ADAPTIVE CODING OF NON-NEGATIVE FACTORIZATION PARAMETERS WITH APPLICATION TO INFORMED SOURCE SEPARATION

Max Bläser *Christian Rohlfing* *Yingbo Gao* *Mathias Wien*

Institut für Nachrichtentechnik, RWTH Aachen University, Germany

ABSTRACT

Informed source separation (ISS) uses source separation for extracting audio objects out of their downmix given some pre-computed parameters. In recent years, non-negative tensor factorization (NTF) has proven to be a good choice for compressing audio objects at an encoding stage. At the decoding stage, these parameters are used to separate the downmix with Wiener-filtering. The quantized NTF parameters have to be encoded to a bit stream prior to transmission.

In this paper, we propose to use context-based adaptive binary arithmetic coding (CABAC) for this task. CABAC is widely used in the video coding community and exploits local signal statistics. We adapt CABAC to the task of NTF-based ISS and show that our contribution outperforms reference coding methods.

***Index Terms*—source separation, NMF, CABAC, arithmetic coding, audio object coding**

I. INTRODUCTION

Informed source separation (ISS) [1], [2] is a research topic bridging the areas of source separation and audio object coding. ISS consists of two stages: at the *encoder*, the audio objects, here recordings of single instruments, singing voice or effects, are perfectly known and used to compute a compact set of parameters which is transmitted to the *decoder*. Here, only the downmix of the audio objects is available. The transmitted parameters are used to assist a source separation step, estimating the audio objects given the downmix. This procedure enables numerous applications such as active listening or karaoke and is independent of the loudspeaker setup at the user side. Spatial Audio Object Coding (SAOC) [3] is a main building block of the recently proposed MPEG-H 3D Audio standard [4] which provides similar functionality. SAOC yields a parameter bit rate close to 3 kbps/object.

The baseline ISS method [5] models the spectrograms of the objects with non-negative tensor factorization (NTF) [6] and uses Wiener-filtering to separate the sources at the decoder yielding bit rates around 1 kbps/object. A recent variation of this method [7], [8] introduces a more complex decoding scheme as it uses a second NTF block in the decoder to refine coarsely quantized parameters and yields bit rates of around 0.5 kbps/object. Similar bit rates are obtained by the compressive sampling-based approach of [9]. Few source samples are selected randomly at the encoder. At the decoder, these samples are used for model estimation with NTF. [10], [11] combine NTF-based ISS and high-rate waveform coding, called the Coding-based ISS framework which is not bound by oracle estimators [12] anymore and leads to high separation quality with bit rates around 5 kbps/object. In contrast to the aforementioned methods, [13] is not using Wiener-filtering at all but encodes magnitude and phase spectrograms of the audio objects independently. However, NTF is used for magnitude modeling.

For encoding the quantized NTF parameters to a bit stream, different choices of encoding algorithms were made in prior work: Huffman coding (HC) [14] of the NTF parameters is used e.g. in [13]. In [5], this is achieved by GZIP which uses LZ77 in combination with HC [14]. The methods in e.g. [7], [10] use arithmetic coding which encodes sequences of symbols more efficiently than HC [14], [15]. In this paper, we propose to use the very efficient context-based adaptive binary arithmetic coding (CABAC) [16] which is widely used in the video coding community, e.g. in High Efficiency Video Coding (HEVC) [17]. CABAC is able to approach *conditional entropy* by exploiting dependencies within the signal's statistics. We show how to adapt this method to encode NTF parameters more efficiently than the aforementioned reference methods by proposing suitable contexts for NTF matrices.

This paper is structured as follows: In Section II, we present the baseline ISS system [5] on which we improve. In Section III, we summarize the CABAC scheme. In Section IV, we present contexts designed for coding the ISS parameters with CABAC. Finally, we evaluate the impact of using CABAC for ISS in Section V and summarize our contribution in Section VI.

Max Bläser and Christian Rohlfing are co-first authors.

II. NON-NEGATIVE-FACTORIZATION-BASED INFORMED SOURCE SEPARATION

The baseline ISS method [5] based on non-negative tensor factorization (NTF) consists of two stages: In the *encoder*, the sources are perfectly known and used for computing compact parameters with NTF. These parameters are quantized, encoded to a bit stream and then transmitted to the *decoder*. At the decoder side only the mix and the compressed parameters are available. The parameters are decompressed from the bit stream and the sources are finally estimated with Wiener-filtering.

II-A. Parameters estimation and coding at encoder

The complex time-frequency representations of the J sources and of the mixture are denoted as $\underline{\mathbf{S}}_j$ and $\underline{\mathbf{X}}$ respectively, with one particular time-frequency (TF) bin $\underline{s}_{f,t,j}$ and $\underline{x}_{f,t} = \sum_j \underline{s}_{f,t,j}$ respectively. The magnitude of the source at TF bin, $s_{f,t,j} = \left|\underline{s}_{f,t,j}\right|$, may be approximated by NTF with

$$s_{f,t,j} \approx \hat{s}_{f,t,j}\left(\Theta\right) = \sum_{k=1}^{K} w_{f,k} h_{t,k} q_{j,k}, \tag{1}$$

where $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{Q}$ are $F \times K$, $T \times K$ and $J \times K$ non-negative matrices and gathered as $\Theta = \{\mathbf{W}, \mathbf{H}, \mathbf{Q}\}$. Here, we use NTF with multiplicative update rules minimizing the β-divergence between the spectrogram $\mathbf{S}_j$ and the approximation given by Equation (1):

$$d_\beta\left(\mathbf{S}_j \mid \Theta\right) \triangleq \sum_{f,t} d_\beta\left(s_{f,t,j} \mid \hat{s}_{f,t,j}\left(\Theta\right)\right). \tag{2}$$

The β-divergence includes e.g. Itakura-Saito distance ($\beta = 0$), Kullback-Leibler divergence ($\beta = 1$) and Euclidean distance ($\beta = 2$). The corresponding multiplicative update rules for all parameters $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{Q}$ are given in detail e.g. in [5], [18]. Note that the overall NTF performance strongly depends on the choices of the initial parameters as well as the number of components K.

Applied on audio spectrograms, the NTF parameters can be interpreted as follows. As given in (1), NTF factorizes the input TF bin $s_{f,t,j}$ into a sum of K *components*. $\mathbf{W}$ consists of K

Conditions	$n \leq N_{\mathrm{LBP}}$			$n > N_{\mathrm{LBP}}$
	–	$b_n^{f-1,k} = 0$	$b_n^{f-1,k} = 1$	–
Context Selection	$\mathrm{ctx}_{n,\mathrm{na}}$	$\mathrm{ctx}_{n,\mathrm{up0}}$	$\mathrm{ctx}_{n,\mathrm{up1}}$	$\mathrm{ctx}_{\mathrm{rst}}$
Context Initialization	$p\left(b_n^{f,k} = 0\right)$	$p\left(b_n^{f,k} = 0 \mid b_n^{f-1,k} = 0\right)$	$p\left(b_n^{f,k} = 0 \mid b_n^{f-1,k} = 1\right)$	$p\left(b_n^{f,k} = 0\right)$

Table I: Proposed conditional context selection and initialization for bin $b_n^{f,k}$ at position n of bin-string $\mathbf{b}^{f,k}$.

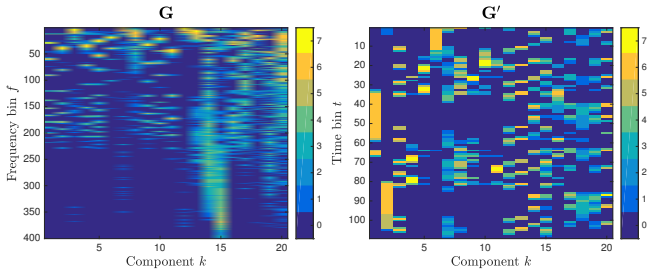


Fig. 1: Quantization indices $\mathbf{G}$ and $\mathbf{G}'$ corresponding to $\mathbf{W}$ and $\mathbf{H}$ for an exemplary guitar drum mixture with $K = 20$ components and $N_{\mathrm{Q}} = 8$ quantization intervals.

spectral basis functions (one for each component) and $\mathbf{H}$ holds the corresponding temporal activations. The activity of each component in each source is stored in $\mathbf{Q}$.

For transmission, quantization of the source parameters is conducted after parameter estimation by NTF in the logarithmic domain as proposed in [5], [10] using scalar quantization $q(\cdot)$ on each element independently:

$$[\mathbf{G}, \mathbf{c}] \triangleq q(\log \mathbf{W}), \tag{3}$$

where the $N_{\mathrm{Q}} \times 1$ vector $\mathbf{c}$ denotes the N_{Q} corresponding quantization centroids. The $F \times K$ matrix $\mathbf{G}$ consists of the corresponding integer-valued quantization indices with $0 \leq g_{f,t} \leq N_{\mathrm{Q}} - 1$. The quantized version of $\mathbf{H}$ is obtained in the same way. Exemplary matrices are shown in Figure 1. $\mathbf{Q}$ only has few elements compared to $\mathbf{W}$ and $\mathbf{H}$. Here, we focus only on the costly coding of $\mathbf{W}$ and $\mathbf{H}$ and quantize $\mathbf{Q}$ with high resolution in the linear domain.

In [5], the quantization indices of $\mathbf{W}$ and $\mathbf{H}$ are fed into the GZIP algorithm[1]. In this paper, we propose to use context-based adaptive binary arithmetic coding (CABAC) instead in Section III.

II-B. Source reconstruction at decoder

The NTF parameters Θ describing the sources are quantized and encoded to a bit stream. At the decoder, the quantization indices and centroids are extracted out of the bit stream by decoding. The inverse quantization operation

$$\bar{w}_{f,k} = c_i, \qquad i = g_{f,k} + 1, \quad \text{with } 1 \leq i \leq N_{\mathrm{Q}}$$

yields the reconstructed matrix $\bar{\mathbf{W}}$; $\bar{\mathbf{H}}$ is obtained accordingly. All reconstructed matrices are gathered under $\bar{\Theta}$, including reconstructed $\bar{\mathbf{Q}}$ which we encode as well as the quantization centroids as single precision floats with GZIP.

The mix $\underline{\mathbf{X}}$ is assumed to be available at the decoder. Given $\bar{\Theta}$, the estimated sources $\tilde{\underline{\mathbf{S}}}_j$ are obtained by Wiener-filtering [19] where $\hat{s}_{f,t,j}(\Theta)$ is given in (1):

$$\tilde{\underline{s}}_{f,t,j} \leftarrow \mathbb{E}\left[\underline{s}_{f,t,j} \mid \underline{x}_{f,t}, \bar{\Theta}\right] = \frac{\hat{s}_{f,t,j}\left(\bar{\Theta}\right)}{\sum_{j'} \hat{s}_{f,t,j'}\left(\bar{\Theta}\right)} \underline{x}_{f,t}. \tag{4}$$

[1]https://www.gnu.org/software/gzip/

III. CONTEXT-BASED ADAPTIVE BINARY ARITHMETIC CODING

Context-based adaptive binary arithmetic coding (CABAC) is used for entropy coding in state of the art video compression standards, such as H.264/AVC [16] and its successor HEVC [17]. CABAC combines adaptive arithmetic coding (AAC), allowing the assignment of non-integer numbers of bits to source symbols, with modeling of higher order dependencies within the source statistics. In contrast to AAC which only tracks the global (non-binary) symbol distributions, CABAC exploits the typically lower conditional entropy by adapting to local conditional symbol probabilities for even higher compression performance. The CABAC engine, also called the M-coder, is a normative part of HEVC and provides high coding throughput due to its multiplication free implementation.

In CABAC, the core steps of interval subdivision and probability update of AAC are only performed for a binary source. Thus, each non-binary source symbol g is binarized using a prefix-free code $C(\cdot)$, resulting in a *bin-string* $\mathbf{b}$

$$\mathbf{b} = C(g) = (b_1, \ldots, b_n, \ldots, b_{N_{\mathrm{C}}})^{\top}. \tag{5}$$

Each bin-string, an $N_{\mathrm{C}} \times 1$ vector, may consist of a variable number N_{C} of individual *bins* b_n which can each assume the value of '0' or '1'. It is clear that depending on the binarization method and the distribution of non-binary source symbols, bin-strings of vastly different length and distributions of '0's and '1's may result. Thus, the initial binarization should correctly map to the initial distribution of the source which is in turn depending on the quantizer setting. For sources of geometric symbol distribution, typical binarizations are Truncated Unary (TU) or Exponential Golomb (EG) codes. For more information, we refer to [20].

As mentioned before, CABAC adapts to local conditional statistics of the input. CABAC uses *contexts* to model states of information (e.g. already coded bins) in the *context modeler*. Each context maps uniquely to one particular state of information which has to be at hand both at encoder and decoder. The probability for the value of a current bin b_n to be en- or decoded is estimated by the conditional probability for the value of b_n given the selected context, $p(b_n \mid \mathrm{ctx})$ which is then used for *binary arithmetic coding* (BAC) of bin b_n. Here, the probability given by the selected context and the bin value is used to perform the necessary interval subdivision. After en- or decoding b_n, the probability model for ctx is updated with the value of b_n. Typically, multiple contexts are associated with a particular bin, each modeling a different probabilistic belief. In the actual coding step however, only a single context is chosen and updated. Contexts are commonly designed based on the bin position or the value of the preceding symbol or bin which represents the belief about the value of the current bin given the value of the neighboring bin. This conditional probability modeling is directly motivated by the underlying structure of the data [16].

In practice, CABAC uses an efficient way to model $p(b_n \mid \mathrm{ctx})$ using a finite state machine with 64 discrete probability states for each context [20]. CABAC offers another coding engine next to BAC, the faster bypass engine which is used to encode bins with nearly equiprobable distribution in HEVC [20] such as sign flags. Since such bins are not regarded in our approach, the bypass engine is deactivated and not further explained.

f	$g_{f,3}$	$C\left(g_{f,3}\right)$	Selected contexts for bin $b_n^{f,3}$			
			$n=1$	$n=2$	$n=3$	$n>3$
1	0	0	$\mathrm{ctx}_{1,\mathrm{na}}$	–	–	–
2	0	0	$\mathrm{ctx}_{1,\mathrm{up0}}$	–	–	–
3	6	1111110	$\mathrm{ctx}_{1,\mathrm{up0}}$	$\mathrm{ctx}_{2,\mathrm{na}}$	$\mathrm{ctx}_{3,\mathrm{na}}$	$\mathrm{ctx}_{\mathrm{rst}}$
4	2	110	$\mathrm{ctx}_{1,\mathrm{up1}}$	$\mathrm{ctx}_{2,\mathrm{up1}}$	$\mathrm{ctx}_{3,\mathrm{up1}}$	–
5	6	1111110	$\mathrm{ctx}_{1,\mathrm{up1}}$	$\mathrm{ctx}_{2,\mathrm{up1}}$	$\mathrm{ctx}_{3,\mathrm{up0}}$	$\mathrm{ctx}_{\mathrm{rst}}$

Table II: Context selection for each bin $b_n^{f,k}$ of bin strings $\mathbf{b}^{f,k} = C\left(g_{f,k}\right)$ given the third component ($k=3$) of exemplary $\mathbf{G}$ shown in Figure 1 with $1 \leq f \leq 5$, and $N_{\mathrm{LBP}} = 3$.

IV. CONTEXT MODELING FOR NTF PARAMETERS

The structure of our entropy coding scheme using CABAC is depicted in Figure 2b. Input to the coding stage are non-negative valued matrices. In the following, we show the encoding process for $\mathbf{W}$. The same scheme is used for $\mathbf{H}$ and the bit streams for both matrices are finally concatenated. The matrix elements are first quantized to integer numbers as described in Section II-A. As CABAC requires a binary input, each element of matrix $\mathbf{G}$ is subject to a *binarization*, resulting in variable-length code words which are fed into the *context modeler*. Here, local statistical dependencies of the source symbols are utilized to steer the subsequent *BAC*.

The typical structure of $\mathbf{G}$ for exemplary NTF parameters $\mathbf{W}$ (and $\mathbf{G}'$ for $\mathbf{H}$) is depicted in Figure 1. Note that both matrices are strongly structured: First, the quantization index 0, representing the lowest value, is by far the most frequent value, due to the sparse structure of spectrograms. Second, for each component k, long runs of zeroes and other values are also common. In the literature, this typical structure of NTF matrices is used for example in the design of NTF constraints, adapting NTF to sparseness or continuity properties [21], [22]. We exploit these properties for the proper selection of binarization methods and the design of several contexts. Due to the similar structure of $\mathbf{W}$ and $\mathbf{H}$, we propose to use the same context modeling for both matrices.

IV-A. Context selection

For binarization of $\mathbf{G}$, we keep the information about the position of the corresponding $g_{f,k}$ by adding the exponent (f,k) to each bin-string, composed of bins $b_n^{f,k}$. Thus, (5) becomes

$$\mathbf{b}^{f,k} = C\left(g_{f,k}\right) = \left(b_1^{f,k}, \ldots, b_n^{f,k}, \ldots, b_{N_{\mathrm{C}}}^{f,k}\right)^{\top}. \tag{6}$$

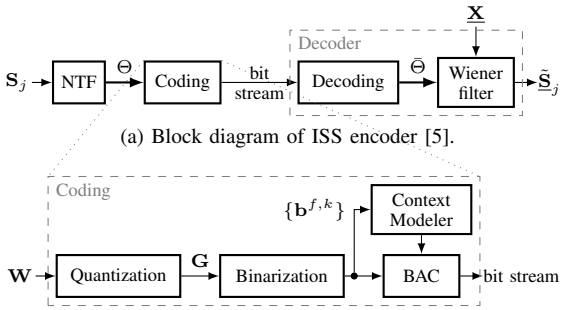
We use TU coding for binarization as the quantizer for $\mathbf{W}$ and $\mathbf{H}$ is operating at lower bit rates. For the exact parametrization, refer to Section V. TU coding is optimal for a geometric shaped distribution that we observed for such a coarse quantization and is achieved by coding the positive integer symbol $g_{f,k}$ with a '1'-sequence of length $g_{f,k}$ terminated with a single '0' [20]:

$$C\left(g_{f,k}\right) = \underbrace{11\ldots1}_{g_{f,k}\text{ times}}0. \tag{7}$$

For the maximal value of $g_{f,k} = N_{\mathrm{Q}} - 1$, the terminal '0' is omitted. Although we choose TU as binarization method, the following context design is also applicable for the prefix part of Exponential Golomb codes, generalizing TU coding.

Our context modeling is applied at the bin level of each bin-string given in (6). For each bin, a specific context is chosen, depending on the following conditions:

- The value of the corresponding bin $b_n^{f-1,k}$ of the preceding bin-string within component (or column) k.
- The position n within the current bin-string, if the other condition does not apply.



(a) Block diagram of ISS encoder [5].

(b) Proposed parameter *coding* block for $\mathbf{W}$ with CABAC.

Fig. 2: Block diagrams of ISS encoder and parameter coding block.

An overview of the entire conditional context selection is given in Table I. Here, $\mathrm{ctx}_{n,\mathrm{na}}$ represents the default context being selected if no condition is met (not available). This can occur if no corresponding bin at position n of a preceding bin-string is available. This context models the global distribution $p\left(b_n^{f,k}\right)$ for all f and k. If the corresponding bin $b_n^{f-1,k}$ of the preceding bin-string within column k is available and has the value '0' or '1', the contexts $\mathrm{ctx}_{n,\mathrm{up0}}$ or $\mathrm{ctx}_{n,\mathrm{up1}}$, respectively, are chosen which model the current conditional probabilities given the bin value of the preceding bin-string $p\left(b_n^{f,k} \mid b_n^{f-1,k}\right)$.

In order to limit the total number of contexts, we introduce the parameter N_{LBP} which is the last bin position using the described conditional context modeling. All bins $b_n^{f,k}$ with $n > N_{\mathrm{LBP}}$ are coded using the same context $\mathrm{ctx}_{\mathrm{rst}}$.

Table I also shows the initialization of the proposed contexts. Note that each context has to be initialized at the encoder and decoder side with the same initial values. Other context designs, for example contexts modeling the behavior across components (between $b_n^{f,k}$ and $b_n^{f,k-1}$) or across more than one preceding bin-string ($b_n^{f,k}$, $b_n^{f-1,k}$ and $b_n^{f-2,k}$), did not improve performance significantly and are therefore not evaluated in Section V.

Given our choice of TU coding as binarization (7), the conditional contexts $\mathrm{ctx}_{n,\mathrm{up0}}$ and $\mathrm{ctx}_{n,\mathrm{up1}}$ can be interpreted as follows: $\mathrm{ctx}_{n,\mathrm{up0}}$ models the probability of runs of identical values in a column of $\mathbf{G}$ which for $n=1$ is a sequence of quantized zero values. The context $\mathrm{ctx}_{n,\mathrm{up1}}$ models a sequence of values in a column of $\mathbf{G}$ which are larger or equal than value n.

An example is given in Table II, how quantization indices from a specific section occurring in the third column of $\mathbf{G}$ (shown in Figure 1) map to bin-strings using TU coding. We further show how contexts are chosen for each bin in the given example, depending on the aforementioned conditions.

IV-B. Complete system

The encoder depicted in Figure 2a resembles the structure of the baseline ISS method [5]. The source spectrograms $\mathbf{S}_j$ are factorized by NTF as described in Section II-A. The NTF parameters Θ are quantized and the resulting quantization indices are encoded. This procedure is shown in detail for $\mathbf{W}$ in Figure 2b: After quantization, $\mathbf{G}$, the integer-valued representation of $\mathbf{W}$, is binarized yielding bin-strings $\{\mathbf{b}^{f,k}\}$, one for each element at position (f,k). Given the already coded bin-strings, the context modeler chooses the appropriate context for bin $b_n^{f,k}$ and passes this information to the binary arithmetic coder. The resulting bit stream is transmitted to the decoder. Here, the estimated sources $\underline{\tilde{\mathbf{S}}}_j$ are obtained by Wiener-filtering mixture $\underline{\mathbf{X}}$ as shown in Section II-B.

V. EVALUATION

V-A. Data sets and setup

In this paper we are using two independent test sets for evaluating the proposed method. For evaluation of the context design

Method	GBAC	CABAC			
Conditional Contexts	–	–	$\mathrm{ctx}_{n,\mathrm{up0}}$	$\mathrm{ctx}_{n,\mathrm{up1}}$	$\mathrm{ctx}_{n,\mathrm{up0}}$, $\mathrm{ctx}_{n,\mathrm{up1}}$
Mean saving, %	-5.86	-15.30	-23.94	$\mathbf{-26.08}$	-25.64
Mean BD saving, %	-8.92	-18.28	-27.14	$\mathbf{-28.99}$	-28.53

Table III: Bit rate savings with respect to GZIP for GBAC ($N_{\mathrm{LBP}} = 0$) and CABAC ($N_{\mathrm{LBP}} = 10$) for the QUASI test set.

for CABAC, we used 10 mixtures consisting of 4 – 7 sources (e.g. vocals, guitar, drums and effects) of the QUASI database[2]. We then fixed the parameters of the proposed methods and compare it against reference coding methods on a second database: We used 100 mixtures consisting of 4 sources (bass, drums, vocals, other) of the DSD100 database[3] for this task. For both databases, each recording is sampled at 44100 Hz and is 30 s long. The separation quality is measured with the signal-to-distortion ratio (SDR, in dB) between original and estimated sources. After taking the mean over the sources for each mix, the resulting SDR value is set in reference to the SDR obtained by an oracle estimator [12] which estimates optimal Wiener filter masks. The resulting measure is denoted with δSDR. The side-information bit rate R of the bit stream (see Figure 2a) is measured in kbps per object.

We use STFT as time-frequency transform with a 93 ms window size and 50 % overlap. The spectral dimension of the spectrograms is filtered with a Mel-filterbank with $F = 400$ Mel-filters [18]. Different numbers of NTF components per source $K/J \in \{1, 2, \ldots, 10\}$ are evaluated with J number of sources and $\beta = 1$ (Kullback–Leibler divergence). The NTF parameters are initialized with an SVD-based method [23] and quantized with $N_{\mathrm{Q}} \in \{2, 4, 8, 16\}$. Note that K/J and N_{Q} have strong influence on the bit rate R. For CABAC, Truncated Unary (TU) as binarization is chosen given the comparatively small values for N_{Q}. We determined experimentally that EG coding only gives small bit rate savings for the highest values of N_{Q}. We fixed $N_{\mathrm{LBP}} = 10$, resulting in a total number of 31 contexts.

For each mixture, all combinations of parameters (K/J, N) result in multiple$(R, \delta\mathrm{SDR})$-points which are optimized to yield the Pareto front per mixture. The optimal points are smoothed using the locally weighted scatter plot smoothing method [24] to obtain rate/quality curves. We also calculate bit rate savings with respect to the reference GZIP by averaging the rate differences of all Pareto points. Additionally, bit rate savings averaged per mixture following the Bjøntegaard Delta (BD) measurement method [25] are given.

For comparison, we encode $\mathbf{W}$ and $\mathbf{H}$ with GZIP, Huffman Coding (HC), Arithmetic Coding (AC) and Run-length Encoding (RLE) [14]. To investigate the influence of conditional context modeling, we also test deactivating the context-modeler and using the BAC with only one global context $\mathrm{ctx}_{\mathrm{rst}}$ (GBAC). We choose to evaluate RLE additionally because sequences of equal values are present in the data quite often as exemplified in Figure 1. RLE is able to approach conditional entropy if the Markov property of the input data is fulfilled [14]. RLE was already introduced to NMF-based ISS in [7] for coding $\mathbf{H}$. Here, we use RLE for coding both $\mathbf{W}$ and $\mathbf{H}$ and choose EG codes for binarization of the resulting integer-valued run-lengths and -symbols.

V-B. Experiments and discussion

First, we evaluate CABAC and our context design in reference to GZIP as it was used in the baseline [5] on the QUASI database. To assess CABAC's performance, mean bit rate savings with respect to GZIP are calculated. Table III gives an overview of the mean bit rate savings of GBAC and CABAC using different context settings: For both methods GBAC and CABAC, $\mathrm{ctx}_{\mathrm{rst}}$ is activated. For CABAC, $\mathrm{ctx}_{n,\mathrm{na}}$ is activated at all times additionally.

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

[3]"MUS 2016" task, http://sisec.inria.fr.

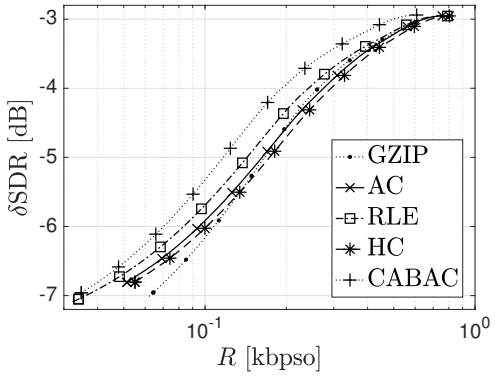


Fig. 3: Rate-quality curves for the proposed CABAC coding scheme with $N_{\mathrm{LBP}} = 10$ and $\mathrm{ctx}_{n,\mathrm{up1}}$ activated in comparison with reference methods for the DSD100 database.

- Activating $\mathrm{ctx}_{n,\mathrm{na}}$ without any other conditional context already yields a noticeable decrease in rate compared to GBAC from -5.86% to -15.30%.
- Regarding the conditional contexts, the best results are achieved by activating $\mathrm{ctx}_{n,\mathrm{up1}}$ in addition to $\mathrm{ctx}_{n,\mathrm{na}}$ and $\mathrm{ctx}_{\mathrm{rst}}$ (-26.08%), closely followed by $\mathrm{ctx}_{n,\mathrm{up0}}$ (-23.94%).
- Activating both $\mathrm{ctx}_{n,\mathrm{up0}}$ and $\mathrm{ctx}_{n,\mathrm{up1}}$ gives a slightly less reduction compared to activating $\mathrm{ctx}_{n,\mathrm{up1}}$ alone.

Regarding the ablation study, under which different conditional contexts are disabled, it can be seen that the addition of contexts, which do not provide significant conditional probabilities, does not necessarily manifest in higher coding efficiency.

Figure 3 shows rate-quality curves for CABAC with $\mathrm{ctx}_{n,\mathrm{up1}}$ activated for the DSD100 test set. First, we conclude that CABAC consistently outperforms all tested reference methods considerably. Compared to GZIP, average rate reductions of around -28% can be achieved. Run-length encoding (RLE) is the method with the second highest rate reduction of around -12%. AC and HC outperform GZIP at lower bit rates whereas AC outperforms HC.

A full MATLAB implementation of the proposed algorithm and a standalone CABAC interface for MATLAB can be found on the companion website for this paper[4].

VI. CONCLUSION

We proposed the application of context-based adaptive binary arithmetic coding (CABAC) in the field of informed source separation (ISS). We designed contexts fitting the typical structure of parameters obtained by non-negative tensor factorization (NTF). We evaluated different context settings for CABAC and showed experimentally on a second test set that CABAC outperforms widely used reference methods such as GZIP or conventional arithmetic coding. Run-length encoding proved to be an adequate low-complexity alternative to CABAC although being less efficient.

Future work could include investigating if prediction methods could decrease the parameter bit rate in a high rate setting. Methods for rate-distortion optimized quantization (RDOQ) could be tested as well as NTF with sparseness constraints as proposed in [13].

[4]http://www.ient.rwth-aachen.de/cms/icassp2018/

VII. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] C. Rohlfing, A. Liutkus, and J. M. Becker, "Quantization-aware parameter estimation for audio upmixing," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017.

[9] Ç. Bilen, A. Ozerov, and P. Pérez, "Compressive sampling-based informed source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, Oct 2015.

[10] 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.

[11] 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.

[12] 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.

[13] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," in *128th Audio Engineering Society Convention (AES 2010)*, London, UK, May 2010.

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

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

[16] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 620–636, July 2003.

[17] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[18] 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].

[19] 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.

[20] M. Wien, *High Efficiency Video Coding – Coding Tools and Specification*, Signals and Communication Technology. Springer, Berlin, Heidelberg, Sept. 2014.

[21] 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.

[22] J. Becker, *Nonnegative Matrix Factorization with Adaptive Elements for Monaural Audio Source Separation*, vol. 16 of *Aachen Series on Multimedia and Communications Engineering*, Shaker Verlag, Aachen, Oct. 2016.

[23] 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.

[24] 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.

[25] G. Bjøntegaard, "Calculation of average PSNR differences between RD curves," Tech. Rep. VCEG-M33, ITU-T SG16/Q6 VCEG, Austin, USA, Apr. 2001.