t) / \sum_{i'} \mathbf{C}_{i'}(f,t) \tag{2}$$

with component index i, spectral index f, time bin t and $\mathbf{C}_i(f,t) = \mathbf{B}(f,i)\,\mathbf{G}(t,i)$ denoting the spectrogram of the ith component. The grouping assignment $\mathbf{r}(i) \in \mathbb{N}_+^I$ links the components to the corresponding sources[2].

Note that the overall NMF performance is strongly dependent on the choice of parameters I, β, α_{tc} and initial matrices $\mathbf{B}_0$ and $\mathbf{G}_0$ (refer to Section 4.2). I has also a strong impact on the resulting bit rate as the rate increases for larger values of I.

[2]The assignment is computed with knowledge of the original sources by maximizing the overall separation quality in a hill-climbing manner [9].

4. NMF-ISS ENCODER

4.1. Source NMF Model

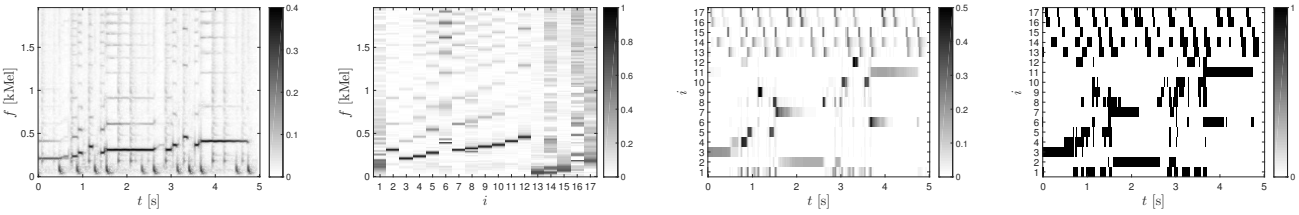
It is necessary to distribute the given number of components I over the available sources to be able to describe the sources $\mathbf{S}_m$ with independent NMF models. First, β-NMF with automatic relevance determination (β-ARD) [12] is used to estimate the optimal number of components for each source[3] which needs to be computed only once per mixture. The ratio of these numbers is then used to distribute I over the sources to obtain the number of components per source $\mathbf{I}_{\mathrm{src}}(m) \in \mathbb{N}_+^M$ with $\sum_m \mathbf{I}_{\mathrm{src}}(m) = I$.

β-NMF is used to factorize each source magnitude spectrogram $\mathbf{S}_m$ into $\mathbf{I}_{\mathrm{src}}(m)$ components independently of the other sources. The parameters are chosen to match the parameters of the decoder NMF, apart from the number of components. The basis and gain matrices describing each source spectrogram are stacked together in the component dimension in the concatenation-block (denoted with "C" in Fig. 1) to form $\mathbf{B}_{\mathrm{src}} \in \mathbb{R}_+^{F\times I}$ and $\mathbf{G}_{\mathrm{src}} \in \mathbb{R}_+^{T\times I}$ which have the same dimensions as the decoder NMF matrices (cf. Eq. (1)). As initialization for both β-ARD and source β-NMF, a fixed semantic initialization is used modelling the spectral behaviour of the 88 piano keys [9, 13]. A source NMF model is shown in Fig. 2b and 2c for an exemplary guitar-drum mixture depicted in Fig. 2a. Using NTF to factorize the sources jointly (as done in e.g. [3]) could result in components describing multiple sources which would deteriorate the NMF performance in the decoder.

4.2. Decoder NMF Model Initialization

The source NMF model is used to calculate an initialization for the decoder NMF ("Init Coding" block in Fig. 1) to enhance the separation quality of the SBSS algorithm. In the following, we introduce a very compact initialization scheme which can be coded very efficiently. We assume that it is sufficient to transmit a binary activity pattern instead of a quantized version of $\mathbf{G}_{\mathrm{src}}$, as the decoder NMF is able to extract a finer-structured $\mathbf{G}$ out of $\mathbf{X}$ given the components activity information. The initial gain matrix $\mathbf{G}_0$ is obtained by simple thresholding of $\mathbf{G}_{\mathrm{src}}$ in dB with threshold $\tau_{\mathbf{G}_0}$ as $\mathbf{G}_0(t,i) = 1$ if $\mathbf{G}'_{\mathrm{src}}(t,i) > \tau_{\mathbf{G}_0}$ and $\mathbf{G}_0(t,i) = 0$ otherwise with $\mathbf{G}'_{\mathrm{src}}(t,i) = 10\log_{10}(\mathbf{G}^2_{\mathrm{src}}(t,i)/\max_{t,i}\mathbf{G}^2_{\mathrm{src}}(t,i))$ denoting $\mathbf{G}_{\mathrm{src}}$ in dB. The structure of the resulting binary matrix $\mathbf{G}_0$, depicted in Fig. 2d, motivates the usage of componentwise run-length

[3]Compared to β-NMF, β-ARD estimates the relevance of each component additionally to $\mathbf{B}$ and $\mathbf{G}$. Non-relevant components are discarded. All additional β-ARD parameters are chosen as proposed as default in [12].



(a) Mixture spectrogram $\mathbf{X}$. (b) Source basis matrix $\mathbf{B}_{\mathrm{src}}$. (c) Source gain matrix $\mathbf{G}_{\mathrm{src}}^{\mathrm{T}}$. (d) Initial gain matrix $\mathbf{G}_{0}^{\mathrm{T}}$.

Fig. 2. Mixture spectrogram $\mathbf{X}$, source model matrices $\mathbf{B}_{\mathrm{src}}$, $\mathbf{G}_{\mathrm{src}}$ and initial temporal gain matrix $\mathbf{G}_0$ for an exemplary guitar-drum mixture. Components $i \in [1, 12]$ correspond to the guitar and $i \in [13, 17]$ to the drum recording.

coding [14] and subsequent adaptive arithmetic coding [15] of the run-lengths.

The initialization for the basis matrix $\mathbf{B}_0$ is constructed out of frames $d_{\mathbf{B}_0}(i)$ of $\mathbf{X}$ for which the ith component is most dominant. These dominant frames are detected within the temporal gain matrix $\mathbf{G}_{\mathrm{src}}$ of the source NMF model

$$d_{\mathbf{B}_0}(i) = \arg\max_t \mathbf{G}_{\mathrm{src}}(t,i) / \sum_{j \neq i} \mathbf{G}_{\mathrm{src}}(t,j) \tag{3}$$

and directly converted to binary numbers as the frame indices are integer numbers.

After transmission, the initial gain matrix $\mathbf{G}_0$ is obtained by arithmetic and subsequent run-length decoding whereas the dominant frame indices are converted back to integer numbers $d_{\mathbf{B}_0}$ at the decoder side ("Init Decoding" block in Fig. 1). The indices $d_{\mathbf{B}_0}$ are then used to construct the initial basis matrix out of the mixture spectrogram

$$\mathbf{B}_0(f,i) = \mathbf{X}(f, d_{\mathbf{B}_0}(i)) \,. \tag{4}$$

If the encoder decides to skip the transmission of $d_{\mathbf{B}_0}$, the initial basis matrix is extracted out of $\mathbf{G}_0$ and $\mathbf{X}$ with a simple matrix multiplication or a median operation

$$\mathbf{B}_0 = \mathbf{X}\mathbf{G}_0^{\mathrm{T}} \quad \text{or} \quad \mathbf{B}_0(f,i) = \mathrm{median}\left[\mathbf{X}(f, \boldsymbol{\Theta}(i))\right] \tag{5}$$

with $\boldsymbol{\Theta}(i) = \{t \,|\, \mathbf{G}_0(t,i) > 0\}$ denoting all time frames for which the ith column of $\mathbf{G}_0$ is active.

The encoder tests SBSS with different initializations (different values of $\tau_{\mathbf{G}_0}$ and initialization schemes for $\mathbf{B}_0$) as well as different NMF parameters β and α_{tc} without calculating a residual NMF model and takes the configuration which yields the best SBSS performance. This procedure is denoted in the following as parameter optimization (PO).

4.3. Residual NMF Model

The encoder calculates a residual NMF model to take possible errors of the SBSS algorithm in the decoder into account. In the following, only the calculation for the residual gain matrix is shown as the corresponding calculations for the residual basis matrix can be derived in the same manner. To obtain a residual temporal gain matrix, it is necessary to align both gain matrices of the NMF source and mixture model. Here, all components are normalized to unit energy. The residual temporal gain matrix $\mathbf{G}_{\mathrm{res}}$ is calculated as

$$\mathbf{G}_{\mathrm{res}}(t,i) = \mathbf{G}_{\mathrm{src}}(t,i) / \mathbf{E}_{\mathbf{G}_{\mathrm{src}}}(i) - \mathbf{G}(t,i) / \mathbf{E}_{\mathbf{G}}(i) \tag{6}$$

with $\mathbf{E}_{\mathbf{G}_{\mathrm{src}}}$ and $\mathbf{E}_{\mathbf{G}}$ denoting the corresponding energies. In the decoder, the residuals are obtained by inverse quantization (marked with a hat symbol) and the NMF model is refined by

$$\mathbf{G}(t,i) \leftarrow \left[\mathbf{G}(t,i) / \mathbf{E}_{\mathbf{G}}(i) + \hat{\mathbf{G}}_{\mathrm{res}}(t,i)\right] \hat{\mathbf{E}}_{\mathbf{G}_{\mathrm{src}}}(i) \tag{7}$$

where $\mathbf{G}$ is again normalized to unit energy before the refinement and afterwards normalized by the transmitted energy $\hat{\mathbf{E}}_{\mathbf{G}_{\mathrm{src}}}$. The refinement of $\mathbf{B}$ is done accordingly. The updated NMF model is then used for synthesis with Eq. (2).

In [3, 6, 5], the source NTF matrices are quantized in the logarithmic domain with scalar quantization. As shown in Section 5, linear quantization gives better results for quantizing the residual NMF matrices of NMF-ISS. Here, we use an additional A-law companding step [14] prior to scalar quantization. The A-law compression curve mimics logarithmic behaviour for A-law parameter $A = 87.5$ whereas $A = 0$ implies linear scalar quantization. The resulting symbols are coded with adaptive arithmetic coding (as used for the run-lengths of $\mathbf{G}_0$ in Section 4.2).

The transmitted side-information consists of quantized and encoded versions of $d_{\mathbf{B}_0}$ and $\mathbf{G}_0$ for initialization of the decoder NMF and the residual NMF model $\mathbf{B}_{\mathrm{res}}$, $\mathbf{G}_{\mathrm{res}}$. In case of suboptimal SBSS performance, the encoder is able to skip the NMF in the decoder. In this case, the residual NMF model is equal to the source NMF model which is then coded as described in Section 4.3. Skipping the decoder NMF is indicated with a flag: $\mathrm{skip} = 1$ (cf. Fig 1). For very low bit rates, the encoder is also able to not only skip the transmission of the NMF residuals but of the NMF initialization as well. In this case, SBSS works completely blind.

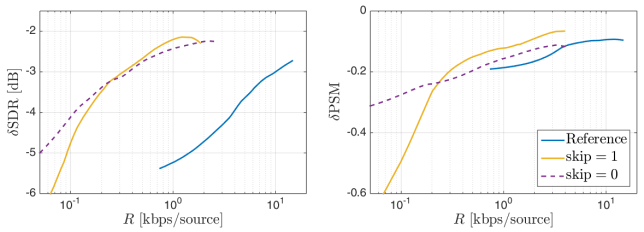
5. EXPERIMENTS

5.1. Setup

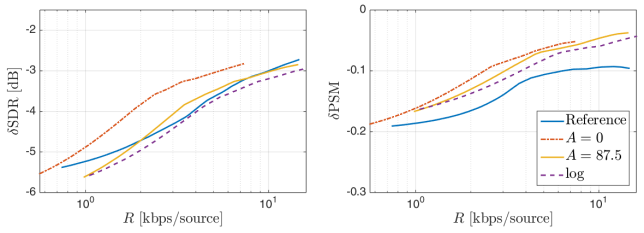
For evaluation of the proposed method, NMF-ISS is performed on five monaural mixtures sampled at $44\,100\,\mathrm{Hz}$ taken from the QUASI database[4]. The mixtures consist of 3 to 6 sources (e.g. vocals, guitar, drums, effects) and are about $20\,\mathrm{s}$ long. As quality measures, the signal-to-distortion ratio (SDR) calculated using the "BSS Eval" toolbox [16] and the perceptual similarity measure (PSM) of PEMO-Q [17] are obtained for each source. The mean measures are calculated over all sources per mixture in reference to the performance of an oracle estimator [18] which yields an upper bound for separation with Wiener-like filtering. The resulting measures are denoted as $\delta\mathrm{SDR}$ and $\delta\mathrm{PSM}$ respectively and given for bit rate R which is normalized per source.

Regarding the STFT, we chose a window size of $93\,\mathrm{ms}$ with $50\,\%$ overlap and Mel-filtering with $F = 400$. The encoder is tested with different numbers of NMF components I normalized per source $I/M \in \{2, 3, 4, 5, 10, 15, 20, 30\}$ with M denoting the number of sources. The encoder estimates optimal SBSS parameters at runtime without calculating the residuals by testing the SBSS algorithm with combinations of the following parameters: Regarding β-NMF,

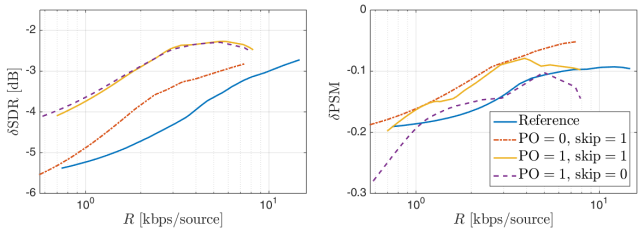
[4] http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/



(a) Influence of decoder NMF (either skipped or unskipped) in comparison with reference [3, 8].



(b) Comparison of quantization and coding schemes: Linear ($A = 0$) and logarithmic ($A = 87.5$) scalar quantization with adaptive arithmetic encoding and scalar quantization in logarithmic domain with Huffman encoding ("log") as used in reference method. Optimal parameter estimation disabled ($\mathrm{PO} = 0$) and decoder NMF skipped ($\mathrm{skip} = 1$) in all cases.



(c) Influence of optimal parameter estimation (PO). In case of disabled parameter optimization ($\mathrm{PO} = 0$), the decoder NMF is disabled ($\mathrm{skip} = 1$).

Fig. 3. δSDR and δPSM results over bit rate R for NMF-ISS. Influence of SBSS, different quantization schemes and parameter estimation as well as comparison with reference [3, 8].

the β-Divergence and temporal continuity parameters are chosen as $\beta \in \{0, 1, 2\}$ and $\alpha_{\mathrm{tc}} \in \{0, 25, 50, 100\}$. The initialization procedure is tested with different thresholds $\tau_{\mathbf{G}_0} \in \{-15, -30, -60\}$ dB for the calculation of $\mathbf{G}_0$ and either transmission of $d_{\mathbf{B}_0}$ or calculation out of $\mathbf{G}_0$ as denoted in Eq. (5) for obtaining $\mathbf{B}_0$. The optimal NMF and initialization parameter combination is then chosen as the one with the highest SBSS SDR. After the parameter optimization (PO), residual coding is conducted with different step sizes for scalar quantization of the NMF residuals. The corresponding number of quantization bins are calculated as $2^{Q_{\mathbf{B}}}$ and $2^{Q_{\mathbf{G}}}$ with $Q_{\mathbf{B}}, Q_{\mathbf{G}} \in [0, 8]$ bit.

All $(R, \delta\mathrm{SDR})$ and $(R, \delta\mathrm{PSM})$ points were optimized per mixture and I/M independently and then smoothed using the locally weighted scatter plot smoothing (LOESS) method to obtain rate/quality curves.

5.2. NMF-ISS performance

Fig. 3a shows results for our algorithm with either skipped or unskipped decoder NMF (cf. Fig. 1). In case of skipped decoder NMF ($\mathrm{skip} = 1$), the full source NMF model is coded and transmitted (as the residual model contains the full source NMF model). For lower and midrange bit rates, running the decoder NMF ($\mathrm{skip} = 0$) yields better results for both δSDR and δPSM than just by coding the source model as the additional amount of rate spent for the SBSS initialization is rewarded by a more sparse residual model. In the extreme case of $R \to 0$, SBSS is still able to estimate the sources with adequate quality. For higher bit rates, encoding only the source model is sufficient. Note that the encoder is able to skip the decoder NMF at run-time to make an optimal decision between these two modes. Additionally, results for the reference implementation [3, 8] are shown and discussed in the following section.

5.3. Comparison with reference

The reference implementation [3, 8] compresses the source spectrograms $\mathbf{S}_m$ directly with β-NTF and $\beta = 0$. For a fair comparison, NMF-ISS is simulated with a similar parameter and quantization configuration as the reference method in this section: The STFT window size is decreased to 46 ms and Mel-filtering is disabled. The residuals are quantized with $Q_{\mathbf{B}} = Q_{\mathbf{G}} = 8$ bit in logarithmic domain and Huffman encoded ("log") afterwards. However, this scheme is only applicable to nonnegative matrices. To cope with real-numbered residuals when using the decoder NMF ($\mathrm{skip} = 0$), the aforementioned scheme is replaced with our A-law quantization and adaptive arithmetic coding scheme (cf. Section 4.3).

Fig. 3b shows results for these different quantization and coding schemes with skipped decoder NMF ($\mathrm{skip} = 1$). Instead of the optimal parameter estimation (PO) which is done in the NMF-ISS encoder to determine optimal decoder parameters (cf. Sec. 4.2), a fixed NMF parameter setting is used with $(\beta, \alpha_{\mathrm{tc}}) = (0, 0)$ which is identical to the NTF setup for the reference method and denoted as $\mathrm{PO} = 0$. Comparing the reference method with NMF-ISS using the same coding scheme ("log") shows that the choice of NMF over NTF comes with a small loss of quality regarding δSDR and an increase for δPSM. Replacing this scheme with companding and adaptive arithmetic coding ($A = 87.5$) results in similar δSDR results for higher bit rates whereas δPSM is slightly improved. Disabling companding ($A = 0$) improves the quality significantly for both δSDR and δPSM and motivates the choice of linear scalar quantization for the NMF-ISS setup.

The influence of the optimal decoder parameter estimation of the encoder (cf. Section 4.2) is shown in Fig. 3c. Enabling estimation with ($\mathrm{PO} = 1$) yields a δSDR increase of about 1 dB for all rates whereas δPSM decreases slightly. Enabling the decoder NMF ($\mathrm{skip} = 0$) does not improve the quality any further. Comparing results for the two different STFT window sizes with either Mel-filtering enabled or disabled shown in Fig. 3a and 3c with $\mathrm{PO} = 1$ shows that Mel-filtering and larger STFT window size results in a bit rate reduction by a factor of 2 at similar δSDR values.

6. CONCLUSIONS

In this paper, we proposed a novel ISS algorithm, NMF-ISS, which makes use of an NMF-based semi-blind source separation (SBSS) algorithm in the decoder. The encoder chooses the optimal initial NMF parameters and calculates NMF residuals to enhance the separation quality further. This procedure introduces flexibility as the encoder is able to skip SBSS or residual coding. We showed experimentally that performing SBSS yields in higher quality at lower bit rates and show, that NMF-ISS outperforms a reference method.

Future work could include introduction of an additional residual coding step in the spectrogram domain to be able to yield higher performance than oracle estimators for source separation as well as further optimization of residual quantization and coding similar to [19].

7. REFERENCES

[1] Mathieu Parvaix, Laurent Girin, and Jean-Marc 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.

[2] Jürgen Herre, Heiko Purnhagen, Jeroen Koppens, Oliver Hellmuth, Jonas Engdegård, Johannes Hilper, Lars Villemoes, Leon Terentiv, Cornelia Falch, Andreas Hölzer, et al., "MPEG spatial audio object coding – the ISO/MPEG standard for efficient coding of interactive audio scenes," *Journal of the Audio Engineering Society*, vol. 60, no. 9, pp. 655–673, 2012.

[3] Antoine Liutkus, Jonathan Pinel, Roland Badeau, Laurent Girin, and Gaël Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[4] Antoine Liutkus, Stanislaw Gorlow, Nicolas Sturmel, Shuhua Zhang, Laurent Girin, Roland Badeau, Laurent Daudet, Sylvain Marchand, and Gaël Richard, "Informed audio source separation: A comparative study," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2397–2401.

[5] Joonas Nikunen, Tuomas Virtanen, and Miikka Vilermo, "Multichannel audio upmixing based on non-negative tensor factorization representation," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 33–36.

[6] Alexey Ozerov, Antoine Liutkus, Roland Badeau, and Gaël Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1699–1712, 2013.

[7] Serap Kirbiz, Alexey Ozerov, Antoine Liutkus, and Laurent Girin, "Perceptual coding-based informed source separation," in *2014 Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 959–963.

[8] Antoine Liutkus, Roland Badeau, and Gaël Richard, "Informed source separation using latent components," in *Latent Variable Analysis and Signal Separation*, pp. 498–505. Springer, 2010.

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

[10] Martin Spiertz and Volker Gnann, "Beta divergence for clustering in monaural blind source separation," in *128th AES Convention*, London, UK, May 2010.

[11] Tuomas 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.

[12] Vincent Y.F. Tan and Cédric Févotte, "Automatic relevance determination in nonnegative matrix factorization with beta-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, 2013.

[13] Martin Spiertz and Volker Gnann, "Note clustering based on 2D source-filter modeling for underdetermined blind source separation," in *Proceedings of the AES 42nd International Conference on Semantic Audio*, Ilmenau, Germany, July 2011.

[14] Jens-Rainer Ohm, *Multimedia Signal Coding and Transmission*, Signals and Communication Technology. Springer-Verlag Berlin Heidelberg, 2015.

[15] Mark Nelson and Jean-Loup Gailly, *The Data Compression Book, 2nd Edition*, M&T Books, New York, 1996.

[16] Emmanuel Vincent, Rémi Gribonval, and Cédric 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] Rainer Huber and Birger 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.

[18] Emmanuel Vincent, Rémi Gribonval, and Mark D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, Dec 2007.

[19] Alexey Ozerov, Antoine Liutkus, Roland Badeau, and Gaël Richard, "Informed source separation: source coding meets source separation," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 257–260.