Paper #8883

Ray-Space-Based Multichannel Nonnegative Matrix Factorization for Audio Source Separation

10.1109/LSP.2021.3055463

M. Pezzoli, J. J. Carabias-Orti*, M. Cobos**, F. Antonacci, A. Sarti
Politecnico di Milano, Italy - Dipartimento di Elettronica, Informazione e Bioingegneria
*Universidad de Jaén, Spain – Telecommunication Engineering Department
**Universitat de València, Spain – Departamento de Informática

POLITECNICO MILANO 1863 · VNIVERSITAT DE VALÈNCIA · UJA. Universidad de Jaén · spat · ISPL Image and Sound Processing Lab

Abstract

Nonnegative matrix factorization (NMF) has been traditionally considered a promising approach for audio source separation. While standard NMF is only suited for single-channel mixtures, extensions to consider multi-channel data have been also proposed. Among the most popular alternatives, multichannel NMF (MNMF) and further derivations based on constrained spatial covariance models have been successfully employed to separate multi-microphone convolutive mixtures. This letter proposes a MNMF extension by considering a mixture model with Ray-Space-transformed signals, where magnitude data successfully encodes source locations as frequency-independent linear patterns. We show that the MNMF algorithm can be seamlessly adapted to consider Ray-Space-transformed data, providing competitive results with recent state-of-the-art MNMF algorithms in a number of configurations using real recordings.

1. Related Works

Multichannel NMF model (MNMF)

- We consider a uniform linear array (ULA) of I channel acquiring the mixture of J acoustic sources.
- Under the local Gaussian model MNMF describes the mixture at the *ith* channel as

$$y_i(\omega,n) \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_{j=1}^{J} \boldsymbol{G}_j(\omega) \sum_{k}^{K} w_{j,k}(\omega) h_{j,k}(n)\right)$$

- $\boldsymbol{G}_j \in \mathbb{C}^{I\times I}$ is the spatial covariance matrix of the jth source
- $w_{j,k}(\omega), h_{j,k}(n)$ are the basis functions and the activation modeling the source PSD $p_j(\omega,n)$.

Ray Space Transofrm (RST)

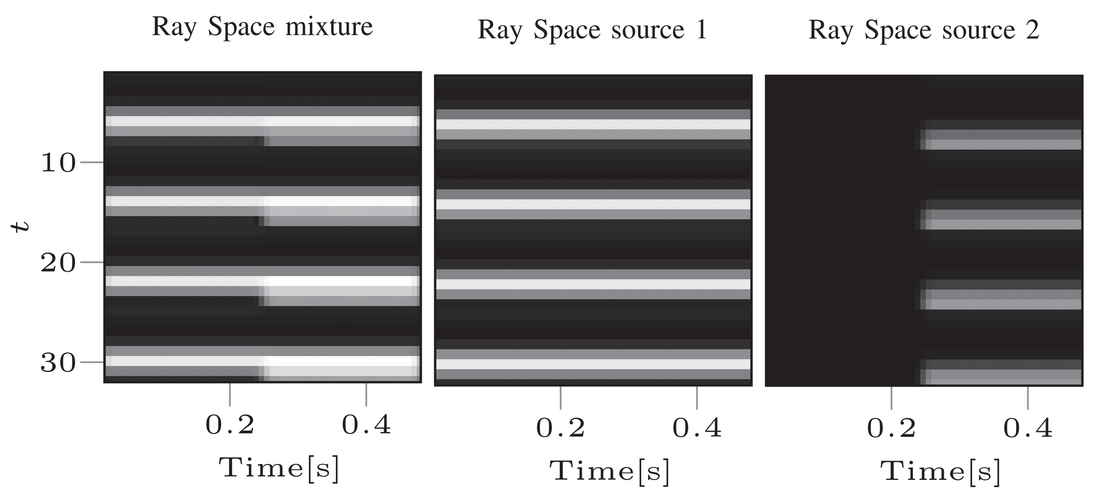
RST [1] is a linear operator $\boldsymbol{\Psi} \in \mathbb{C}^{I\times LD}$ that maps the signals of a ULA onto the Ray Space

$$\boldsymbol{Z}(\omega,n) = \boldsymbol{\Psi}^{\mathrm{H}}(\omega)\boldsymbol{y}(\omega,n).$$

Array signals can be recovered using the inverse RST $\widetilde{\boldsymbol{\Psi}} = (\boldsymbol{\Psi}\boldsymbol{\Psi}^H)^{-1}\boldsymbol{\Psi}$

$$\boldsymbol{y}(\omega,n) \approx \widetilde{\boldsymbol{\Psi}}(\omega)\boldsymbol{Z}(\omega,n).$$

GOAL:
exploit RS representation of source's position as input domain for MNMF separation



- $[\boldsymbol{\Psi}]_{i,t} = e^{-j\omega\frac{d\mu_W}{c\sqrt{1+\mu_W^2}}(i-1)} \psi^*_{l,i}$, $t = 1,\dots,LD$ is the ray space index spanning D directions in L locations of the ULA
- Ray Space consists in the parametrization of line equation $z = \mu x + \nu$ as a function of slope μ and intercept ν
- **Main feature**: acoustic rays emitted by **point sources** are **mapped** onto **lines** in the Ray Space encoding their **location.**

2. The Ray Space MNMF (RS-MNMF)

In presence of J sources the Ray Space data is modelled as

$$Z_t(\omega,n) = \sum_{j=1}^{J} r_{t,j}\, s_j(\omega,n) + b_t(\omega,n),$$

- $r_{t,j}$ describes the jth source contribution at tth Ray Space element.

We employ a general β-divergence cost function

$$\mathcal{C}_{\mathrm{RS}}(\Theta) = \sum_{t,\omega,n} d_\beta(|Z_t(\omega,n)|^2\, \hat{y}_t(\omega,n))$$

- $\hat{y}_t(\omega,n) = \sum_j g_{t,j} \sum_{k\in K_j} w_k(\omega) h_k(n)$ is the square magnitude of the ray space modeled using NMF
- Similarly to instantaneous algorithm in [2] the components are estimated using MU method.

The estimate of the jth source image in the Ray Space is given in terms of MMSE as

$$\tilde{s}_j^{(t)\mathrm{im}}(\omega,n) = \frac{g_{t,j}\, p_j(\omega,n)}{\hat{y}_t(\omega,n)} Z_t(\omega,n)$$

Estimate of the jth **sources** at the **microphones** is obtained using the **inverse RST**:

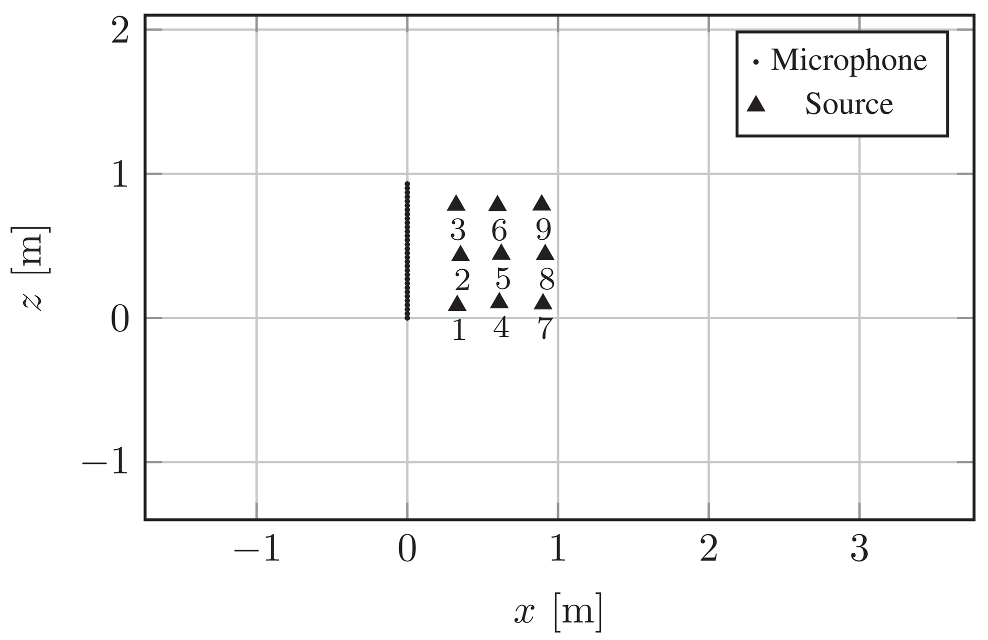
$$\hat{s}_j^{\mathrm{im}}(\omega,n) = \widetilde{\boldsymbol{\Psi}}(\omega,n)\tilde{s}_j^{\mathrm{im}}(\omega,n)$$

3. Results

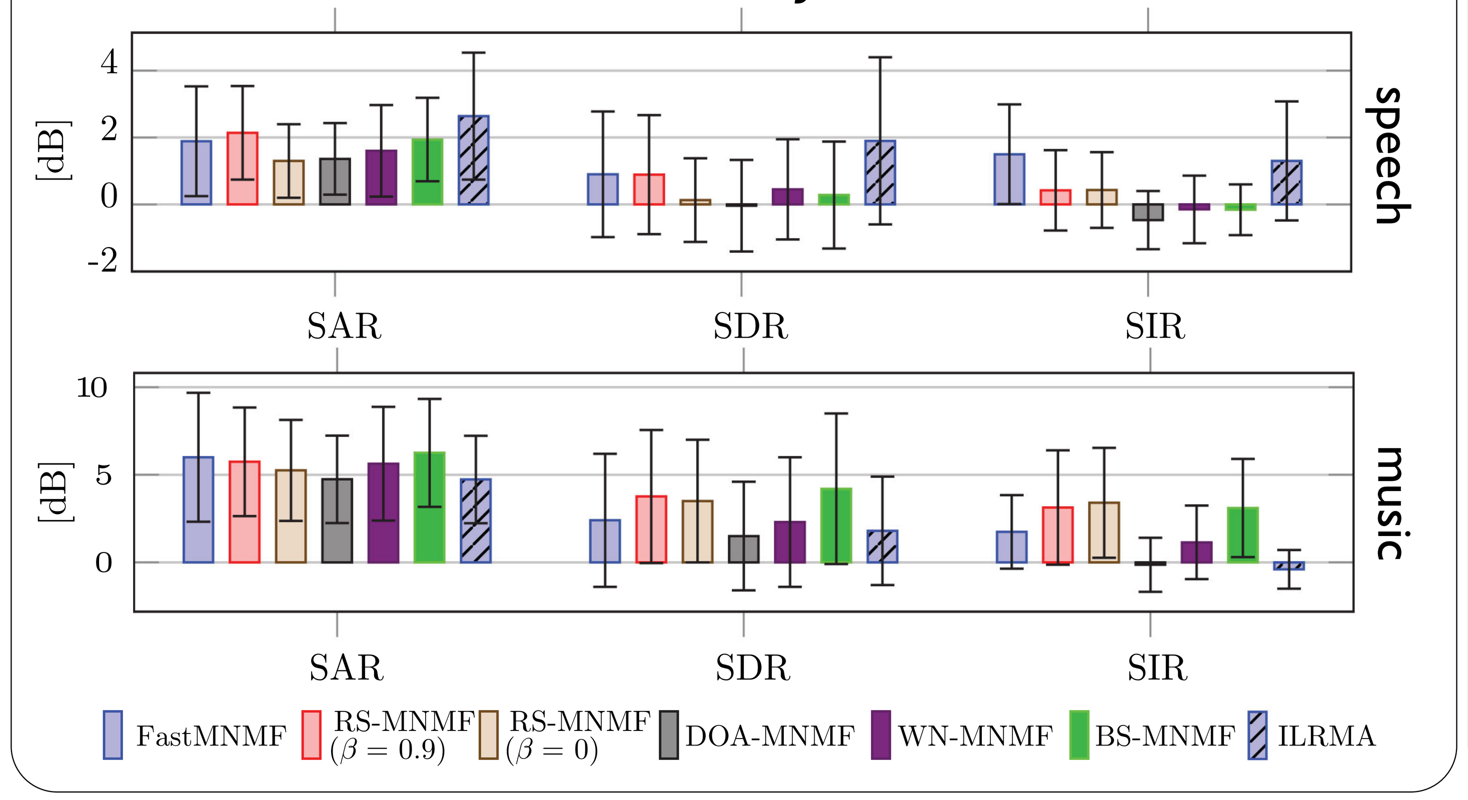
Setup and metrics

- 5.5m×3.4m×3.3m Room with $T60 \approx 0.4$s
- ULA of $I = 32$ microphones and 9 source locations
- Mixtures with $J = \{2,3\}$ sources with 3s signals of male/female speech and music
- Results compared with BS-MNMF[2,] FastMNMF[3], DOA-MNMF[4], WN-MNMF[5] and ILRMA[6].
- Performance is evaluated in terms of :
  - Signal-to-artifacts ratio (SAR),
  - Signal-to-distortion ratio (SDR),
  - Signal-to-interference ratio (SIR).

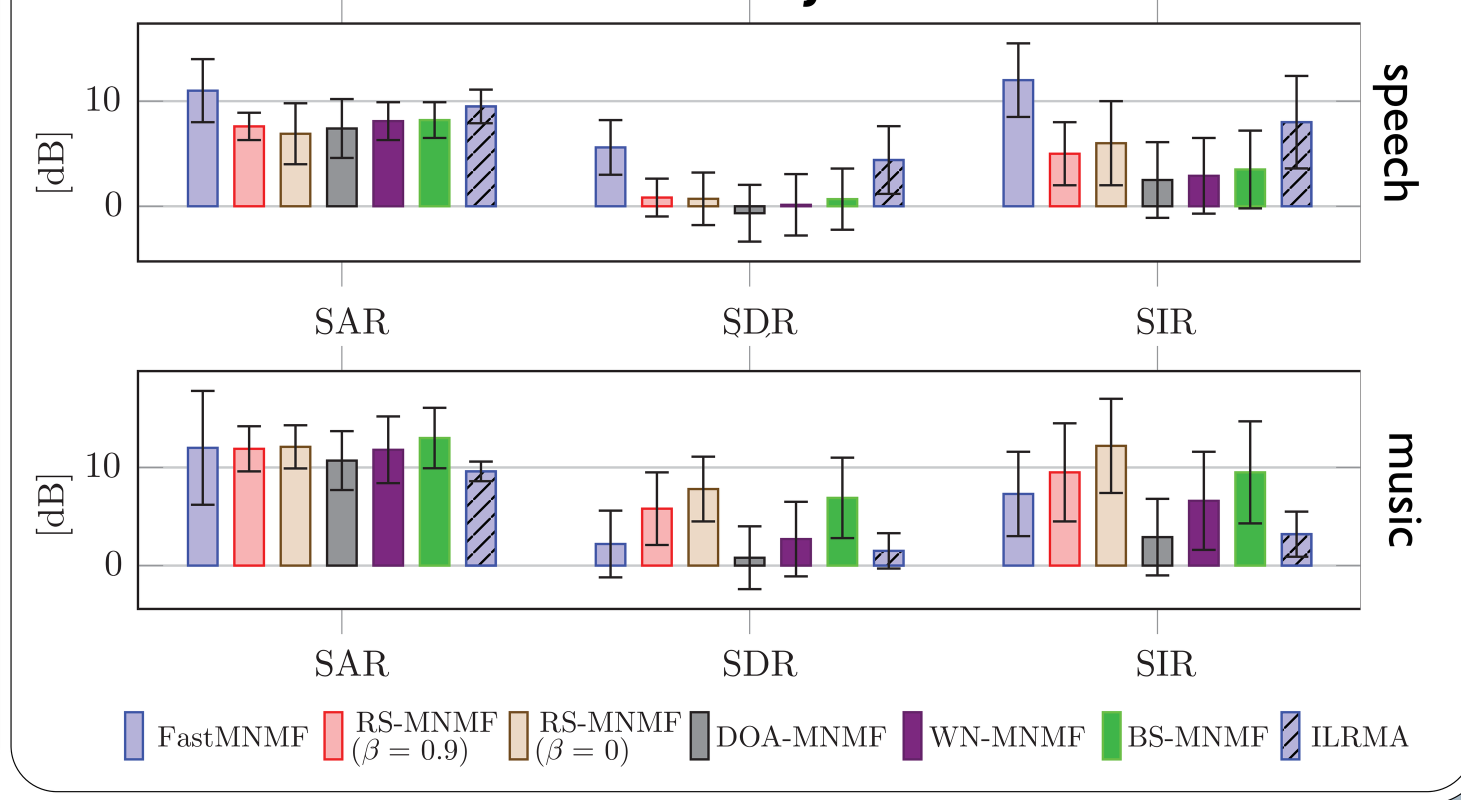
https://github.com/polimi-ispl/rs-mnmf



Results with J=3 sources



Results with J=2 sources



References

[1] L.Bianchi,F.Antonacci,A.Sarti,andS.Tubaro,"The rayspace transform: A new framework for wave field processing," IEEE Trans. Signal Process., vol. 64, no. 21, pp. 5696–5706, Nov. 2016.
[2] S.Lee,S.H.Park,andK.Sung,"Beamspace-domain multichannelnon-negative matrix factorization for audio source separation," IEEE Signal Process. Lett., vol. 19, no. 1, pp. 43–46, Jan. 2012.
[3] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 2610–2625, 2020
[4] J. J. Carabias-Orti, J. Nikunen, T. Virtanen, and P. Vera-Candeas, "Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 26, no. 9, pp. 1512–1527, Sep. 2018.
[5] Y. Mitsufuji, S. Uhlich, N. Takamune, D. Kitamura, S. Koyama, and H. Saruwatari, "Multichannel non-negative matrix factorization using banded spatial covariance matrices in wavenumber domain," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 28, pp. 49–60, 2020.
[6] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 9, pp. 1626–1641, Sep. 2016.

2022 IEEE International Conference on Acoustics Speech and Signal Processing, Singapore, Singapore, 22-27 May 2022

icassp 2022 Singapore