Compressive sampling-based informed source separation

Cagdas Bilen, Alexey Ozerov, Patrick Pérez

COMPRESSIVE SAMPLING-BASED INFORMED SOURCE SEPARATION

Çağdaş Bilen, Alexey Ozerov* and Patrick Pérez*

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{cagdas.bilen, alexey.ozerov, patrick.perez}@technicolor.com

ABSTRACT

The paradigm of using a very simple encoder and a sophisticated decoder for compression of signals became popular with the theory of distributed coding and it has been exercised for the compression of various types of signals such as images and video. The theory of compressive sampling later introduced a similar concept but with the focus on guarantees of signal recovery using sparse and low rank priors lying in an incoherent domain to the domain of sampling. In this paper, we bring together the concepts introduced in distributed coding and compressive sampling with the informed source separation, in which the goal is to efficiently compress the audio sources so that they can be decoded with the knowledge of the mixture of the sources. The proposed framework uses a very simple time domain sampling scheme to encode the sources, and a sophisticated decoding algorithm that makes use of the low rank non-negative tensor factorization model of the distribution of short-time Fourier transform coefficients to recover the sources, which is a direct application of the principles of both compressive sampling and distributed coding.

Index Terms— Informed source separation, low complexity encoder, compressive sampling, nonnegative tensor factorization, generalized expectation-maximization

1. INTRODUCTION

Audio source separation is a challenging task in audio signal processing [1], in which the quality of the reconstructed sources depends strongly on the particular task and the amount of prior information that can be exploited. Informed source separation (ISS) [2–5], which is also strongly related to spatial audio object coding (SAOC) [6], is a new trend in source separation, where some *side-information* about the sources and/or the mixing system is extracted at a stage where the clean sources are still available, e.g., during the mixing of a music recording by a sound engineer. A natural constraint is that this side-information should be small enough as compared to encoding the sources independently. More precisely, an ISS method is based on a so-called *encoding* stage, where the side-information is extracted, given both the sources and their mixture, and a so-called *decoding* stage, where the sources are not available any more and estimated from the mixture, given the side-information. As such, the ISS being at the crossroads of source separation and compression [7], it usually leads to much better quality of reconstructed sources than the conventional audio source separation at the expense of some bitrate required for side-information transmission. Indeed, the quality of reconstructed sources can be fully controlled during the encoding stage [5, 7], and perceptual psycho-acoustic aspects can be taken into account [6, 8].

One of the limits of all existing ISS and SAOC schemes is that the encoding process is not very fast. All these approaches [2–6] require at least computing some time-frequency transform, estimating some models or parameters and optionally encoding some residual signals [5, 6]. Moreover, the decoding is usually faster than the encoding, since it relies on some models or parameters that are already estimated or pre-computed at the encoder.

The goal of this work is to propose an ISS approach where the computational load is shifted from the encoder to the decoder, i.e., the encoder should be extremely fast, possibly at the expense of a slower decoder. Possible advantages of such a low complexity encoding are as follows. First, this would allow performing encoding using very low power devices. Second, even if the devices are not low power, for the archiving purposes (e.g., archiving music or movie audio multitracks) the encoding is performed for every archived piece, while the decoding may only be needed occasionally, when there is a necessity. Thus, having a very low complexity encoder would lead to overall energy, time and cost savings.

The approach we propose in this work is inspired by distributed source coding [9] and in particular distributed video coding [10] paradigms, where the goal is also to shift the complexity from the encoder to the decoder. Our approach relies on the compressive sensing/sampling (CS) principles [11–13], since we are projecting the sources on a linear subspace spanned by a randomly selected subset of vectors of a basis that is incoherent [13] with a basis where the audio sources are sparse. Even though CS emerged as a field relying on sparse representations for signal reconstruction, it is later discovered that it is not only possible with sparse models but also with group sparse and low rank models, hence our proposed approach is directly related to CS. We baptize our approach *compressive sampling-based ISS (CS-ISS)*.

More specifically, we propose to encode the sources by a simple random selection of a subset of temporal samples of the sources[1] followed by a uniform quantization and an entropy encoder. This is the only side-information transmitted to the decoder. To recover the sources at the decoder from the quantized source samples and the mixture, we propose using a model-based approach that is inline with model-based compressive sensing [14]. Notably, we use the Itakura-Saito (IS) nonnegative tensor factorization (NTF) model

* The first and second authors have contributed equally for this work.

This work was partially supported by ANR JCJC program MAD (ANR-14-CE27-0002).

[1] Note that the advantage of sampling in time domain is double. First, it is faster than sampling in any transformed domain. Second, temporal basis is incoherent enough with the short time Fourier transform (STFT) frame where audio signals are sparse and it is even more incoherent with the low rank NTF representation of STFT coefficients. It is shown in compressive sensing theory that the incoherency of the measurement and prior information domains is essential for the recovery of the sources [13].

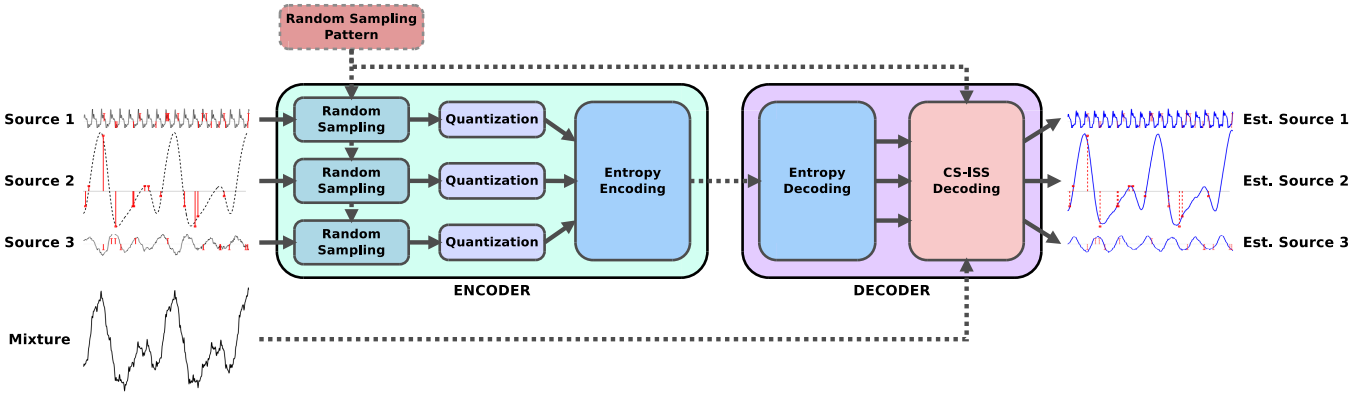

Figure 1: The overall structure of the encoder and the decoder in the proposed CS-ISS scheme. An example of three original (dashed black) and reconstructed sources (in blue) is shown, along with the 6-bit quantized random samples (in red) extracted by the encoder and used by the decoder to separate source mixture (in black).

of source spectrograms as in [4, 5]. We show that, thanks to its Gaussian probabilistic formulation [15], this model may be estimated in the maximum-likelihood (ML) sense from the mixture and the transmitted quantized portion of source samples. To estimate the model we develop a new generalized expectation-maximization (GEM) algorithm [16] based on multiplicative update (MU) rules [15]. Given the estimated model and all other observations, the sources can be estimated by Wiener filtering [17].

2. OVERVIEW OF THE CS-ISS FRAMEWORK

The overall structure of the proposed CS-ISS encoder/decoder is depicted in Figure 1. The encoder randomly subsamples the sources with a desired rate, using a predefined randomization pattern, and quantize these samples. The quantized samples are then ordered in a single stream to be compressed with an entropy encoder to form the final encoded bitstream. The random sampling pattern (or a seed that generates the random pattern) is known by both the encoder and the decoder and therefore not transmitted. The audio mixture is also assumed to be known by the decoder. The decoder performs entropy decoding to retrieve the quantized samples of the sources followed by CS-ISS decoding which will be discussed in detail in Section 3. The use of random samples and quantization for the purpose of compression and signal reconstruction is not new and is already used in compressive sampling applications, however using it for audio sources and informed source separation is proposed for the first time in this paper.

The proposed CS-ISS framework has several advantages over traditional ISS which can be summarized as follows:

- The simple encoder in Figure 1 can be used for low complexity encoding such as needed in low power devices. Low complexity encoding scheme is also advantageous for applications where encoding is used frequently but only few encoded streams need to be decoded. An example of such an application is music production in a studio where the sources of each produced music are kept for future use but seldom needed. Hence significant savings in terms of processing power and processing time is possible with CS-ISS.
- Performing sampling in time domain (and not in a transformed domain) provides not only a simple sampling scheme but also the possiblity to perform the encoding in an online fashion when needed, which is not always as straightforward for other methods [4, 5]. Furthermore, the encoding of each source being independent of the encoding of other sources enables the possibility of encoding sources in a *distributed* manner without compromising the decoding efficiency.
- The encoding step is performed without any assumptions on the decoding step, therefore it is possible to use other decoders than the one proposed in this paper. This provides a significant advantage over classical ISS [2–5] in the sense that when a better performing decoder is designed the encoded sources can directly benefit from the improved decoding without the need for re-encoding. This is made possible by the random sampling used in the encoder. It is shown by the compressive sensing theory that the random sampling scheme provides incoherency with a large number of domains so that it becomes possible to design efficient decoders relying on different prior information on the data.

3. CS-ISS DECODER

Let us indicate the support of the random samples of source $j \in \{1, \ldots, J\}$ with Ω''_j such that it is sampled at time indices $t \in \Omega''_j \subseteq \{1, \ldots, T\}$. After the entropy decoding stage, the CS-ISS decoder has the subset of quantized samples of the sources,[2] $y''_{jt}(\Omega''_j)$, where the quantized samples are defined as $y''_{jt} = s''_{jt} + b''_{jt}$ in which s''_{jt} indicates the true source signal and b''_{jt} is the quantization noise. In this work, the noise is modeled as zero mean i.i.d. Gaussian with the variance[3] σ^2. The mixture is assumed to be the sum of the original sources such that

$$x''_t = \sum_{j=1}^{J} s''_{jt}, \quad t \in \{1, \ldots, T\}, j \in \{1, \ldots, J\} \qquad (1)$$

[2]Throughout this paper the time-domain signals will be represented by letters with two primes, e.g., x'', framed and windowed time-domain signals will be denoted by letters with one prime, e.g., x', and complex-valued STFT coefficients will be denoted by letters with no prime, e.g., x.

[3]Even though it is known that the actual quantization noise distribution is not exactly Gaussian, this approximation is known to work well in practice while greatly simplfying the derivations.

Algorithm 1 GEM algorithm for CS-ISS Decoding using the NTF model

1: **procedure** CS-ISS DECODING($\mathbf{x}, \{\bar{\mathbf{y}}'_j\}_1^J, \{\Omega'_j\}_1^J, K$)
2: Initialize non-negative $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ randomly
3: **repeat**
4: Estimate $\hat{\mathbf{s}}$ (sources) and $\widehat{\mathbf{P}}$ (posterior power spectra), given $\mathbf{Q}, \mathbf{W}, \mathbf{H}, \mathbf{x}, \{\bar{\mathbf{y}}'_j\}_1^J, \{\Omega'_j\}_1^J$ ▷ E-step, see section 3.1
5: Update $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ given $\widehat{\mathbf{P}}$ ▷ M-step, see section 3.2
6: **until** convergence criteria met
7: **end procedure**

and known at the decoder. In order to compute the STFT coefficients, the mixture and the sources are first converted to windowed-time domain with a window length of M and a total of N windows. Resulting coefficients, denoted by y'_{jmn}, s'_{jmn}, b'_{jmn} and x'_{mn}, represent the quantized sources, the original sources, the quantization noise and the mixture in windowed-time domain respectively for $j = 1, \dots, J$, $n = 1, \dots, N$ and $m = 1, \dots, M$ (only for m in appropriate subset Ω'_{jn} in case of quantized source and quantization noise samples). Due to multiplying with a windowing function with the value, ω_m, at index m within the window, the variance of the noise in windowed time domain is, $\sigma^2_{b,m} = \omega_m^2 \sigma^2$. The STFT coefficients of the sources, s_{jfn}, of the noise, b_{jfn}, and of the mixture, x_{fn}, are computed by applying the unitary Fourier transform, $\mathbf{U} \in \mathbb{C}^{F \times M} (F = M)$, to each window of the windowed-time domain counterparts. For example,[4] $[x_{1n}, \cdots, x_{Fn}]^T = \mathbf{U}[x'_{1n}, \cdots, x'_{Mn}]^T$.

The sources are modelled in the STFT domain with a normal distribution ($s_{jfn} \sim \mathcal{N}_c(0, v_{jfn})$) where the variance tensor $\mathbf{V} = [v_{jfn}]_{j,f,n}$ has the following low-rank NTF structure (with a small K) [18]:

$$v_{jfn} = \sum_{k=1}^{K} q_{jk} w_{fk} h_{nk}. \tag{2}$$

This model is parametrized by $\boldsymbol{\theta} = \{\mathbf{Q}, \mathbf{W}, \mathbf{H}\}$, with $\mathbf{Q} = [q_{jk}]_{j,k} \in \mathbb{R}_+^{J \times K}$, $\mathbf{W} = [w_{fk}]_{f,k} \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H} = [h_{nk}]_{n,k} \in \mathbb{R}_+^{N \times K}$.

We propose to recover the source signals with a GEM algorithm that is briefly described in Algorithm 1. The algorithm estimates the sources and source statistics from the observations using a given model $\boldsymbol{\theta}$ via Wiener filtering at the expectation step, and then updates the model using the posterior source statistics at the maximization step. The details on each step of the algorithm are given the Sections 3.1 and 3.2 respectively.

3.1. Estimating the sources

Since all the underlying distributions are Gaussian and all the relations between the sources and the observations are linear, the sources may be estimated in the minimum mean square error (MMSE) sense via the Wiener filter [17] given the covariance tensor $\mathbf{V}$ defined in (2) by the model parameters $\mathbf{Q}, \mathbf{W}, \mathbf{H}$.

Let us define the observed data vector for the n-th frame, $\bar{\mathbf{o}}'_n$, as $\bar{\mathbf{o}}'_n \triangleq \left[\bar{\mathbf{y}}'^T_{1n}, \dots, \bar{\mathbf{y}}'^T_{Jn}, \mathbf{x}_n^T\right]^T$, where $\bar{\mathbf{y}}'_{jn} \triangleq \left[y'_{jmn}, m \in \Omega'_{jn}\right]^T$. We can write the posterior distribution of each source frame $\mathbf{s}_{jn}$ given the corresponding observed data $\bar{\mathbf{o}}'_n$ and the NTF model $\boldsymbol{\theta}$ as $\mathbf{s}_{jn}|\bar{\mathbf{o}}'_n; \boldsymbol{\theta} \sim \mathcal{N}_c(\hat{\mathbf{s}}_{jn}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}})$ with $\hat{\mathbf{s}}_{jn}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}$ being, respectively, posterior mean and posterior covariance matrix, each of which can be computed by Wiener filtering as [17]

$$\hat{\mathbf{s}}_{jn} = \boldsymbol{\Sigma}^H_{\bar{\mathbf{o}}'_n \mathbf{s}_{jn}} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{o}}'_n \bar{\mathbf{o}}'_n} \bar{\mathbf{o}}'_n, \tag{3}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \boldsymbol{\Sigma}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} - \boldsymbol{\Sigma}^H_{\bar{\mathbf{o}}'_n \mathbf{s}_{jn}} \boldsymbol{\Sigma}^{-1}_{\bar{\mathbf{o}}'_n \bar{\mathbf{o}}'_n} \boldsymbol{\Sigma}_{\bar{\mathbf{o}}'_n \mathbf{s}_{jn}}, \tag{4}$$

given the definitions

$$\boldsymbol{\Sigma}_{\bar{\mathbf{o}}'_n \bar{\mathbf{o}}'_n} = \begin{bmatrix} \boldsymbol{\Sigma}_{\bar{\mathbf{y}}'_{1n}\bar{\mathbf{y}}'_{1n}} & \cdots & \underline{\mathbf{0}} & \boldsymbol{\Sigma}^H_{\mathbf{x}_n \bar{\mathbf{y}}'_{1n}} \\ \vdots & \ddots & \vdots & \vdots \\ \underline{\mathbf{0}} & \cdots & \boldsymbol{\Sigma}_{\bar{\mathbf{y}}'_{Jn}\bar{\mathbf{y}}'_{Jn}} & \boldsymbol{\Sigma}^H_{\mathbf{x}_n \bar{\mathbf{y}}'_{Jn}} \\ \boldsymbol{\Sigma}_{\mathbf{x}_n \bar{\mathbf{y}}'_{1n}} & \cdots & \boldsymbol{\Sigma}_{\mathbf{x}_n \bar{\mathbf{y}}'_{Jn}} & \boldsymbol{\Sigma}_{\mathbf{x}_n \mathbf{x}_n} \end{bmatrix}, \tag{5}$$

$$\boldsymbol{\Sigma}_{\bar{\mathbf{o}}'_n \mathbf{s}_{jn}} = \left[\underline{\mathbf{0}}^T_{S_{1,jn} \times F}, \boldsymbol{\Sigma}^T_{\bar{\mathbf{y}}'_{jn}\mathbf{s}_{jn}}, \underline{\mathbf{0}}^T_{S_{2,jn} \times F}, \boldsymbol{\Sigma}^T_{\mathbf{x}_n \mathbf{s}_{jn}}\right]^T, \tag{6}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{\bar{\mathbf{y}}'_{jn}\bar{\mathbf{y}}'_{jn}} &= \mathbf{U}(\Omega'_{jn})^H \mathrm{diag}\left(\left[v_{jfn}\right]_f\right) \mathbf{U}(\Omega'_{jn}) \\ &\quad + \mathrm{diag}\left(\left[\sigma^2_{b,m} \left| m \in \Omega'_{jn}\right.\right]_m\right), \end{aligned} \tag{7}$$

$$\boldsymbol{\Sigma}_{\mathbf{s}_{jn}\mathbf{s}_{jn}} = \boldsymbol{\Sigma}_{\mathbf{x}_n \mathbf{s}_{jn}} = \mathrm{diag}\left(\left[v_{jfn}\right]_f\right), \tag{8}$$

$$\boldsymbol{\Sigma}_{\bar{\mathbf{y}}'_{jn}\mathbf{s}_{jn}} = \boldsymbol{\Sigma}^H_{\mathbf{x}_n \bar{\mathbf{y}}'_{jn}} = \mathbf{U}(\Omega'_{jn})^H \mathrm{diag}\left(\left[v_{jfn}\right]_f\right), \tag{9}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}_n \mathbf{x}_n} = \mathrm{diag}\left(\left[\sum\nolimits_j v_{jfn}\right]_f\right), \tag{10}$$

where $\mathbf{U}(\Omega'_{jn})$ is the $F \times |\Omega'_{jn}|$ matrix of columns from $\mathbf{U}$ with index in Ω'_{jn} and $S_{1,jn} \triangleq \sum_{\hat{j}=1}^{j-1} |\Omega'_{\hat{j}n}|$, $S_{2,jn} \triangleq \sum_{\hat{j}=j+1}^{J} |\Omega'_{\hat{j}n}|$.

Therefore the posterior power spectra, $\widehat{\mathbf{P}} = [\hat{p}_{jfn}]_{j,f,n}$, which will be used to update the NTF model as described in the following section, can be computed as

$$\hat{p}_{jfn} = \mathbb{E}\left[\left|s_{jfn}\right|^2 \middle| \bar{\mathbf{o}}'_n; \boldsymbol{\theta}\right] = |\hat{s}_{jfn}|^2 + \widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_{jn}\mathbf{s}_{jn}}(f, f). \tag{11}$$

3.2. Updating the model

NTF model parameters can be re-estimated using the multiplicative update (MU) rules minimizing the IS divergence [15] between the the 3-valence tensor of estimated source power spectra $\widehat{\mathbf{P}}$ and the 3-valence tensor of the NTF model approximation $\mathbf{V}$ defined as $D_{IS}(\widehat{\mathbf{P}} \| \mathbf{V}) = \sum_{j,f,n} d_{IS}(\hat{p}_{jfn} \| v_{jfn})$, where $d_{IS}(x\|y) = x/y - \log(x/y) - 1$ is the IS divergence; and $\hat{p}_{jfn}$ and v_{jfn} are specified respectively by (11) and (2). As a result, $\mathbf{Q}, \mathbf{W}, \mathbf{H}$ can be updated with the MU rules presented in [18]. These MU rules can be repeated several times to improve the model estimate.

4. RESULTS

In order to assess the performance of our approach, 3 (11 seconds length) sources of a music recording at 16 kHz are encoded and then decoded using the proposed CS-ISS with different levels of quantization (16 bits, 11 bits, 6 bits and 1 bit) and different raw sampling bitrates[5] per source (0.64, 1.28, 2.56, 5.12 and 10.24 kbps/source). Since uniform quantization is used, the noise variance in time domain is $\sigma^2 = \Delta^2/12$ where Δ is the quantization step size. It is assumed that the random sampling pattern is *pre-defined* and known during both encoding and decoding. The quantized samples are

[4] $\mathbf{x}^T$ and $\mathbf{x}^H$ represent the non-conjugate transpose and the conjugate transpose of the vector (or matrix) $\mathbf{x}$ respectively.

[5] The raw sampling bitrate is defined as the bitrate before the entropy encoding step.

Bits per Sample	Raw rate (kbps / source)				
	0.64	**1.28**	**2.56**	**5.12**	**10.24**
	Compressed Rate / SDR (% of Samples Kept)				
16 bits	0.50 / -1.64 dB (0.25%)	1.00 / 4.28 dB (0.50%)	2.00 / 9.54 dB (1.00%)	4.01 / 16.17 dB (2.00%)	8.00 / 21.87 dB (4.00%)
11 bits	0.43 / 1.30 dB (0.36%)	0.87 / 6.54 dB (0.73%)	1.75 / 13.30 dB (1.45%)	**3.50 / 19.47 dB** (2.91%)	**7.00 / 24.66 dB** (5.82%)
6 bits	**0.27 / 4.17 dB** (0.67%)	**0.54 / 7.62 dB** (1.33%)	**1.08 / 12.09 dB** (2.67%)	2.18 / 14.55 dB (5.33%)	4.37 / 16.55 dB (10.67%)
1 bit	0.64 / -5.06 dB (4.00%)	1.28 / -2.57 dB (8.00%)	2.56 / 1.08 dB (16.00%)	5.12 / 1.59 dB (32.00%)	10.24 / 1.56 dB (64.00%)

Table 1: The final bitrates (in kbps per source) after the entropy coding stage of CS-ISS with corresponding SDR (in dBs) for different (uniform) quantization levels and different raw bitrates before entropy coding. The percentage of the samples kept is also provided for each case in parentheses. Results corresponding to the best rate-distortion compromise are in bold.

truncated and compressed using an arithmetic encoder with a zero mean Gaussian distribution assumption. At the decoder side, following the arithmetic decoder, the sources are decoded from the quantized samples using 50 iterations of the GEM algorithm with STFT computed using a half-overlapping sine window of 1024 samples (64 ms) and the number of components fixed at $K = 18$, i.e. in average 6 components per source. The quality of the reconstructed samples is measured in signal to distortion ratio (SDR) as described in [19]. The resulting encoded bitrates and SDR of decoded signals are presented in Table 1 along with the percentage of the encoded samples in parentheses. Note that the compressed rates in Table 1 differ from the corresponding raw bitrates due to the variable performance of the entropy coding stage, which is expected.

The performance of CS-ISS is compared to a classical ISS approach with a more complicated encoder and a simpler decoder presented in [4]. The ISS algorithm is used with NTF model quantization and encoding as in [5], i.e., NTF coefficients are uniformly quantized in logarithmic domain, quantization step sizes of different NTF matrices are computed using equations (31)-(33) from [5] and the indices are encoded using an arithmetic coder based on a two-states Gaussian mixture model (GMM) (see Fig. 5 of [5]). The approach is evaluated for different quantization step sizes and different numbers of NTF components, i.e., $\Delta = 2^{-2}, 2^{-1.5}, 2^{-1}, \ldots, 2^{4}$ and $K = 4, 6, \ldots, 30$. The results are generated with 250 iterations of model update. The performance of both CS-ISS and classical ISS are shown in Figure 2 in which CS-ISS clearly outperforms the ISS approach, even though the ISS approach can use optimized number of components as opposed to our decoder which uses a fixed number of components (the encoder is very simple and does not compute or transmit this value). The performance difference is due to the high efficiency achieved by the CS-ISS decoder thanks to the incoherency of random sampled time domain and of low rank NTF domain. Also, the ISS approach [4] is unable to perform beyond an SDR of 10 dBs due to the lack of additional information about STFT phase as explained in [5]. Even though it was not possible to compare to the ISS algorithm presented in [5] in this paper due to time constraints, the results indicate that the rate distortion performance exhibits a similar behaviour. It should be reminded that the proposed approach distinguishes itself by it low complexity encoder and hence can still be advantageous against other ISS approaches with better rate distortion performance.

The performance of CS-ISS in Table 1 and Figure 2 indicates that different levels of quantization may be preferable in different rates. Even though neither 16 bits nor 1 bit quantization seem well performing, the performance indicates that 16 bits quantization may be superior to other schemes when a much higher bitrate is available. Coarser quantization such as 1 bit, on the other hand, had very

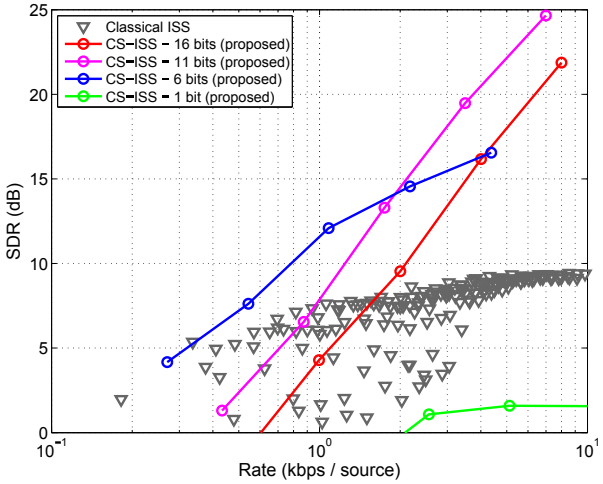

Figure 2: The rate-distortion performance of CS-ISS using different quantization levels of the encoded samples. The performance of ISS algorithm from [4] is also shown for comparison.

poor performance in the experiments. The choice of quantization can be performed in the encoder with a simple look up table as a reference. One must also note that even though the encoder in CS-ISS is very simple, the proposed decoder is significantly high complexity, typically higher than the encoders of traditional ISS methods. However, this can also be overcome by exploiting the independence of Wiener filtering among the frames in the proposed decoder with parallel processing, e.g., using graphical processing units (GPUs).

5. CONCLUSION

In this paper we proposed a novel low complexity informed source separation encoder that is based on compressed sampling principles. The encoded bitstream is decoded by an algorithm that exploits the low rank NTF structure in the distribution of the STFT coefficients and that is shown to compete with traditional ISS methods in terms of rate distortion performance. The proposed compressive sampling based informed source separation approach is first of its kind in the literature and has several advantages over traditional ISS approaches: it has a low complexity encoder, which is of practical interest in certain set-ups and it can benefit from better performance decoders in the future without the need to re-encode the sources. A more comprehensive performance assessment and comparison to other ISS algorithms is considered as future work.

6. REFERENCES

[1] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. Gowreesunker, D. Lutter, and N. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

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

[3] M. Parvaix and L. Girin, "Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1721 – 1733, 2011.

[4] 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.

[5] A. Ozerov, A. Liutkus, R. Badeau, and G. 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, Aug. 2013.

[6] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," in *124th Audio Engineering Society Convention (AES 2008)*, Amsterdam, Netherlands, May 2008.

[7] 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'11)*, New Paltz, New York, USA, Oct. 2011, pp. 257–260.

[8] S. Kirbiz, A. Ozerov, A. Liutkus, and L. Girin, "Perceptual coding-based informed source separation," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 959–963.

[9] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80–94, September 2004.

[10] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71 – 83, January 2005.

[11] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[12] R. Baraniuk, "Compressive sensing," *IEEE Signal Processing Mag.*, vol. 24, no. 4, pp. 118–120, July 2007.

[13] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, 2008.

[14] R. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Info. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[15] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[16] A. Dempster, N. Laird, and D. Rubin., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[18] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*, Prague, May 2011, pp. 257–260.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.