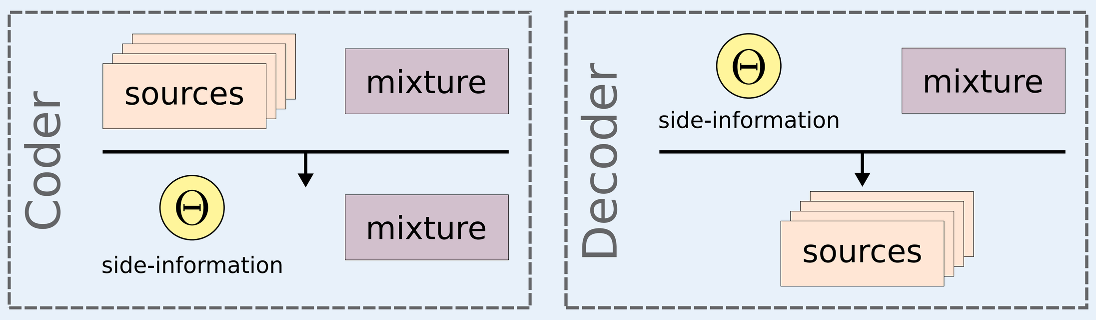
VERY LOW BITRATE SPATIAL AUDIO CODING WITH DIMENSIONALITY REDUCTION

Christian Rohlfing[1], Jeremy E.Cohen[2], Antoine Liutkus[3]

[1]Institut für Nachrichtentechnik, RWTH Aachen University
[2]Department of Image and Signal-processing, Gipsa-lab, CNRS, Grenoble, France
[3]Inria, speech processing team, Villers-lès-Nancy, France

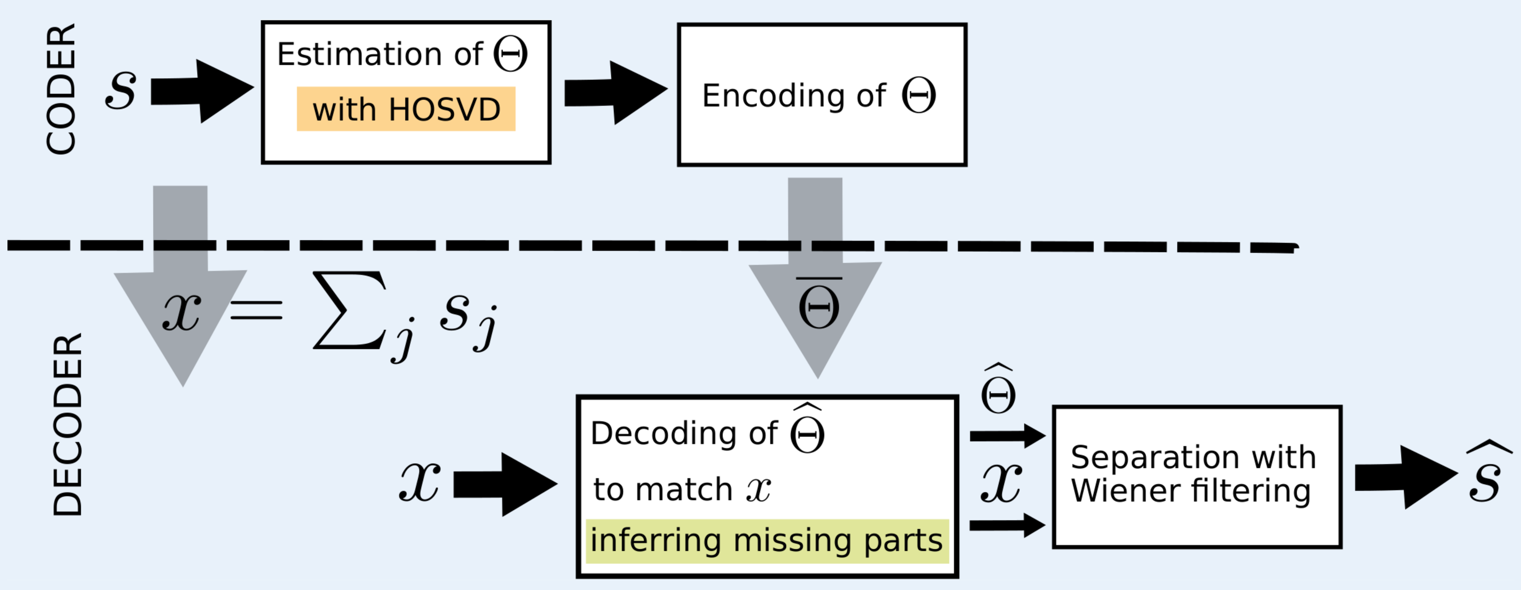
Introduction



Reference ISS method [1] consists of two steps:

- **Coder**: Sources and mix perfectly known. Nonnegative Tensor Factorization (NTF) of source spectrograms with NTF. Encoding of resulting parameters Θ.
- **Decoder**: Only mix available. Wiener filtering used to recover sources from mix assisted by transmitted NTF parameters $\overline{\Theta}$.

Dimension reduction for ISS



Restriction of classical NTF-based ISS methods:
Temporal dimension of source spectrogram can be very large (e.g. for full-length tracks).

$\Rightarrow$ Growth of temporal dimension of NTF parameters (bitrate) and computational complexity of NTF.

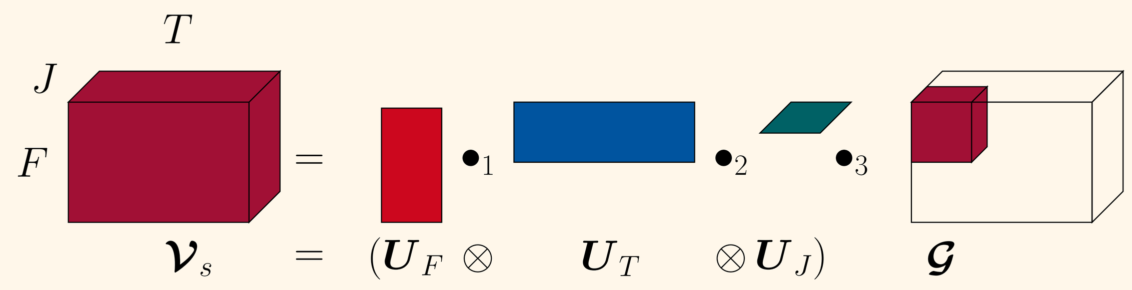
Proposed solution:

- Replace NTF with Higher-order SVD in coder.
- Do not send temporal parameter but estimate it at decoder.

$\Rightarrow$ Size of side-information independent of track length!

Higher-order SVD at coder

Replace classical NTF with higher-order SVD model:



Randomized SVD reduces memory costs and numerical complexity:
$\mathcal{O}(FT(R_F + R_T))$ (vs. classical NTF with R components: $\mathcal{O}(FTJR)$)

Transmitted parameters do *not* include U_T but only U_F, U_J and $\mathcal{G}$.

Parameter estimation at decoder

Model of mixture $\mathcal{V}_x \approx (U_F \otimes U_T \otimes u_J)\,\mathcal{G}$ with $u_J = \sum_{j=1}^{J} [U_J]_{j:}$

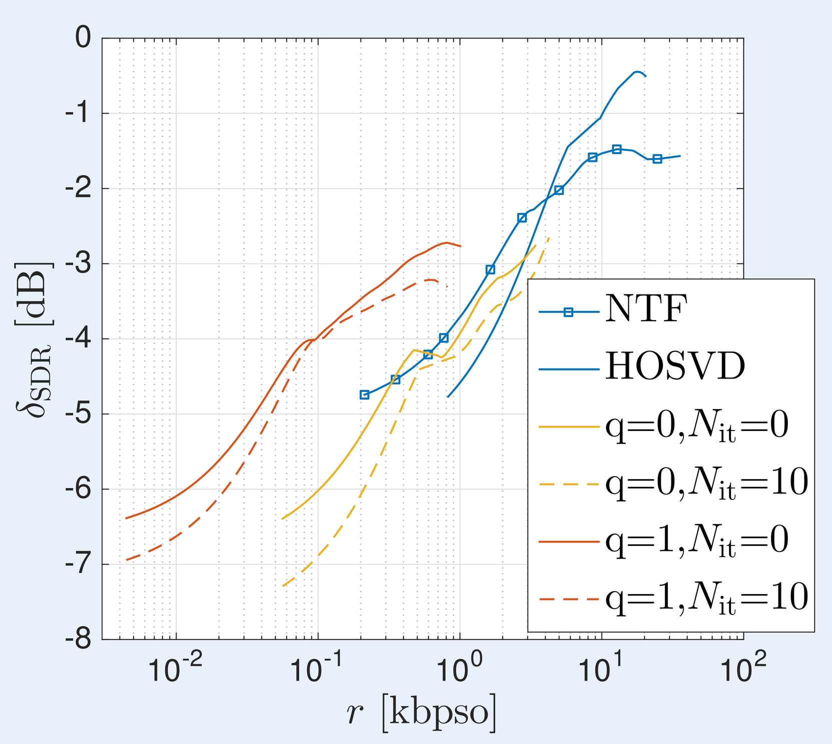
Temporal parameter estimation:
Temporal matrix U_T is not transmitted. Estimate U_T given the other transmitted parameters $\overline{\Theta}$.

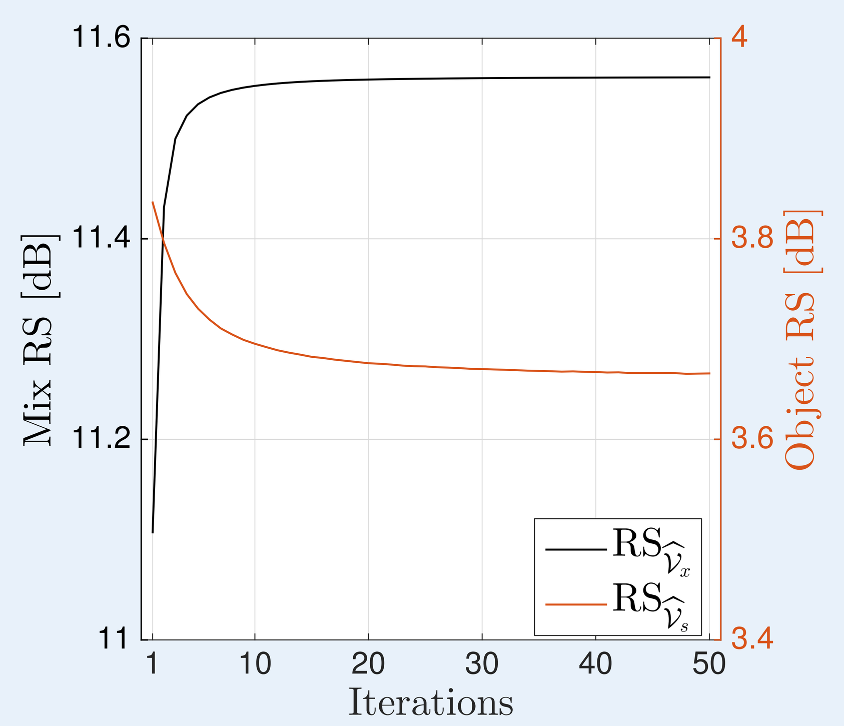
$$\widehat{U}_T \leftarrow \underset{U_T^{\mathsf{T}} U_T = I}{\operatorname{argmin}} \|\mathcal{V}_x - N_T U_T^{\mathsf{T}}\|_F^2$$

with $N_T = U_F\,(u_J \bullet_3 \mathcal{G})$

Iterative re-estimation of parameters U_F and $\mathcal{G}$ initialized with their quantized value obtained from coder.

Experiments





Reference methods:
Performance of NTF method [1] as well as plain HOSVD without quantization and with full transmission of all HOSVD parameters.

Setup:

- Ten full-length tracks (4 sources, taken from DSD100 database)
- HOSVD operates with $R_F, R_T \in [5, 100]$ and $R_J = 4$
- Quantization disabled ($q = 0$) or enabled ($q = 1$) with $N_{\mathcal{G}}, N_F \in [5, 1000]$ centroids for quantization of $\mathcal{G}$ and U_F
- Number of iterations for re-estimation $N_{\mathrm{it}} \in \{0, 10\}$
- Quality measure: δ_{SDR} over parameter bitrate r and reconstruction scores $\mathrm{RS}_{\widehat{\mathcal{V}}_x} = 10\log(\|\mathcal{V}_x\|_F^2 / \|\mathcal{V}_x - \widehat{\mathcal{V}}_x\|_F^2)$ and $\mathrm{RS}_{\widehat{\mathcal{V}}_s}$ over iterations

Results:

- NTF (—•—) yields performance better than HOSVD (—) with full resolution: it has less parameters
- Not sending U_T but estimating it at decoder, with other parameters full resolution (—), permits to reach NTF performance (—•—)
- Coarse quantization of U_F and $\mathcal{G}$ and not sending U_T (—) leads to remarkable decrease of bitrate by a factor of almost 10.
- Iterative re-estimation of quantized parameters (- - -, - - -) is not increasing performance. Score $\mathrm{RS}_{\widehat{\mathcal{V}}_x}$ (—) is improved by 0.5 dB. Generalization to $\widehat{\mathcal{V}}_s$ does not hold.

[1] 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.

rohlfing@ient.rwth-aachen.de
www.ient.rwth-aachen.de
Institut für Nachrichtentechnik, Melatener Str. 23, 52074 Aachen

ICASSP 2017