Applied Acoustics 218 (2024) 109888

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Spherical-harmonics-based sound field decomposition and multichannel NMF for sound source separation ☆,☆☆

Mirco Pezzoli [a,*], Julio Carabias-Orti [b,**], Pedro Vera-Candeas [b], Fabio Antonacci [a], Augusto Sarti [a]

[a] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, via Ponzio 34/5, Milan, 20133 MI, Italy
[b] Telecommunication Engineering Department, University of Jaen, Linares, 23700, Spain

ARTICLE INFO

Keywords:
Sound field separation
Non-negative matrix factorization (NMF)
Spherical harmonics
Blind source separation

ABSTRACT

In the context of source separation solutions for virtual reality applications, several techniques in the spherical harmonics domain have been proposed in the literature. The performance of such methods is limited under high reverberation conditions and the rendering of the obtained spatial sound is fixed to the recording location only. Recently, novel sound field works in the literature proposed a global representation that enables both the direct sound (exterior field) estimation and the reconstruction in locations different from the acquisition ones. In this paper, we propose a signal processing framework based on Multichannel Non-Negative Matrix Factorization in the spherical harmonics domain that operates directly over the exterior field coefficients enabling the reconstruction of the direct sound field of the separated sources. To evaluate our proposal, we compared with other state-of-the-art source separation approaches using several setups and including different reverberation conditions, showing promising results in terms of BSS_eval metrics.

1. Introduction

The high-order ambisonics (HOA) format is a popular choice for virtual reality (VR) applications due to its efficient transmission, uniform sound scene coverage, and ease of acoustic scene rotation [1]. In VR applications, object-based encoding is utilized to enable immersive experiences with six degrees of freedom (6DOF) as proposed in MPEG-I [2]. However, sound source separation (SSS) is necessary to convert from HOA (3DOF) to object-based (6DOF) format. Numerous approaches employing SSS techniques have been proposed for this purpose. While beamforming methods have been employed in the spherical harmonics (SH) domain [3–6], they require prior knowledge or estimation of the direction of arrival (DOA) for each source and may be susceptible to low-frequency noise with small array radii. In [7], Independent Component Analysis (ICA) was proposed as an alternative approach. To address underdetermined scenarios with more sources than channels, some researchers have explored the use of Multichannel NMF (MNMF) in the SH domain. In [8], the authors proposed an MNMF model using a weighted combination of DOA kernels in the ambisonics SH domain to leverage spatial properties, while modeling source spectrograms with a standard NMF structure. In [9], this approach was combined with deep neural network for the task of singing voice separation under a supervised scenario. In [10], the authors extended [8] using a convolutive NTF model to deal with reverberant conditions. DOA information is also exploited for source separation in [11] using an end-to-end deep learning model in the ambisonics SH domain.

A novel approach to efficient blind SSS has been introduced in [12], which utilizes Non-negative Tensor Factorization (NTF). The ap-

☆ This work has been funded by “REPERTORIUM” project. Grant agreement number 101095065. Horizon Europe. Cluster II. Culture, Creativity and Inclusive Society. Call HORIZON-CL2-2022-HERITAGE-01-02.
☆☆ This work was supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”).
* Principal corresponding author.
** Corresponding author.
E-mail addresses: mirco.pezzoli@polimi.it (M. Pezzoli), carabias@ujaen.es (J. Carabias-Orti), pvera@ujaen.es (P. Vera-Candeas), fabio.antonacci@polimi.it (F. Antonacci), augusto.sarti@polimi.it (A. Sarti).

https://doi.org/10.1016/j.apacoust.2024.109888
Received 1 August 2023; Received in revised form 28 December 2023; Accepted 21 January 2024
Available online 30 January 2024

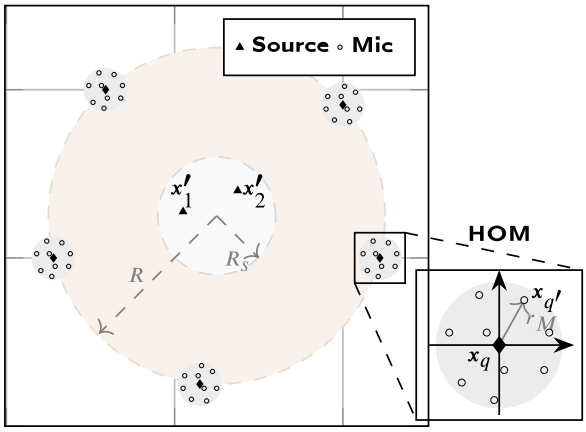

Fig. 1. Graphical representation of HOMs setup. The ROI (light red) is surrounded by a set of HOMs, while the sources are confined inside the region of radius R_s.

proach employs diagonalization over the full-rank signal model in the SH domain to represent the interior sound field, thereby reducing the computational cost needed for MNMF. Unlike the MNMF approaches mentioned earlier, this study incorporates the modeling of near-field sources, in addition to far-field sources, by leveraging the properties of the spherical Hankel function. To improve the performance of the NTF cost function, a masking scheme is utilized to eliminate both a noisy (evanescent) region and the high-power region. However, these methods experience a significant decrease in performance when faced with high reverberation conditions.

The main limitation of the customary techniques lies in the fact that they lack of a global sound field representation, hence their performance is constrained to the recording location. Recently, sound field representations have been adopted in order to combine and exploit the information of multiple arrays [13–16]. In this context, spherical harmonics expansion provides a convenient model of the acoustic field [15,17–19] because the direct sound of sources confined in a region can be inherently described in terms of exterior field. Oppositely, interferers and reverberation coming from the outside of the source region are modelled as an interior field. The authors in [15] employed higher order microphone arrays (HOMs) to split sound field into its interior and exterior elements. Unfortunately, the required dense spatial sampling makes the adoption of [15] difficult for practical implementations, for instance more than 600 sensors are needed to cover a region of 1 m at frequency 1000 Hz. Various approaches have been suggested to address the significant hardware and computational demands by reducing the number of sensors required. For instance, in [17], the authors utilize Room Impulse Responses (RIRs) to represent reverberation as an interior field based on the collected data. Moreover, in [20], synthetic free field impulse responses are combined with the previously measured RIRs to limit the size of the dictionary. It is worth noting that although the sound field reconstruction techniques are able to estimate the acoustic field in arbitrary locations, in the case of multiple sound sources, the contribution of each individual source is not readily available.

In this work, we propose a framework based on the exterior/interior field decomposition to increase the robustness of the state-of-the-art audio signal processing techniques against reverberation. In particular, we propose a MNMF-based signal decomposition approach that operates directly on the exterior field coefficients, allowing the separation of target sources within an acoustic region and the reconstruction of the sound field of a single source at any spot outside the source region. The adoption of a global sound field representation allows us to reduce the reverberant component of the sound field in the array signals performing the separation on the dry signal only. Moreover, the employed sound field model enables the combined use of all the HOMs in the optimization algorithm, instead of limit the MNMF to each single array.

We evaluate the proposed framework with several setups with different number of arrays, sources, locations and T60. To demonstrate the potential of the proposed framework, comparisons with other state-of-the-art methods have been conducted at the microphone level.

The rest of the paper is organized as follows. Section 2 formulates the problem in the spherical harmonics domain and reviews the foundaments of MNMF. The proposed framework to model the exterior field coefficients using NMNF is presented in Section 3. Simulation results and comparisons with other state-of-the-art source separation methods are presented in Section 4. Finally, Section 5 concludes the paper.

2. Data model

2.1. Sound field model

Let us assume that Q HOMs of Vth order are distributed and enclose a circular region of interest (ROI) with radius R (see Fig. 1). S active sound sources are contained in a circular region concentric with the ROI of radius $R_s \leq R$. Within the ROI, the sound field at a generic location $\boldsymbol{x} = [r, \theta, \phi]^T$ is generally given superimposing two components as [19]

$$y(\boldsymbol{x}, f, t) = y_E(\boldsymbol{x}, f, t) + y_I(\boldsymbol{x}, f, t), \tag{1}$$

where $y_E(\boldsymbol{x}, f, t)$ models the exterior (direct) sound from the source region and $y_I(\boldsymbol{x}, f, t)$ is referred to as the interior field from the outside of the ROI, with f temporal frequency and t temporal index. Hence, the interior component models sources outside the ROI and reflections. Conveniently, both components in (1) can be expressed using spherical harmonics expansion [21]

$$y_E(\boldsymbol{x}, f, t) = \sum_{n=0}^{N_E} \sum_{m=-n}^{n} \beta_{nm}(f, t) h_n\left(2\pi \frac{f}{c} r\right) Y_{nm}(\theta, \phi), \tag{2}$$

$$y_I(\boldsymbol{x}, f, t) = \sum_{n=0}^{N_I} \sum_{m=-n}^{n} \alpha_{nm}(f, t) j_n\left(2\pi \frac{f}{c} r\right) Y_{nm}(\theta, \phi), \tag{3}$$

where $Y_{nm}(\cdot)$ defines the spherical harmonic of order n and degree m, $h_n(\cdot)$ is the nth order spherical Hankel function of the second kind and $j_n(\cdot)$ is the nth order spherical Bessel functions of the first kind, with c the sound speed. The limits of the expansions are given by $N_E = \lceil 2\pi \frac{f}{c} e R_s / 2 \rceil$ and $N_I = \lceil 2\pi \frac{f}{c} e R / 2 \rceil$ [22]. The terms $\beta_{nm}(f, t)$ in (2) contains the exterior coefficients, while $\alpha_{nm}(f, t)$ in (3) are the interior ones. The coefficients are known as "global" since they completely determine the acoustic field in the ROI. The signals of the qth HOM, analyzing the sound field, are expressed using spherical harmonics as [18]

$$a_{\nu,\mu}^{(q)}(f, t) = \frac{1}{b_\nu\left(2\pi \frac{f}{c} r_M\right)} \sum_{q'=1}^{Q'} y^{(q)}(\boldsymbol{x}_{q'}, f, t) Y_{\nu\mu}^*(\theta_{q'}, \phi_{q'}), \tag{4}$$

where Q' is the number of sensors in the HOM, $y^{(q)}(\boldsymbol{x}_{q'}, f, t)$ is the signal (1) of the sensor at $\boldsymbol{x}_{q'} = [r_M, \theta_{q'}, \phi_{q'}]^T$ (referred to the origin of the qth array), while $b_\nu(\cdot)$ is a term depending on the array type [21], i.e., consisting of an open or rigid sphere. Note that the expansion order in (4) $\nu = 0, \ldots, V$, is given by the array order V, while $\mu = -\nu, \ldots, \nu$. The coefficients a in (4) are referred to as "local" since they are computed with respect to each HOM reference system. The local coefficients are related to the global sound field ones (2), (3) through the well-known spherical harmonics translation as [15]

$$\mathbf{a}(f, t) = \left[\mathbf{T}_E(f), \mathbf{T}_I(f)\right] \begin{bmatrix} \boldsymbol{\beta}(f, t) \\ \boldsymbol{\alpha}(f, t) \end{bmatrix} \tag{5}$$

where $\mathbf{a} \in \mathbb{C}^{Q(V+1)^2 \times 1}$ is a vector collecting coefficients from all Q HOMs, $\mathbf{T}_E(f) \in \mathbb{C}^{Q(V+1)^2 \times (N_E+1)^2}$ and $\mathbf{T}_I(f) \in \mathbb{C}^{Q(V+1)^2 \times (N_I+1)^2}$ are translation matrices [18] that connect the exterior field coefficients $\boldsymbol{\beta} \in \mathbb{C}^{(N_E+1)^2 \times 1}$ and the interior field coefficients $\boldsymbol{\alpha} \in \mathbb{C}^{(N_I+1)^2 \times 1}$, respectively. It follows that by inverting the relation in (5) and ignoring

the interior components, a naive estimate of the direct sound of the sources (exterior field) can be obtained as

$$\hat{\boldsymbol{\beta}}(f,t) = \mathbf{T}_E^{\dagger}(f)\mathbf{a}(f,t), \tag{6}$$

which however suffers from leakage of the interior sound [20]. Therefore, different approaches for the estimation of the global coefficients have been proposed in order to improve the inversion performance [17,19,20].

2.2. Multichannel NMF

Multichannel NMF (MNMF) can be expressed using the well-known local Gaussian model (LGM) which enables the systematic modeling and combination of spatial and spectral cues. In fact, under LGM modelling [23], the spatial image of the $Q(V+1)^2$-channel mixture of S multiple and mutually independent sources $\mathbf{y}(f,t) \in \mathbb{C}^{Q(V+1)^2}$ is represented as a sum of complex Gaussians, i.e.,

$$\mathbf{y}(f,t) = \sum_{s=1}^{S} y_s(f,t) \sim \mathcal{N}_{\mathbb{C}}\left(0, \sum_s \mathbf{H}_s(f,t)\lambda_s(f,t)\right), \tag{7}$$

with a positive-definite Hermitian covariance matrix which is the sum of components resulting from the multiplication between: 1) a spatial covariance parameter $\mathbf{H}_s(f,t) \in \mathbb{C}^{Q(V+1)^2 \times Q(V+1)^2}$, $s = 1,\dots,S$, modeling the spatial characteristics of the sth source image at the TF point (f,t), and 2) a parameter $\lambda_s(f,t) \in \mathbb{R}$, $s = 1,\dots,S$ representing the spectral variance of source image s i.e., its spectro-temporal structure at (f,t) TF point.

Assuming static sources, it is commonly accepted to consider that the spatial covariances remain time-invariant, i.e., $\mathbf{H}_s(f,t) = \mathbf{H}_s(f)$ [23,24]. Moreover, a classical NMF scheme can be employed to model the spectral variance as

$$\lambda_s(f,t) = \sum_{k=1}^{K} u_{sk} b_k(f) g_k(t), \tag{8}$$

where $b_k(f)$ and $g_k(t)$ represent the basis functions and the time-varying gains, respectively, for each NMF component $k \in [1,K]$. The parameter u_{sk} associates sources and NMF components.

The model parameters $\boldsymbol{\theta} = \{\mathbf{H}_s(f), \lambda_s(f,t)\}$ can be obtained from the observed data using the maximum likehood (ML) criterion as

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \boldsymbol{\theta}'). \tag{9}$$

As a matter of fact, under LGM modeling, (9) is equivalent to minimizing the generalized Itakura Saito (IS) cost function:

$$C_{IS}(\theta) = \sum_{f,t} \operatorname{tr}\left(\hat{\Sigma}_{y,ft}\Sigma_{y,ft}^{-1}\right) - \log\det\left(\hat{\Sigma}_{y,ft}\Sigma_{y,ft}^{-1}\right) - Q(V+1)^2, \tag{10}$$

with $\hat{\Sigma}_{y,ft} = \mathbf{y}_{ft}\mathbf{y}_{ft}^H$ and $\Sigma_{x,ft} = \mathbf{H}_s(f)\lambda_s(f,t)$. Alternatively, generalization of the Euclidean NMF to the multichannel scenario can be formulated using the following observation model:

$$p\left(\hat{\Sigma}_{y,ft}, \theta\right) \sim \prod_{f,t} \mathcal{N}_{\mathbb{C}}\left(\hat{\Sigma}_{y,ft} \mid \Sigma_{y,ft}, 1\right). \tag{11}$$

In fact, maximizing the log of the likelihood (11) is equivalent to minimizing the multichannel Euclidean cost function

$$C_{FRB}(\theta) = \sum_{f,t} \left\| \hat{\Sigma}_{y,ft} - \Sigma_{y,ft} \right\|_F^2, \tag{12}$$

where $\|\cdot\|_F^2$ represents the squared Frobenius norm.

Then, the update rules can be obtained using classical algorithms such as the expectation-maximization (EM) [25] or the majorization-minimization (MM) [26]. However, updating the model parameters incurs a significant computational cost of $O(I^3)$ due to the multiple matrix inversions required during the updates [27]. This limits the practical use of this framework when the number of channels increases. Several techniques based on diagonalization have been proposed to tackle this issue and provide efficient computational solutions [12,28–30]. However, these methods come at a cost of being limited to certain array setups [30], reducing to only the diagonal values of the spatial covariance matrix (SCM) [12], or relying on statistical independence between the sources to derive the spatial characteristics [28,29].

3. Proposed framework

3.1. Spherical harmonics signal model

Let us introduce a model for the exterior field coefficients (2) that similarly to [17,19,20], expresses $\boldsymbol{\beta}$ as a sum of O omnidirectional equivalent sources [31,32]

$$\boldsymbol{\beta}(f,t) = \mathbf{D}(f)\mathbf{w}(f,t) \Rightarrow \mathbf{w}(f,t) = \mathbf{D}^{\dagger}(f)\boldsymbol{\beta}(f,t), \tag{13}$$

where $\mathbf{w} \in \mathbb{C}^{O\times 1}$ are the weights that model the spectro-temporal evolution of the sources, while $\mathbf{D}(f) \in \mathbb{C}^{(\tilde{N}_E+1)^2 \times O}$ is the matrix of elements

$$\mathbf{D}_{\imath,o}(f) = -ik\sqrt{4\pi} j_n\left(2\pi \frac{f}{c} r'_o\right) Y_{nm}^*(\theta'_o, \phi'_o), \tag{14}$$

expressing the translation [33] of the o-th Green's function from $\boldsymbol{x}'_o = [r'_o, \theta'_o, \phi'_o]^T$ to the origin with $\imath = 1,\dots,(\tilde{N}_E+1)^2$ the index of degrees and orders. The matrix $\mathbf{D}(f)$ corresponds to a dictionary of equivalent sources obtained by sampling the source region. Note that, in order to reduce the number of coefficients, it could be convenient to use an expansion order $\tilde{N}_E$ in (13) that is different from the actual order N_E in (2). This comes at the cost of lower reconstruction accuracy. It follows that using (13) we can relate the signals of the sensors with the exterior field as [20]

$$\mathbf{a} = \mathbf{P}\mathbf{D}\mathbf{w} = \boldsymbol{\Gamma}\mathbf{w} = \boldsymbol{\Gamma}\mathbf{D}^{\dagger}\boldsymbol{\beta}, \tag{15}$$

where $\mathbf{P} \in \mathbb{C}^{Q(V+1)^2 \times (\tilde{N}_E+1)^2}$ models the propagation term of SH expansion [19] and $\boldsymbol{\Gamma} = \mathbf{P}\mathbf{D}$. Note that in (15) the dependency on time and frequency has been omitted for simplicity. Therefore, we adopt as input of the MNMF optimization a first estimate of the exterior coefficients obtained through (15), as proposed in [17,20]

$$\hat{\boldsymbol{\beta}}(f,t) = \left[\boldsymbol{\Gamma}(f)\mathbf{D}(f)^{\dagger}\right]^{\dagger}\mathbf{a}(f,t). \tag{16}$$

3.2. Exterior field MNMF model

Let us assume that the exterior field model from (16) can be modeled using a LGM as described in Sec. 2.2:

$$\hat{\mathbf{B}}(f,t) = \mathbb{E}\left[\hat{\boldsymbol{\beta}}(f,t)\hat{\boldsymbol{\beta}}(f,t)^H\right] \approx \sum\nolimits_{s,o} \mathbf{H}_o(f) w_{so}(f,t), \tag{17}$$

where $\mathbf{H}_o(f) = \left[\left[\boldsymbol{\Gamma}\mathbf{D}^{\dagger}\right]^{\dagger}\boldsymbol{\Gamma}\right]\left[\left[\boldsymbol{\Gamma}\mathbf{D}^{\dagger}\right]^{\dagger}\boldsymbol{\Gamma}\right]^H$ is a SCM in the SH domain ($\mathbf{H}_o(f) \in \mathbb{C}^{(\tilde{N}_E+1)^2\times(\tilde{N}_E+1)^2}$) modeling the fixed dictionary of Green's functions (14) and the source weight parameter $w_{so}(f,t) \in \mathbb{R}^+$. Here, we consider modeling the source weight parameters using a classical NMF structure $w_{so}(f,t) = z_{os}u_{sk}b_k(f)g_k(t)$ where z_{os} represents the spatial weight parameter that associates sources and locations, while the other parameters u_{sk}, $b_k(f)$ and $g_k(t)$ model the sources spectro-temporal structure. In fact, this model can be seen as an extension of the original DOA-based model from [34] which has also been applied in the Ambisonics domain [8].

3.3. Derivation of update rules

The free parameters $\theta = \{\mathbf{z}, \mathbf{u}, \mathbf{b}, \mathbf{g}\}$ in (17) can be estimated by minimizing a cost function (e.g., (10) or (12)). Unfortunately, as explained in Section 3.2, the computational cost associated to the matrix inversion operations of the SCM drastically limits the SH order of the model. This drawback can be mitigated using diagonalization techniques such

Table 1
Average SIR and SAR values for the two-sources scenarios.

Setup	T60 = 0.3 s, $Q = 4$		T60 = 0.3 s, $Q = 8$		T60 = 0.3 s, $Q = 16$		T60 = 0.6 s, $Q = 4$		T60 = 0.6 s, $Q = 8$		T60 = 0.6 s, $Q = 16$		T60 = 1.2 s, $Q = 4$		T60 = 1.2 s, $Q = 8$		T60 = 1.2 s, $Q = 16$	
Method	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR
Female Speech																		
Proposed $\tilde{N}' = 1$	9.8	12.8	13.1	12.9	15.3	12.8	12.1	13.1	12.0	13.3	14.5	13.3	12.1	13.2	12.0	**13.6**	15.2	**13.4**
Proposed $\tilde{N}' = 2$	**18.0**	13.5	12.6	**13.6**	**16.9**	13.1	15.8	14.2	12.5	13.3	**16.4**	12.9	13.6	**14.8**	12.5	13.1	**15.9**	12.5
Proposed $\tilde{N}' = 3$	16.0	**14.1**	**15.2**	**13.6**	16.6	**13.5**	**16.2**	**14.5**	**14.7**	**13.4**	15.7	**13.1**	**14.8**	**14.8**	**14.0**	13.1	15.1	12.9
ILRMA	9.4	9.8	1.8	7.4	7.8	7.9	8.2	10.4	7.4	9.9	3	8.1	7.5	11.5	8.9	11.8	4.0	9.0
FastMNMF	13.0	8.6	8.7	5.4	-0.4	4.1	11.7	9.3	7.1	7.1	-2.3	4.8	11.2	9.4	8.5	7.9	1.4	4.5
Male Speech																		
Proposed $\tilde{N}' = 1$	15.7	12.3	14.0	12.2	**13.9**	12.8	14.9	11.4	16.1	11.8	**15.6**	**12.9**	14.3	10.4	15.0	10.8	**14.8**	11.9
Proposed $\tilde{N}' = 2$	17.0	12.9	**17.9**	13.6	6.4	11.2	15.4	12.0	**16.8**	12.9	10.3	11.6	14.5	10.9	**15.3**	11.5	12.5	11.8
Proposed $\tilde{N}' = 3$	**17.4**	12.9	9.6	11.2	11.3	11.4	**15.9**	**11.7**	10.1	11.7	13.0	11.9	**14.9**	10.4	14.7	**11.9**	14.1	**12.9**
ILRMA	10.6	**14.9**	10.6	**15.0**	10.6	**14.9**	2.3	10.8	2.5	**13.7**	2.1	12.7	0.4	10.3	-2.1	11.2	-2.1	11.3
FastMNMF	-1.2	6.9	3.8	0.9	7.0	-0.6	11.1	11.3	10.3	0.8	-3.3	0.2	10.2	**12.3**	11.5	0.4	-4.4	0.9

Table 2
Average SIR and SAR values for the three-sources scenarios.

Setup	T60 = 0.3 s, $Q = 4$		T60 = 0.3 s, $Q = 8$		T60 = 0.3 s, $Q = 16$		T60 = 0.6 s, $Q = 4$		T60 = 0.6 s, $Q = 8$		T60 = 0.6 s, $Q = 16$		T60 = 1.2 s, $Q = 4$		T60 = 1.2 s, $Q = 8$		T60 = 1.2 s, $Q = 16$	
Method	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR	SIR	SAR
Female Speech																		
Proposed $\tilde{N}' = 1$	8.7	12.6	11.3	11.5	**12.5**	12.5	7.3	11.8	9.6	11.0	11.0	12.1	7.0	11.2	8.4	10.5	9.7	**11.6**
Proposed $\tilde{N}' = 2$	7.7	**13.3**	11.2	**12.9**	11.4	**13.4**	10.4	12.6	11.3	12.4	10.7	**12.5**	7.6	12.5	10.0	12.1	9.7	10.4
Proposed $\tilde{N}' = 3$	**10.9**	12.1	**12.2**	11.6	12.4	11.9	**12.3**	12.0	**13.0**	11.9	**13.8**	12.0	4.9	10.2	**11.5**	11.1	**12.3**	11.5
ILRMA	4.9	8.1	0.0	7.8	0.6	7.9	5.2	7.1	-0.8	8.1	-0.5	7.5	3.4	7.6	-0.6	7.8	-1.4	7.5
FastMNMF	10.4	11.3	12.0	10.2	4.7	6.7	11.2	**13.9**	11.5	**12.7**	7.4	6.1	**13.9**	**16.5**	10.9	**13.9**	5.4	6.2
Male Speech																		
Proposed $\tilde{N}' = 1$	-5.4	8.0	-1.3	7.3	**9.5**	**10.7**	9.8	8.8	-2.1	7.4	6.9	9.4	6.9	8.9	4.4	8.1	7.7	9.0
Proposed $\tilde{N}' = 2$	-0.2	8.0	0.2	8.4	2.0	8.4	9.3	8.2	9.3	9.3	0.5	8.4	-5.5	9.3	7.7	9.0	**12.0**	**11.3**
Proposed $\tilde{N}' = 3$	0.8	8.7	**13.4**	10.3	8.8	9.2	1.2	8.2	**12.9**	**10.1**	**11.6**	**11.1**	-1.8	8.6	**11.7**	**10.3**	11.2	10.2
ILRMA	2.3	7.1	5.3	**11.6**	4.0	9.7	-1.7	7.5	-0.5	10.0	-1.4	9.6	-4.8	8.3	-4.7	9.2	-4.3	9.3
FastMNMF	**17.2**	**13.0**	8.8	9.5	7.4	4.9	**15.2**	**13.2**	5.3	7.8	6.9	6.2	**14.2**	**12.9**	3.5	8.3	7.0	5.7

as the ones proposed in [12,29]. Alternatively, since the SCM matrix in the proposed model ($\mathbf{H}_o(f)$ in (17)) is a fixed parameter, we adopt the majorization-minimization algorithm as in [26] to minimize the generalized Euclidean distance (12) to keep the algorithm computationally feasible. The update rules are obtained using an auxiliary function f^+

$$f^+(\theta) = \sum_{f,t} w_{so}(f,t)^2 \mathrm{tr}\left(\mathbf{H}_o(f)\mathbf{R}_{so}(f,t)^{-1}\mathbf{H}_o(f)^H\right) - \sum_{so} w_{so}(f,t)\mathrm{tr}\left(\hat{\mathbf{B}}(f,t)\mathbf{H}_o(f)^H\right) - \sum_{so} w_{so}(f,t)\mathrm{tr}\left(\mathbf{H}_o(f)\hat{\mathbf{B}}(f,t)^H\right), \tag{18}$$

with auxiliary variables $\mathbf{R}_{so}(f,t)$ that satisfy Hermitian positive definiteness and $\sum_{so}\mathbf{R}_{so}(f,t) = I$. The equality of the auxiliary function and the cost function holds when the auxiliary variables satisfy

$$\mathbf{R}_{so}(f,t) = \hat{\mathbf{B}}(f,t)^{-1}\mathbf{H}_o(f)w_{so}(f,t). \tag{19}$$

Then, the free parameters are estimated by computing the partial derivatives of f^+ w.r.t z_{os}, u_{sk}, $b_k(f)$, and $g_k(t)$. Setting these derivatives at zero, we have the following multiplicative update rules:

$$b_k(f) \leftarrow \frac{\sum_{t,s,o} u_{sk}g_k(t)z_{os}\mathrm{tr}(\hat{\mathbf{B}}(f,t)\mathbf{H}_o(f))}{\sum_{t,s,o} u_{sk}g_k(t)z_{os}\mathrm{tr}(\tilde{\mathbf{B}}(f,t)\mathbf{H}_o(f))}. \tag{20}$$

$$g_k(t) \leftarrow \frac{\sum_{f,s,o} u_{sk}b_k(f)z_{os}\mathrm{tr}(\hat{\mathbf{B}}(f,t)\mathbf{H}_o(f))}{\sum_{f,s,o} u_{sk}b_k(f)z_{os}\mathrm{tr}(\tilde{\mathbf{B}}(f,t)\mathbf{H}_o(f))}. \tag{21}$$

$$u_{sk} \leftarrow \frac{\sum_{f,t,o} b_k(f)g_k(t)z_{os}\mathrm{tr}(\hat{\mathbf{B}}(f,t)\mathbf{H}_o(f))}{\sum_{f,t,o} b_k(f)g_k(t)z_{os}\mathrm{tr}(\tilde{\mathbf{B}}(f,t)\mathbf{H}_o(f))}. \tag{22}$$

$$z_{os} \leftarrow \frac{\sum_{f,t,k} u_{sk}b_k(f)g_k(t)\mathrm{tr}(\hat{\mathbf{B}}(f,t)\mathbf{H}_o(f))}{\sum_{f,t,k} u_{sk}b_k(f)g_k(t)\mathrm{tr}(\tilde{\mathbf{B}}(f,t)\mathbf{H}_o(f))}, \tag{23}$$

where $\hat{\mathbf{B}}$ is the observation model and $\tilde{\mathbf{B}}$ the estimation in (17). The algorithm is implemented by first randomly initializing its parameters, and then repeating the update steps (20)-(23) for a fixed number of iterations. Furthermore, the spatial selector z_{os} is re-scaled after each iteration to ensure that it satisfies the condition $\sum_o z_{os} = 1$.

3.4. Sound source extraction

The exterior field coefficients of each source signal $\tilde{\boldsymbol{\beta}}^s(f,t)$ are reconstructed using a generalized Wiener filter [26] for each degree and order index ι as follows:

$$[\tilde{\boldsymbol{\beta}}^s(f,t)]_\iota = \frac{\sum_o [\mathbf{H}_o(f)]_{\iota\iota}[w_{so}(f,t)]_s}{\sum_{so}[\mathbf{H}_o(f)]_{\iota\iota}w_{so}(f,t)}[\hat{\boldsymbol{\beta}}(f,t)]_\iota. \tag{24}$$

Note that the weights of the monopoles $w_{so}(f,t)$ are independent of the SH order. Therefore, to reduce computational requirements, it is possible to decrease the SH order of the observation model $\hat{\boldsymbol{\beta}}(f,t)$ and the fixed dictionary of Green's functions $\mathbf{H}_o(f)$ to $\tilde{N}'_E$ ($\tilde{N}'_E < \tilde{N}_E$) during the factorization while still obtaining $\tilde{N}_E$-order coefficients in (24) by keeping the original order of $\hat{\boldsymbol{\beta}}(f,t)$ (from (16)) and $\mathbf{H}_o(f)$.

Then, using (2), the exterior field of the sources can be estimated by propagating the coefficients to any point $\boldsymbol{x} = [r,\theta,\phi]^T$ outside the source region.

$$\tilde{y}^s(\boldsymbol{x},f,t) = \sum_{n=0}^{\tilde{N}_E}\sum_{m=-n}^{n} \tilde{\beta}^s_{nm}(f,t)h_n(2\pi\frac{f}{c}r)Y_{nm}(\theta,\phi), \tag{25}$$

where $\tilde{y}^s(\boldsymbol{x},f,t)$ can be considered as the free field response at $\boldsymbol{x}$ due to the excitation of a point source s.

4. Experiments

4.1. Setup

The performance of the proposed technique is compared with respect to state-of-the-art MNMF techniques. In particular, we considered FastMNMF [29] and ILRMA [35] since both do not assume any specific single-array setup. As input mixture for all the considered methods we employed the direct sound estimate in (16). In order to evaluate the separation performance, we run an extensive simulation campaign varying acoustic conditions and number of HOMs.

We simulated RIRs using the image-source method [36] in a 5 m × 6 m × 4.5 m rectangular room, while varying the T60 ∈ {0.3, 0.6, 1.2} s. A variable number $Q \in \{4, 8, 16\}$ of *first*-order HOMs ($V = 1$) are positioned on a circle around a ROI of $R = 1$ m. For each setup, we consider 3 realizations of $S = 2$ sources located in random positions within the source region ($R_s = 0.5$ m) and with 3 s source signals (male and female speakers) taken from dev1 dataset of [37]. Furthermore, we considered a setup with $S = 3$ sources simulated with the same aforementioned conditions. The proposed method is implemented in MATLAB with 8 kHz sampling frequency. The STFT hamming window has a length of 512 samples, and we adopt an overlap of 75% and 1024-samples FFT obtained using zero padding of the frames. A dictionary (14) of $O = 27$ Green's functions placed on points of a uniformly sampled grid in the source region is employed (i.e., the distance between neighbour points is $17.7cm$), while the SH expansion order of the reconstruction (24) is limited to $\tilde{N}_E = 15$. In addition, in order to evaluate the impact of SH order in MNMF process (17) we evaluate different $\tilde{N}'_E \in \{1, 2, 3\}$. The number of NMF basis per source has been set to 20 for the proposed model and state-of-the-art techniques. We adopted the MATLAB implementation of ILRMA[1] with partitioning model which exhibits higher performance in [35]. We employed the Pytorch implementation of FASTMNMF[2] with the iterative projection algorithm. We considered both *circular* and *two-steps* initializations and we adopted the latter with 30 steps of gradual initialization [29] which showed improved performance as observed also in [29]. Finally, we considered a total number of 500 iterations for all the considered algorithms. Audio examples of the separation results are available online [38].

4.2. Discussion

4.2.1. Results with two sources

We evaluate the separation performance adopting SDR, SIR and SAR metrics [39] computed for every sensor and averaged. Note that the values combine the results for the 3 random source location setups. Inspecting the SDR results for female speech (upper row of Fig. 2), we can observe that the proposed method provides a rather robust performance independently from the number of HOMs. Conversely, FastMNMF shows its best results with 4 HOMs independently from T60; however, its performance rapidly degrades when increasing the number of HOMS (16 HOMs, i.e., 64 channels in total). This behaviour was reported by the authors and it is due to numerical approximation errors in the original implementation to optimize the computational requirements. The same trend can be observed in Table 1 where the average SIR and SAR results are reported. As far as ILRMA is concerned, it reports low average performance for all the considered setups.

From Fig. 2, one can note that the employed SH order during the MNMF procedure does not substantially influence the overall SDR performance of the separation. As a matter of fact, the proposed method constantly achieves values ≥ 10 dB for all the scenario, excluding SDR with 4 HOMs at T60 = 0.3 s and SH order 1 (≈ 7 dB). In general, improved results can be obtained with higher SH order, but this comes

[1] https://github.com/d-kitamura/ILRMA.
[2] https://github.com/sekiguchi92/SoundSourceSeparation.

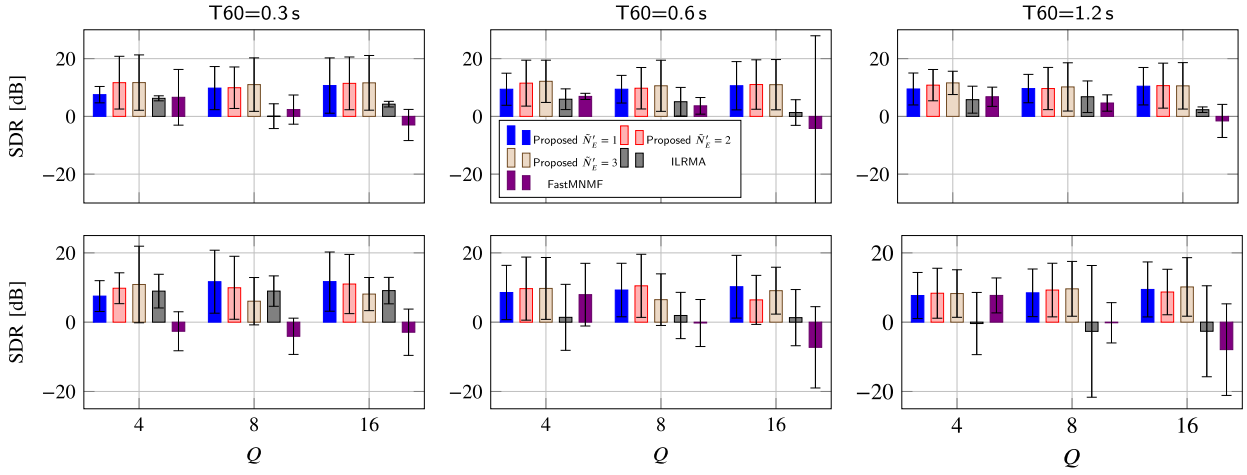

Fig. 2. First row: female 2 sources SDR results; second row male 2 sources SDR results.

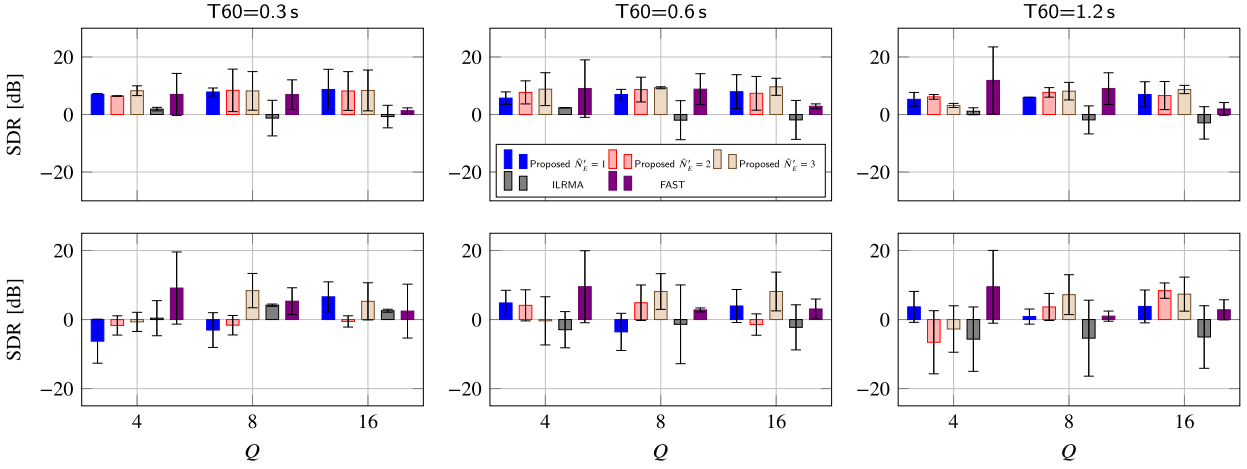

Fig. 3. First row: female 3 sources SDR results; second row male 3 sources SDR results.

at the cost of increased computational cost. As a matter of fact, both SIR and SAR in Table 1 improve at higher $\tilde{N}'_E$. Hence, one could limit the expansion order obtaining comparable performance with a reduced number of channels. From Table 1, we can observe that the scenario with $Q = 4$ HOMs particularly benefits from higher expansion orders, suggesting that, in case of limited HOMs, the order can be increased for improving the performance. Interestingly, the separation is not considerably affected by the reverberation time showing that the prior estimation of the exterior field effectively removes reverberant components of the sound field. Therefore, all the considered MNMF techniques take advantage of dereverberated input signals. Inspecting the results for male sources (bottom row in Fig. 2) we can note a decrease in the metrics for FastMNMF and ILRMA e.g., $\text{SDR}_{\text{FastMNMF}} < 2\,\text{dB}$ or negative. The cause of the performance decrease can be attributed to a rank-deficient SCM problem due to the fact that the considered source signals are more correlated between each other with respect to case of female sources. On the other hand, the proposed technique, achieves SDR values that are similar for the two sources. It follows that including the sound field propagation model into the MNMF allows us to increase the robustness of the separation independently from the characteristics of the source signals.

4.2.2. Results with three sources

We provide additional results considering the same setup described in Sec. 4.1 but using combinations of $S = 3$ male and female speech source signals active simultaneously.

From an overall inspection of Fig. 3 and Table 2, we can observe similar trends with respect to the two-sources scenario. Interestingly, we can see a clear improvement for FastMNMF in the case of 3 male sources and 4 HOMs with respect to the 2 sources scenario. In fact, adding a further source seems to mitigate the problem of dealing with rank-deficient SCM when the sources are highly correlated. Nonetheless, this method continues suffering from approximation errors when the number of HOMs increases. On the contrary, ILRMA provides the worst separation results among the compared methods both in Fig. 3 and Table 2. In fact, ILRMA is a determined approach (#channels = #sources) and thus, it doesn't fully exploit the spatial information provided by the HOMs. We remark that both, FASTMNMF and ILRMA operate over the estimated direct sound signal at each sensor location from the propagated exterior field coefficients obtained using (16) as in [17,20]. Alternatively, in the proposed approach we operate directly over the estimated exterior field coefficients $\hat{\beta}(f,t)$ which allow us to reconstruct the exterior field of the sources in any point outside the source region. In fact, as depicted for the two sources scenario, using the estimated exterior field coefficients clearly limits the negative effect of the reverberation on the separation performance for all the compared methods.

Regarding the proposed approach, we can see that better performance is observed with higher number of HOMs. The best average SIR and SAR in Table 2 are obtained by the proposed approach when $Q = 8, 16$. The separation performance of the proposed approach implicitly relies on the localization which is more accurate with 16 HOMs than with 4 and 8 HOMs. As a matter of fact, the weights w_{s0} (17) select the Green's function in the grid that are located close to the actual

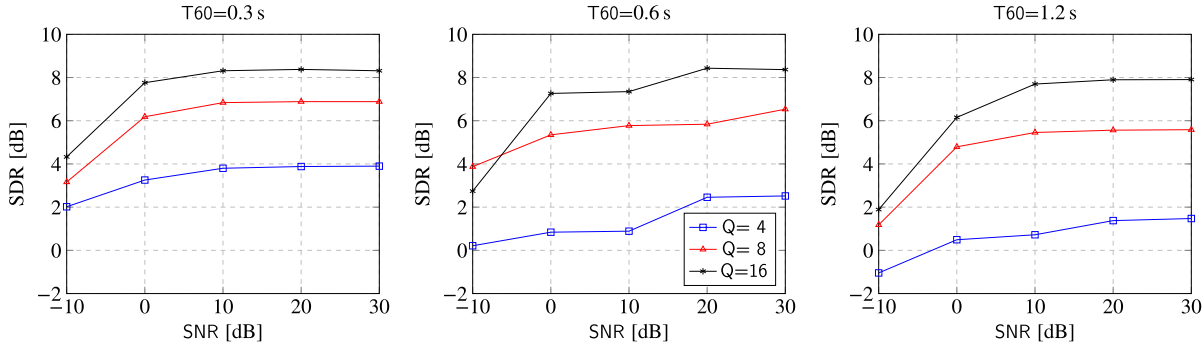

Fig. 4. Effect of background noise on the separation performance for 2 female sources using the Proposed method with $\tilde{N}'_E = 3$.

positions of the sources. On the other hand, the separation performance with respect to the SH order seems to be quite stable for the female scenario while, for the case of male sources, in general, higher order provides better results. Note that, in general, higher SH order allows shifting the spatial alias towards higher frequency values which may benefit the source localization, specially when the sources are highly correlated.

4.2.3. Robustness against background noise

As explained in [15], the sound field model used in this work offers a proficient theoretical solution for dereverberation and noise suppression when the noise is originated outside the region of interest.

To evaluate the effect of interference noise in the separation performance, we simulated an undesired source (e.g. an air conditioning machine on a meeting room) modeled using Gaussian noise randomly located outside the region of interest with different SNR levels with respect to the target mixture. Several tests were performed using the room setup in Sec. 4.1. In these experiments, the number of target sources is limited to two female speech sources, and the SH order in the MNMF process is set to $\tilde{N}'_E = 3$.

Fig. 4 displays the average separation results in terms of SDR with respect to the SNR considering different number of arrays and T60. As can be seen, the separation performance clearly degrades when SNR< 0 dB while for SNR values above 10 dB the proposed system performs similarly. Moreover, as expected, the effect of reverberation and the background noise is mitigated when more microphones are used.

5. Conclusions

We proposed a MNMF-based framework in the SH domain that operates directly on the exterior field coefficients. It does so by including the propagation model of the SH coefficients achieving the separation of the global direct acoustic field independently from the final reconstruction location. This framework can be used to increase the robustness of the audio signal processing techniques against reverberation. We evaluated the proposed framework for the task of source separation using a dataset with several source locations, type and reverberation conditions. The obtained results demonstrate the robustness of this strategy under high reverberation when combined with source separation approaches.

CRediT authorship contribution statement

Mirco Pezzoli: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Julio Carabias-Orti:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing. **Pedro Vera-Candeas:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing, Project administration. **Fabio Antonacci:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Augusto Sarti:** Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Audio examples are available online and are linked in the manuscript

References

[1] Bianchi L, Antonacci F, Sarti A, Tubaro S. Model-based acoustic rendering based on plane wave decomposition. Appl Acoust 2016;104:127–34. https://doi.org/10.1016/j.apacoust.2015.10.010.
[2] Wien M, Boyce JM, Stockhammer T, Peng W-H. Standardization status of immersive video coding. IEEE J Emerg Sel Top Circuits Syst 2019;9(1):5–17. https://doi.org/10.1109/JETCAS.2019.2898948.
[3] Rafaely B. Spatial sampling and beamforming for spherical microphone arrays. In: Hands-free speech communication and microphone arrays. IEEE; 2008. p. 5–8.
[4] Yan S, Sun H, Svensson UP, Ma X, Hovem JM. Optimal modal beamforming for spherical microphone arrays. IEEE Trans Acoust Speech Signal Process 2011;19(2):361–71. https://doi.org/10.1109/TASL.2010.2047815.
[5] Chu Z, Yang Y, Yang Y. A new insight and improvement on deconvolution beamforming in spherical harmonics domain. Appl Acoust 2021;177:107900. https://doi.org/10.1016/j.apacoust.2020.107900.
[6] Kumari D, Kumar L. Optimal beamformer design in spherical sector harmonics domain. Appl Acoust 2022;200:109070. https://doi.org/10.1016/j.apacoust.2022.109070.
[7] Epain N, Jin C. Independent component analysis using spherical microphone arrays. Acta Acust United Acust 2012;98:91–102.
[8] Nikunen J, Politis A. Multichannel nmf for source separation with ambisonic signals. In: Int Workshop Acoust Signal Enhanc. IEEE; 2018. p. 251–5.
[9] Muñoz-Montoro AJ, Carabias-Orti JJ, Vera-Candeas P. Ambisonics domain singing voice separation combining deep neural network and direction aware multichannel nmf. In: 2021 IEEE 23rd international workshop on multimedia signal processing (MMSP); 2021. p. 1–6.
[10] Guzik M, Kowalczyk K. Convolutive ntf for ambisonic source separation under reverberant conditions. In: ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2023. p. 1–5.
[11] Lluís Francesc, Meyer-Kahlen Nils, Chatziioannou Vasileios, Hofmann Alex. Direction specific ambisonics source separation with end-to-end deep learning. Acta Acust 2023;7:29. https://doi.org/10.1051/aacus/2023020.
[12] Mitsufuji Y, Takamune N, Koyama S, Saruwatari H. Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain. IEEE Trans Acoust Speech Signal Process 2021;29:607–17. https://doi.org/10.1109/TASLP.2020.3045528.
[13] Pezzoli M, Carabias-Orti JJ, Cobos M, Antonacci F, Sarti A. Ray-space-based multichannel nonnegative matrix factorization for audio source separation. IEEE Signal Process Lett 2021;28:369–73.
[14] Pezzoli M, Borra F, Antonacci F, Sarti A, Tubaro S. Reconstruction of the virtual microphone signal based on the distributed ray space transform. In: 26th Eur Signal Process Conf. IEEE; 2018. p. 1537–41.

[15] Fahim A, Samarasinghe PN, Abhayapala TD. Sound field separation in a mixed acoustic environment using a sparse array of higher order spherical microphones. In: Hands-free speech communication and microphone arrays. IEEE; 2017. p. 151–5.
[16] Pezzoli M, Borra F, Antonacci F, Tubaro S, Sarti A. A parametric approach to virtual miking for sources of arbitrary directivity. IEEE Trans Audio Speech Lang Process 2020;28:2333–48.
[17] Borra F, Gebru ID, Markovic D. Soundfield reconstruction in reverberant environments using higher-order microphones and impulse response measurements. In: Int Conf Acoust Speech Signal Process. IEEE; 2019. p. 281–5.
[18] Samarasinghe PN, Abhayapala TD, Poletti MA. 3d spatial soundfield recording over large regions. In: Int Workshop Acoust Signal Enhanc. IEEE; 2012. p. 1–4.
[19] Pezzoli M, Cobos M, Antonacci F, Sarti A. Sparsity-based sound field separation in the spherical harmonics domain. In: Int Conf Acoust Speech Signal Process. IEEE; 2022. p. 1051–5.
[20] Borra F, Krenn S, Gebru ID, Marković D. 1st-order microphone array system for large area sound field recording and reconstruction: discussion and preliminary results. In: Workshop Appl Signal Process Audio Acoust. IEEE; 2019. p. 378–82.
[21] Williams EG. Fourier acoustics: sound radiation and nearfield acoustical holography. Elsevier; 1999.
[22] Jones HM, Kennedy RA, Abhayapala TD. On dimensionality of multipath fields: spatial extent and richness. In: Int Conf Acoust Speech Signal Process, vol. 3. IEEE; 2002. p. III-2837–40.
[23] Duong NQK, Vincent E, Gribonval R. Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans Acoust Speech Signal Process 2010;18(7):1830–40. https://doi.org/10.1109/TASL.2010.2050716.
[24] Ozerov A, Fevotte C. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans Acoust Speech Signal Process 2010;18(3):550–63. https://doi.org/10.1109/TASL.2009.2031510.
[25] Sawada H, Kameoka H, Araki S, Ueda N. New formulations and efficient algorithms for multichannel nmf. In: Workshop Appl Signal Process Audio Acoust. IEEE; 2011. p. 153–6.
[26] Sawada H, Kameoka H, Araki S, Ueda N. Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans Acoust Speech Signal Process 2013;21(5):971–82.
[27] Boyd SP, Lieven V. Introduction to applied linear algebra: vectors, matrices, and least squares. Cambridge, UK: Cambridge University Press; 2018.
[28] Sekiguchi K, Nugraha AA, Bando Y, Yoshii K. Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices. In: 27th Eur Signal Process Conf. IEEE; 2019. p. 1–5.
[29] Sekiguchi K, Bando Y, Nugraha AA, Yoshii K, Kawahara T. Fast multichannel non-negative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation. IEEE Trans Acoust Speech Signal Process 2020;28:2610–25.
[30] Mitsufuji Y, Uhlich S, Takamune N, Kitamura D, Koyama S, Saruwatari H. Multichannel non-negative matrix factorization using banded spatial covariance matrices in wavenumber domain. IEEE Trans Acoust Speech Signal Process 2020;28:49–60. https://doi.org/10.1109/TASLP.2019.2948770. https://ieeexplore.ieee.org/document/8878116/.
[31] Zhang F-M, Zhang X-Z, Zhang Y-B, Bi C-X, Zhou R. Sound field reconstruction using sparse bayesian learning equivalent source method with hyperparametric-coupled prior. Appl Acoust 2023;211:109496. https://doi.org/10.1016/j.apacoust.2023.109496.
[32] Tsunokuni I, Kurokawa K, Matsuhashi H, Ikeda Y, Osaka N. Spatial extrapolation of early room impulse responses in local area using sparse equivalent sources and image source method. Appl Acoust 2021;179:108027. https://doi.org/10.1016/j.apacoust.2021.108027.
[33] Ben Hagai I, Pollow M, Vorländer M, Rafaely B. Acoustic centering of sources measured by surrounding spherical microphone arrays. J Acoust Soc Am 2011;130(4):2003–15. https://doi.org/10.1121/1.3624825.
[34] Nikunen J, Virtanen T. Direction of arrival based spatial covariance model for blind sound source separation. IEEE Trans Acoust Speech Signal Process 2014;22(3):727–39.
[35] Kitamura D, Ono N, Sawada H, Kameoka H, Saruwatari H. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. IEEE Trans Acoust Speech Signal Process 2016;24(9):1626–41.
[36] Habets EAP. Room impulse response generator. Technische Universiteit Eindhoven, Tech Rep 2 (2.4). 2006. p. 1.
[37] Ono N, Koldovskỳ Z, Miyabe S, Ito N. The 2013 signal separation evaluation campaign. In: Int Workshop Mach. Learn Signal Process. IEEE; 2013. p. 1–6.
[38] Pezzoli M, Carabias-Orti JJ, Vera-Candeas P, Antonacci F, Sarti A. Spherical-harmonics-based sound field decomposition and multichannel nmf for sound source separation. https://polimi-ispl.github.io/spherical-harmonics-mnmf. [Accessed 7 February 2023].
[39] Vincent E, Gribonval R, Févotte C. Performance measurement in blind audio source separation. IEEE Trans Acoust Speech Signal Process 2006;14(4):1462–669.