Ray-Space constrained multichannel Nonnegative Matrix Factorization for Audio Source Separation

Antonio J. Muñoz-Montoro[1], Marco Olivieri[2], Mirco Pezzoli[2],
Julio Carabias-Orti[1], Fabio Antonacci[2], and Augusto Sarti[2]
[1]*Departamento de Ingeniería de Telecomunicación*, *Universidad de Jaén*, Linares, Spain
[2]*Dipartimento di Elettronica, Informazione e Bioingegneria*, *Politecnico di Milano*, Milan, Italy

***Abstract*—Non-negative matrix factorization (NMF) has been widely adopted for the blind separation of acoustic sources. In the context of VR and AR applications, when several microphones are available the adoption of sound field representations such as spherical harmonics or ray space is shown to be effective in the NMF context. In this work, we propose a dictionary-based NMF-based model considering ray-space-transformed signals. The novelty of this approach is to account for the explicit modelling of the frequency dependency of the sound propagation from the source positions to the sensors. Spatially-constrained approaches aim at exploiting spatial information to improve the separation performance; however, they may not take advantage of possible priors given by representations of the data. The proposed approach allows us to exploit the source location model of the Ray Space through a predefined frequency-dependent dictionary of Ray Space patterns. Results demonstrate the competitive performance of the proposed method with respect to state-of-the-art NMF-based algorithms using real recordings.**

***Index Terms*—Ray Space, Non-negative matrix factorization (NMF), blind source separation (BSS), microphone array processing**

I. Introduction

Sound source separation (SSS), a fundamental task in the field of audio signal processing, aims to disentangle multiple audio sources from a mixtures thereof with applications like speech enhancement or instrument isolation in music recordings. More recently, augmented reality applications and object-based (6DOF) audio formats [1], [2] further increased interest in multichannel audio processing, including SSS.

Traditionally, Nonnegative Matrix Factorization (NMF) has been a popular choice for both blind and informed SSS, where it factorizes an input matrix, typically the magnitude spectrogram, into rank-1 components. Initially designed for single-channel scenarios, extending NMF to multichannel setups involved aggregating channels into matrices [3] or considering parallel factor models [4]. In order to exploit the spatial information in the multichannel acquisition, the so-called multichannel NMF (MNMF) used a Gaussian probabilistic model [5]–[7] where the observation and the mixing system are encoded using spatial covariance matrices (SCMs), modeling both, magnitude correlations and phase differences, between channels. Although effective, the technique suffers from high computational costs and sensitivity to parameter initialization. To reduce the computational cost caused by full-rank SCMs, rank-1 restriction based on independent vector analysis has been combined with NMF in the so-called independent low-rank matrix analysis (ILRMA) [8]. Alternatively, to keep the model full-rank, several techniques based on diagonalization have been proposed to mitigate the initialization dependency and provide efficient computational solutions [9]. However, these methods relied on statistical independence between the sources to derive spatial characteristics. Recently, unsupervised deep learning solutions trained directly over the mixtures have been developed for full-rank spatial covariance analysis (FCA) using a generative model consisting of the full-rank spatial model and the deep spectral model [10].

Several NMF-based approaches [11]–[14] propose to project the array signals onto sound field domains to exploit the properties of transformed signals. In [12], the discrete Fourier transform is adopted to map the multichannel acquisitions on the wave-number domain (WN-MNMF). The spherical harmonics (SH) domain was first employed in [15] using a plane wave modeling of the source, while [16] extended the approach to deal with the near-field source. In fact, to deal with the high number of SH signals, this method relies only on the diagonal information in the SCMs, and therefore, the MNMF model converges to NTF. More recently, in [14] a global description of the sound field in the spherical harmonics is introduced in order to exploit multiple microphone arrays in a single representation for MNMF. Using an NTF approach, Lee et al. [11] proposed a beamspace data model suitable for far-field scenarios that consider the projection of the input signal onto a set of steered directions while accounting for the inherent phase-difference information that is present in this type of recordings. Similarly, in [13] the Ray Space Transform (RST) [17], [18] has been adopted to project the signals of a uniform linear array (ULA) onto the ray space. This representation allows the authors to take advantage of the inherent modeling of the source location in ray space [19] to overcome the far-field DOA assumption typically adopted in the literature.

In this work, we propose to extend the RS-NTF model in

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", by the Universidad de Málaga under project "Ayudas para la movilidad y perfeccionamiento del personal investigador y difusión de la actividad investigadora" and under project "Ayudas para proyectos dirigidos por jóvenes investigadores" with grant reference "B1-2023_035", and by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

[13] by adopting a set of pre-computed source locations in the ray space. The adoption of a dictionary of ray space patterns allows us to improve the separation performance letting the optimization select the best matching sources. The results on real measurements show that the proposed constrained RS-NTF is able to provide improved separation with respect to state-of-the-art MNMF-based techniques.

The rest of the paper is organized as follows. Sec. II formulates the data model and revised the NTF models using spatial information in the literature. The proposed constrained NTF model in the Ray Space domain is presented in Sec. III. The obtained results and comparisons with other state-of-the-art source separation methods are presented in Sec. IV. Finally, Sec. V concludes the paper.

II. Background

A. Data Model

Consider a ULA comprising I microphones. In the presence of J acoustic sources, we can express the time-frequency representation of the ith microphone signal as [5]

$$y_i(\omega,n)=\sum_{j=1}^{J}a_{i,j}(\omega)s_j(\omega,n)+b_i(\omega,n), \tag{1}$$

where, $i=1,\ldots,I$ denotes the microphone index, n represents the time frame index, $\omega=2\pi f$ is the angular frequency with $f>0$ the temporal frequency. Additionally, $a_{i,j}(\omega)$ models the transfer function between the ith microphone and the jth source, $s_j(\omega,n)$ denotes the jth source signal, and $b_i(\omega,n)$ characterizes the self-noise of the ith microphone.

The RST, introduced in [17], [18], serves as a linear operator that maps microphone signals (1) onto the Ray Space [19]. The Ray Space is essentially a domain based on the parameterization of the line equation $z=\mu x+\nu$ within the xz plane in terms of μ and ν, namely the slope and the intercept on the z-axis, respectively. Consequently, a ray within the geometric space is transformed into a point within the Ray Space. Notably, a key characteristic of ray space is its ability to map acoustic rays emanating from point-like sources onto lines within the Ray Space, as demonstrated in [17]. This feature enables the development of algorithms for source localization [19] and separation [13], [20] as operations on linear patterns within Ray Space.

The narrowband RST is defined through the linear operator [17] denoted as $\mathbf{\Psi}(\omega)$, with its (i,t)th element given as:

$$[\mathbf{\Psi}]_{i,t}(\omega)=e^{-\jmath\omega\frac{d\mu_{\mathrm{w}}}{c\sqrt{1+\mu_{\mathrm{w}}^2}}(i-1)}\psi^*_{l,i},\ t=(l-1)+(\mathrm{w}-1)L+1 \tag{2}$$

where $\mathbf{\Psi}(\omega)\in\mathbb{C}^{I\times LD}$, and $t=1,\ldots,LD$ represents the index of the sampled point within the Ray Space. Here, $\mathrm{w}=1,\ldots,D$ is the index for the D sampled directions that discretize the μ parameter space, and $l=1,\ldots,L$ denotes the index indicating spatial displacement across different subarrays, related to the discretization of the ν parameter space. The term $\psi_{l,i}$ in (2) represents a Gaussian spatial window defining the lth subarray [17]. It follows that ray-space-transformed signals of the ULA (1) are given as

$$\mathbf{z}(\omega,n)=\mathbf{\Psi}^H(\omega)\mathbf{y}(\omega,n), \tag{3}$$

where $\mathbf{z}(\omega,n)=[z_1(\omega,n),\ldots,z_{LD}(\omega,n)]^T$ is the vector of the Ray Space data. The array signals can be recovered as

$$\mathbf{y}(\omega,n)\approx\mathbf{\Psi}^\dagger\mathbf{z}(\omega,n), \tag{4}$$

where $\mathbf{\Psi}^\dagger=\left(\mathbf{\Psi}\mathbf{\Psi}^H\right)^{-1}\mathbf{\Psi}$ is the pseudo-inverse RST matrix [17].

B. Non Negative Tensor Factorization (NTF)

When multichannel audio signals are transformed into a set of spectrograms (i.e. one for each channel), they can be treated as a three-way tensor denoted as $\mathbf{Y}$. NTF aims to estimate an approximation using $\hat{\mathbf{Y}}$, assuming an instantaneous mixture, which is constructed as a superposition of k feature matrices $q_{j,i}$, $w_k(\omega)$ and $h_k(n)$ modeling the tensor channel, frequency, and time components, respectively:

$$\hat{y}_i(\omega,n)=\sum_j q_{j,i}w_{k,j}(\omega)h_{k,j}(n), \tag{5}$$

The free parameters can be estimated by solving the following optimization problem:

$$\min_{Q,W,H}\sum_{iwn}g_i(\omega,n)d_\beta\left(y_i(\omega,n)|\hat{y}_i(\omega,n)\right), \tag{6}$$

where $g_i(\omega,n)$ is a weighting parameter that controls the impact of the error observed in the different elements of $y_i(\omega,n)$. d_β represents the β-divergence, allowing the separation quality to be changed, subject to the parameter β [21]. When β equals 2, 1, or 0, the NTFs are called EUC-NTF, KL-NTF, or IS-NTF, respectively.

Nevertheless, given that these methods exclusively deal with non-negative power spectrograms, they depend solely on amplitude information, completely disregarding the phase details of the short-time Fourier transforms (STFTs). However, phase information can hold significant relevance in multichannel source separation, especially when dealing with far-field scenarios where the information carried by the interchannel level differences (ILD) becomes almost non-discriminating.

Several approaches have been proposed to account for the phase differences between channels. One approach is the so-called Multichannel NMF (MNMF) that employs a semi-nonnegative Gaussian probabilistic modeling applied directly to the complex-valued STFTs of all channels [5], [6], [15] using a full-rank spatial covariance matrix (SCM) representation. Moreover, several techniques based on diagonalization have been proposed to provide efficient computational solutions [9], [12], [16]. Another approach consist in integrating the spatial information into power spectrogram observation using beamforming [11] or ray-space domain [13] transforms.

C. The Beamspace-Domain NTF (BS-NTF)

The beamspace domain, introduced in [11], describes the microphone signals by focusing on the directional attributes of the sound field, which represent the interchannel phase differences (IPDs) within the data. In practical terms, this approach commences with a plane-wave decomposition [22] of the signals across a specified set of M directions, a method known as the beamspace transform [11].

$$\tilde{\mathbf{y}}(\omega, n) = \mathbf{G}_{\mathrm{BT}}^{H} \mathbf{y}(\omega, n), \tag{7}$$

where $\mathbf{y}(\omega, n) = [y_1(\omega, n), \ldots, y_I(\omega, n)]^T$ is the vector of the array signals, $\tilde{\mathbf{y}}(\omega, n) = [\tilde{y}_1(\omega, n), \ldots, \tilde{y}_M(\omega, n)]^T$ is the vector of the beamspace signals and $\mathbf{G}_{\mathrm{BT}} \in \mathbb{C}^{I \times M}$ is the beamspace transform matrix [23], composed by the steering vectors whose elements are [24]

$$p_i(\theta_m, \omega) = e^{-\jmath\omega \frac{d \sin(\theta_m)}{c}(i-1)}, \quad i = 1, \ldots, I, \tag{8}$$

with d the distance between two consecutive microphones, c the speed of sound and $\jmath$ the imaginary unit. Then, inspired by [5], the square magnitude for each of the mth beamspace signal $\hat{y}_m(\omega, n)$, assuming an instantaneous mixture, can be modelled as

$$\hat{y}_m(\omega, n) = \sum_{j=1}^{J} g_{m,j} \sum_{k} w_{k,j}(\omega) h_{k,j}(n), \tag{9}$$

where $g_{m,j} \in \mathbb{R}_+$ represents the mixing weights of the jth source in the beamspace domain and the mth beamspace bin, $w_{k,j}(\omega) \in \mathbb{R}_+$ and $h_{k,j}(n) \in \mathbb{R}_+$ denote both the basis functions and their corresponding time-varying gains for each kth source-dependent component. Note that $g_{m,j}$ in (9) is frequency independent since all the frequency components pertaining to the same signals are assumed to have shared DOA. In general, $g_{m,j}$ reaches its maximum when θ_m (8) equals the DOA of the jth source [11].

Although BS-NTF exploits the IPDs in the optimization, its source location information is limited to the DOA only due to the far-field model of (8). This limitation can be overcome thanks to the adoption of the Ray Space signal model [17], [20].

D. Ray Space NTF model

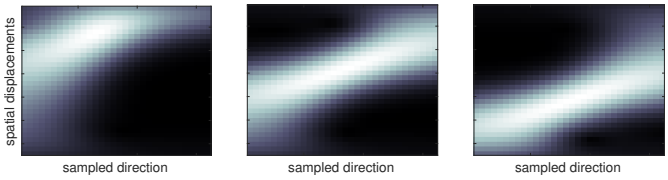
As described in [13], in the presence of J sources active at any time, the data in the Ray Space can be expressed as

$$z_t(\omega, n) = \sum_{j=1}^{J} r_{t,j} s_j(\omega, n) + b_t(\omega, n), \tag{10}$$

where $r_{t,j}$ describes the contribution of the jth source to the tth Ray Space element. In [13], the authors proposed an NTF model to estimated square magnitude of the Ray Space data as follows,

$$\hat{z}_t(\omega, n) = \sum_{j} g_{t,j} \sum_{k} w_{k,j}(\omega) h_{k,j}(n), \tag{11}$$

with $g_{t,j} = |r_{t,j}|^2$. Note that (10) assumes an instantaneous mixture model, which does not hold in practice since the propagation of the Green functions is not invariant for all frequencies. In this work we mitigate this limitation by accounting for the frequency dependency of the propagation of the Green functions from the source location to the sensors.



(a) (b) (c)

Fig. 1: Example of three different elements within the Ray Space dictionary for $\omega = 1$kHz, each associated with a different grid position. (a) corresponds to the grid position $(0.3, 0.1)$ m, (b) to $(0.2, 0.4)$ m, and (c) to $(0.2, 0.6)$ m.

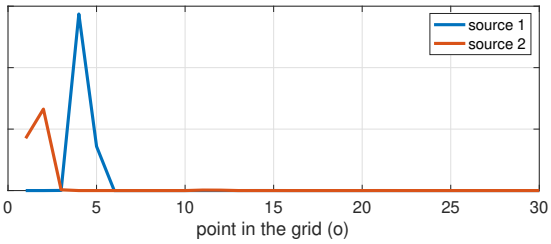


Fig. 2: Example of the estimation of parameter $c_{j,o}$ for two sources. Note that each source is associated with specific positions within the position grid.

III. Dictionary-based Ray Space NTF

In this work, we propose a novel signal model that integrates a Ray Space dictionary, linked to positions within a predefined grid of locations. Based on (11) and using a tensor notation, we propose to approximate the observed data in the Ray Space domain as follows:

$$\hat{\mathbf{Z}}_t = \sum_{j,o} \mathbf{g}_{t,o} c_{j,o} \mathbf{W}_j \mathbf{H}_j. \tag{12}$$

where $c_{j,o}$ establishes the relationship between the spatial position denoted by o within the grid and the corresponding sources identified by j and $[\mathbf{g}_{t,o}]_\omega = g_{t,o}(\omega)$ is the frequency vector of the dictionary associated to the position o and Ray Space element t.

To construct this dictionary, we have modeled the propagation of the position grid to the sensors by using Green functions, followed by the application of the Ray Space transform. An inherent advantage of this dictionary lies in its ability to effectively capture the frequency-dependent characteristics of each Ray Space element, which represents a significant and noteworthy contribution in contrast to [13]. Figure 1 shows three different elements from the Ray Space dictionary. Each of these illustrative instances corresponds to a different position within the grid, resulting in the emergence of various ray types. Figure 2 shows an example of the estimation of parameter $c_{j,o}$ for two sources. As can be observed, this parameter associates each source to different positions in the grid.

Furthermore, for an accurate representation of the parameter $c_{j,o}$, we propose the incorporation of regularization during the parameter estimation process. This regularization is pivotal in

preventing the concurrent activation of multiple sources at the same grid position. This is a significant contribution since it enhances the fidelity of source activation modeling within the positional grid, which is a critical aspect. Thus, the cost function can be formulated as follows:

$$C_{\mathrm{RS}}(\Theta) = \sum_{t,\omega,n} d_\beta \left(|\mathbf{Z}_t|^2 \middle| \hat{\mathbf{Z}}_t \right) + \alpha(c_{j,o}), \tag{13}$$

where $\alpha(c_{j,o}) = \lambda \sum_j \left(c_{j,o} c_{j,o}^T - \mathrm{tr}(c_{j,o} c_{j,o}^T) \right)$ is the penalty function of the single activation regularization and $\mathrm{tr}(\cdot)$ is the trace operator of a square matrix. As proposed in [13], we can exploit the similarity with the instantaneous algorithm of [5] to derive the updated algorithm of the multiplicative updates (MU) method:

$$c_{j,o} \leftarrow c_{j,o} \frac{\sum_{t\omega} \mathbf{g}_{t,o} \sum_n \left(\hat{\mathbf{Z}}_t^{\cdot \beta-2} \cdot |\mathbf{Z}_t|^2 \cdot (\mathbf{W}_j \mathbf{H}_j) \right) + 2\lambda c_{j,o}}{\sum_{t\omega} \mathbf{g}_{t,o} \sum_n \left(\hat{\mathbf{Z}}_t^{\cdot \beta-1} \cdot (\mathbf{W}_j \mathbf{H}_j) \right) + 2\lambda \sum_j \mathbf{1}^J c_{j,o}}, \tag{14}$$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \cdot \frac{\sum_{t,o} \mathbf{g}_{t,o} c_{j,o} \left(\hat{\mathbf{Z}}_t^{\cdot \beta-2} \cdot |\mathbf{Z}_t|^2 \right) \mathbf{H}_j^{\mathrm{T}}}{\sum_{t,o} \mathbf{g}_{t,o} c_{j,o} \hat{\mathbf{Z}}_t^{\cdot \beta-1} \mathbf{H}_j^{\mathrm{T}}}, \tag{15}$$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j \cdot \frac{\sum_{t,o} \left(\mathbf{g}_{t,o} c_{j,o} \mathbf{W}_j \right)^{\mathrm{T}} \left(\hat{\mathbf{Z}}_t^{\cdot \beta-2} \cdot |\mathbf{Z}_t|^2 \right)}{\sum_{t,o} \left(\mathbf{g}_{t,o} c_{j,o} \mathbf{W}_j \right)^{\mathrm{T}} \hat{\mathbf{Z}}_t^{\cdot \beta-1}}, \tag{16}$$

where $\mathbf{1}^J$ is a $J \times J$ matrix of all ones and $\cdot$ represents the element-wise matrix operations. Once the factorization concludes, we can obtain an estimate of the Ray Space source image in terms of the minimum mean squared error (MMSE) as in [5], [11]

$$\tilde{\mathbf{S}}_t^{(j)} = \frac{\sum_o \mathbf{g}_{t,o} c_{j,o} [\mathbf{W}]_j [\mathbf{H}]_j}{\hat{\mathbf{Z}}_t} \mathbf{Z}_\mathbf{t}, \tag{17}$$

where $[\tilde{\mathbf{S}}_t^{(j)}]_{\omega,n} = [\tilde{s}_1^{(j)}, \ldots, \tilde{s}_{LD}^{(j)}]^T$ is the estimated contribution of the jth source at the tth Ray Space bin. Finally, an estimate of the sources at each microphone can be obtained by applying the inverse RST (4)

$$\hat{\mathbf{S}}_i^{(j)} = \mathbf{\Psi}_{it}^\dagger \tilde{\mathbf{S}}_t^{(j)}, \tag{18}$$

where $[\hat{\mathbf{S}}_i^{(j)}]_{\omega,n} = [\hat{s}_1^{(j)}, \ldots, \hat{s}_I^{(j)}]^T$ is the vector of the jth estimated source signal at the ith microphone.

IV. Experiments

To evaluate the proposed algorithm, we conducted experiments in a reverberant environment using a 32-microphone ULA with an inter-mic spacing of $0.03\,\mathrm{m}$. We enriched our dataset by estimating RIRs between microphones and multiple source positions using sweep excitation techniques [25], while maintaining an average T60 reverberation time of approximately 0.4 s. These RIR measurements[1] were acquired in a $5.5\,\mathrm{m} \times 3.4\,\mathrm{m} \times 3.3\,\mathrm{m}$ office room, using a Genelec 8020C[2] loudspeaker for sound source emissions. The source positions were organized in a grid configuration, spanning distances from 0.3 to 0.9 meters relative to the array's reference line. In the formulation of the dictionary, we considered a square grid of spatial coordinates, extending from 0 to 1 meter with respect to the array's reference line. For comparative analysis, we included FastMNMF [9], ILRMA [8], WN-MNMF [12] and DOA-MNMF [7]. All of these methods are based on a local Gaussian model that reduces IS divergence ($\beta = 0$). RS-NMF [13] is also considered in the comparison with $\beta = 0.9$, since it is an optimal value for this approach as shown in [13]. Furthermore, to illustrate the impact of incorporating regularization into the proposed model, we have included results for scenarios both with and without regularization, denoted as (Prop) and (Prop-Reg), respectively. It is essential to emphasize that in both cases, the dictionary remains fixed throughout the estimation process. Moreover, through empirical experimentation, we determined that using $\beta = 1.4$ for our model resulted in improved overall performance compared to other divergence measures.

We generated the array signals by convolving the acquired RIR with 3-second segments of source signals, including male and female speakers, as well as non-percussive music signals, extracted from the dev1 dataset [26]. For our experimentation, we considered combinations of 2 and 3 active sources. The signals were processed at an 8 kHz sampling rate, employing a STFT with a Hamming window of size 256 and 75% overlap, along with 512 FFT points. The number of iterations and the number of bases for each source were empirically set to 300 and 12, respectively, and these values were uniformly applied across all the algorithms considered. The number of Ray Space points (2) is set to be equal to the number of microphones $LD = 32$, with $L = 8$ subarrays and $D = 4$ directions uniformly sampled.

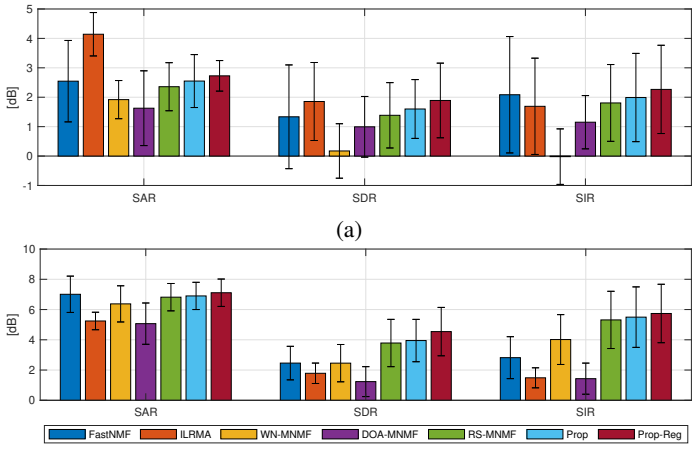
To evaluate the performance of the proposed algorithm, we computed the metrics SAR, SDR, and SIR for each microphone signal, followed by calculating the average values of these metrics in all microphone signals.

A. Results

Figure 3 presents the averaged results and standard deviations of the metrics computed for separation of two and three simultaneous active sources using the algorithms reported. In Figure 3a, we depict the averaged results obtained for male and female speech source signals. Notably, our proposed approach, when applied with regularization (Prop-Reg), demonstrates superior performance in terms of SDR and SIR, showcasing an improvement of approximately 0.4 dB over the non-regularized proposal. Overall, it is evident that Ray-Space-based models yield promising outcomes, with ILRMA being the sole technique that achieves competitive results in terms of SAR. Figure 3b shows the outcomes for music source signals. In this context, our proposed model consistently outperforms other algorithms across all metrics. It is worth highlighting the benefits of applying the regularization in achieving improved source separation. Notably, Ray-Space-based models, which employ a fixed transformation irrespective of the source type, exhibits greater consistency in results compared to other tech-

[1] Available: https://github.com/polimi-ispl/rs-mnmf

[2] Available: https://www.genelec.com/previous-models/8020c



(a)

(b)

Fig. 3: The average and standard deviation of SAR, SDR, and SIR for each algorithm when applied to (a) speech and (b) music signals.

niques, thanks to its inherent representation of source positions within the Ray Space.

V. Conclusion

In this work we have proposed a novel approach for audio source separation in the Ray Space domain, with a primary emphasis on integrating a predefined dictionary of Ray Space patterns. With the inclusion of this dictionary, we aim to model the frequency dependency of the propagation of the source position to the sensors. Moreover, we have introduced regularization techniques to enhance the spatial modeling of source activations within the position grid, ensuring accurate source localization. The empirical findings presented in this paper show evidence of the competitive prowess of our proposed method. In direct comparison with state-of-the-art NMF-based algorithms, our approach consistently outperforms them in real-world scenarios, as reflected by SDR, SIR, and SAR results. In the future, we plan to account for the room impulse responses to enhance the robustness of the model in the presence of reverberation.

References

[1] J. G. Tylka and E. Y. Choueiri, "Fundamentals of a parametric method for virtual navigation within an array of ambisonics microphones," *Journal of the Audio Engineering Society*, vol. 68, no. 3, pp. 120–137, 2020.

[2] M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, and A. Sarti, "A parametric approach to virtual miking for sources of arbitrary directivity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2333–2348, 2020.

[3] R. M. Parry and I. Essa, "Estimating the spatial position of spectral components in audio," in *International Conference on Independent Component Analysis and Signal Separation*, pp. 666–673, Springer, 2006.

[4] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *IEE conference publication*, vol. 511, p. 8, Citeseer, 2005.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

[6] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[7] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[8] 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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[9] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2610–2625, 2020.

[10] Y. Bando, K. Sekiguchi, Y. Masuyama, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Neural full-rank spatial covariance analysis for blind source separation," *IEEE Signal Processing Letters*, vol. 28, pp. 1670–1674, 2021.

[11] S. Lee, S. H. Park, and K. Sung, "Beamspace-domain multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Processing Letters*, vol. 19, no. 1, pp. 43–46, 2011.

[12] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 49–60, 2020.

[13] M. Pezzoli, J. J. Carabias-Orti, M. Cobos, F. Antonacci, and A. Sarti, "Ray-space-based multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 369–373, 2021.

[14] M. Pezzoli, J. Carabias-Orti, P. Vera-Candeas, F. Antonacci, and A. Sarti, "Spherical-harmonics-based sound field decomposition and multichannel nmf for sound source separation," *Applied Acoustics*, vol. 218, p. 109888, 2024.

[15] J. Nikunen and A. Politis, "Multichannel nmf for source separation with ambisonic signals," in *Int. Workshop Acoust. Signal Enhanc.*, pp. 251–255, IEEE, 2018.

[16] Y. Mitsufuji, N. Takamune, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on evanescent-region-aware non-negative tensor factorization in spherical harmonic domain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, pp. 607–617, 2021.

[17] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro, "The ray space transform: A new framework for wave field processing," *IEEE Transactions on Signal Processing*, vol. 64, pp. 5696–5706, Nov. 2016.

[18] F. Borra, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, S. Tubaro, and A. Sarti, "A fast ray space transform for wave field processing using acoustic arrays," in *28th European Signal Processing Conference (EUSIPCO)*, pp. 186–190, IEEE, 2021.

[19] D. Markovic, F. Antonacci, A. Sarti, and S. Tubaro, "Soundfield imaging in the ray space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2493–2505, 2013.

[20] D. Marković, F. Antonacci, L. Bianchi, S. Tubaro, and A. Sarti, "Extraction of acoustic sources through the processing of sound field maps in the ray space," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 2481–2494, Dec 2016.

[21] 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, 2009.

[22] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*. Elsevier, 1999.

[23] M. D. Zoltowski, "Beamspace root-music," *IEEE Trans. Signal Processing*, vol. 41, no. 1, pp. 344–364, 1993.

[24] W. Liu and S. Weiss, *Wideband beamforming: concepts and techniques*, vol. 17. John Wiley & Sons, 2010.

[25] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, 2001.

[26] N. Ono, Z. Koldovskỳ, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," in *2013 IEEE International workshop on machine learning for signal processing (MLSP)*, pp. 1–6, IEEE, 2013.