Multichannel Blind Source Separation Based on Evanescent-Region-Aware Non-Negative Tensor Factorization in Spherical Harmonic Domain

Yuki Mitsufuji, *Member, IEEE*, Norihiro Takamune, Shoichi Koyama, *Member, IEEE*,
and Hiroshi Saruwatari, *Member, IEEE*

***Abstract*—There is growing interest in new audio formats in the context of virtual reality (VR), and higher-order ambisonics (HOA) is preferred for VR systems to transmit recorded scenes owing to its transmission efficiency and its flexibility to work with different loudspeaker setups. However, the conversion between another well-known format, i.e., object format, and the HOA format is not fully addressed in the literature. To address this issue, blind source separation in a spherical harmonic (SH) domain can be considered as the best way to extract objects in terms of efficiency, i.e., decoding HOA signals for separation can be omitted. A few authors attempted to extract objects from encoded HOA signals directly by using multichannel non-negative matrix factorization (MNMF), but these approaches either assume only far-field sources or do not take array characteristics into account, which make these methods difficult to use for VR in practical situations where singers or speakers often perform close to microphones. Furthermore, MNMF generally requires a huge computational cost, although dimensional reduction to the SH domain is performed. In this work, we also model near-field sources by estimating the model parameters of non-negative tensor factorization (NTF) in the SH domain assuming that microphone signals can be obtained with a rigid spherical array. We propose a masking scheme to exclude noisy evanescent regions in the SH domain from the NTF cost function. Evaluations show that our method outperforms existing methods devised for the HOA format and that our masking approach is effective in improving the separation quality.**

***Index Terms*—Non-negative matrix factorization, spherical harmonics, spherical microphone array.**

I. Introduction

THE ambisonics format is widely used for virtual reality (VR) applications these days, owing to its transmission efficiency and its flexibility to work with different loudspeaker setups. The higher-order extension of the ambisonics format, higher-order ambisonics (HOA), has become a part of the Moving Picture Experts Group (MPEG)-H 3D Audio format [1]. In the HOA recording workflow, microphones are assumed to be placed on a sphere or on a circle, and an HOA encoder that converts microphone signals to the spherical harmonic (SH) domain is subsequently applied. The main advantage of the HOA format is that the reproduction speaker layout does not have to be known at the recording or encoder side, meaning that its reproduction is not limited to a specific loudspeaker setup.

The object-based audio format is also widely used and regarded as another standard VR format. It consists of pairs of object metadata and corresponding waveform signals. Placing an audio object anywhere in space gives listeners an immersive experience that can serve as the audio part of the six degrees of freedom (6DOF) system proposed in MPEG-I [2]. Although the HOA format can be rotated with respect to the origin in the spherical coordinate system, and many authors have made considerable effort to reduce the computational cost to make the system feasible in the three degree of freedom (3DOF) system [3]–[5], the format cannot be easily adapted to the 6DOF framework where translations along the xyz axes are required. Thus, a conversion from the HOA format to the object format is highly desired.

To address this issue, several works based on non-blind source separation methods such as beamformers have been reported. Exploiting the sum of weighted SHs for a beamformer design is referred to as a modal beamformer, and a lot of research has been carried out on this beamformer [6]–[13]. It is well-known that this type of beamformer is prone to suffering from low-frequency noise when the radius of the array is small; the trade-off between the modal beamformer and delay-and-sum beamformer was discussed in [8]. Furthermore, since a modal beamformer requires knowledge of the directions of arrival (DOAs) for target sources, the separation quality depends on the accuracy of the source localization method deployed as pre-processing.

As a blind source separation method, independent component analysis (ICA) in the SH domain was investigated in [14]. One limitation of an ICA-based method is that it cannot handle the underdetermined case where the number of sources is higher than the number of microphones.

Several authors attempted integrating multichannel non-negative matrix factorization (MNMF) in the SH domain to tackle the underdetermined problem [15], [16]. The method in [15] provides a fixed spatial covariance matrix (SCM) as

Manuscript received August 17, 2020; revised November 27, 2020; accepted December 7, 2020. Date of publication December 21, 2020; date of current version January 11, 2021. This work was supported in part by JSPS KAKENHI under Grants 19H01116 and 19H04131. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding author: Yuki Mitsufuji.)*

The authors are with the University of Tokyo, Tokyo 113-8654, Japan (e-mail: yuhki.mitsufuji@sony.com; norihiro_takamune@ipc.i.u-tokyo.ac.jp; koyama.shoichi@ieee.org; hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/TASLP.2020.3045528

a kernel that consists of an outer product of SH functions, assuming that the sum of weighted kernels can approximate the SCM of a mixture represented in the SH domain. Since a plane wave in the SH domain consists of an SH function that is frequency-independent, the number of kernels for the method can be greatly reduced from that in the previous model, which provided frequency-dependent kernels [17]. Unfortunately, this method cannot model a near-field source efficiently because a spherical wave represented in the SH domain is dependent not only on the SH order but also on the frequencies, especially lower frequencies. This makes the method difficult to use for VR in practical situations where singers or speakers often perform close to microphones. Moreover, owing to decayed time-frequency (TF) regions in the low-frequency regime, also known as evanescent regions, the huge amount of information in such regions is not fully exploited [18]. In the context of sound field reproduction or modal beamformers, power compensation by the radial function is generally applied to make up for the loss of low-frequency energy during the acquisition [19]. In another work, the flexible audio source separation toolbox (FASST) [20] was employed in an HOA framework [16]. However, MNMF generally requires a huge computational cost even though dimensional reduction to the SH domain is performed.

In parallel with the advent of MNMF in the SH domain, an efficient blind source separation method based on non-negative tensor factorization (NTF) [21] in the wavenumber domain was proposed in [22]. This method exploits the sparse nature of an SCM in the wavenumber domain; the power of an SCM is concentrated at the diagonal part in the case of a uniform linear array, and the diagonal elements can be extracted to simplify the computation of its update rules, i.e., the replacement of matrix inversion with element-wise divisions. This work was further extended to tridiagonal modeling to cope with sidelobes from the diagonal elements caused by the finite size of a uniform linear array [23]. Another NTF method projects microphone signals to a beamspace domain [24]. However, the methods proposed so far do not provide a solution to cope with noisy evanescent regions when a small spherical or circular array is used as a recording device.

In this paper, we introduce an efficient NTF-based blind source separation method that can be used in practical situations to address the huge computational cost required for MNMF in the SH domain. As proposed in the literature for sound field reproduction, we first apply the inverse radial function to a mixed spectrogram tensor converted to the SH domain to amplify evanescent regions, so that the higher-order SH coefficients in the low-frequency regime can be recovered. Although the information in the evanescent region is important to improve the separation quality, the region in the cost function should be carefully treated because not only valid information but also measurement noise could be amplified. With this in mind, we devise a masking scheme to appropriately exclude TF bins from the NTF cost function, making the system sensitive only to parts of the region with a high signal-to-noise ratio (SNR). The masking scheme is also effective in modeling near-field sources because the frequency-dependent part can be eliminated from the NTF cost function and the mixture spectrogram tensor can be approximated with an efficient frequency-independent model.

The main contributions of this work are as follows:

- The HOA-encoding workflow is taken into account by applying the inverse radial function to compensate for the loss of beneficial information in the evanescent region;
- A frequency-independent NTF-based method in the SH domain with a masking scheme is devised to efficiently model both near-field and far-field sources; and
- The efficiency of the proposed method is evaluated in a realistic scenario where a mixture of spherical waves from near-field sources is obtained with a small practical array.

Note that the ideas of using a SH conversion or TF masking can also be integrated with other methods such as spatial clustering [25] or deep clustering [26]. In this paper, we focus our attention on NTF because it allows us to model the TF characteristics for each source without requiring training a model on a large dataset.

The paper is organized as follows: Section II shows the signal representation in the SH domain; Section III explains an SCM-based separation scheme; Section IV introduces our separation model based on NTF and a masking scheme to appropriately exploit the evanescent regions; and in Section V, we evaluate our proposed method under various environmental conditions.

The following notations are used throughout this paper: $\mathbf{x}$ denotes a column vector and $\mathbf{X}$ a matrix. The conjugate operator, trace, determinant, matrix transpose, conjugate matrix transpose, and Moore–Penrose pseudoinverse are denoted by $(.)^*$, $\mathrm{tr}\,(.)$, $\det\,(.)$, $(.)^T$, $(.)^H$, and $(.)^\dagger$ respectively.

II. Spherical-Harmonic-Domain Representation

An interior sound field in spherical coordinates can be expressed as a weighted sum of spherical Bessel functions $j_n(\cdot)$ and SH functions, i.e., $Y_{nm}(\cdot)$:

$$X({\mathchar'26\mkern-9mu k}, r, \Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}({\mathchar'26\mkern-9mu k}) j_n({\mathchar'26\mkern-9mu k} r) Y_{nm}(\Omega), \qquad (1)$$

where $X({\mathchar'26\mkern-9mu k}, r, \Omega)$ denotes an arbitrary narrowband sound field at radius r and angle Ω. The weight of the basis functions $\alpha_{nm}({\mathchar'26\mkern-9mu k})$ denotes the interior SH coefficient and ${\mathchar'26\mkern-9mu k}$ denotes the wavenumber. The SH order and degree are denoted by n and m, respectively. The SH function is defined with $\Omega = \{\theta, \phi\}$ [27] as

$$Y_{nm}(\Omega) \equiv \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\theta) e^{\mathrm{i}m\phi}, \qquad (2)$$

where $P_{nm}(\cos\theta)$, θ, and ϕ denote the associated Legendre function, the elevation, and the azimuth, respectively. The imaginary unit is denoted by i.

Given a sound field on the boundary of a sphere, the SH coefficient $\alpha_{nm}({\mathchar'26\mkern-9mu k})$ can be obtained by exploiting the orthonormality of the SH function:

$$\alpha_{nm}({\mathchar'26\mkern-9mu k}) = \frac{1}{j_n({\mathchar'26\mkern-9mu k} R_\mathrm{b})} \int_\Omega X({\mathchar'26\mkern-9mu k}, R_\mathrm{b}, \Omega) Y_{nm}^*(\Omega) d\Omega, \qquad (3)$$

with

$$\int_{\Omega} Y_{nm}(\Omega) Y^{*}_{n'm'}(\Omega) d\Omega = \delta_{nn'} \delta_{mm'}, \tag{4}$$

where R_{b} denotes the radius of the boundary and $\delta_{nn'}$ denotes the Kronecker delta function defined as

$$\delta_{nn'} = \begin{cases} 1 \,(n = n') \\ 0 \,(n \neq n') \end{cases}. \tag{5}$$

This process has been referred to as the HOA encoding process in the literature [28].

In practice, the SH coefficients cannot be retrieved from (3) owing to the numerical instability of the spherical Bessel function in the denominator, which is often referred to as the Bessel zero problem or forbidden frequency problem [29]. To avoid this problem, several microphone arrays have been proposed [30]–[33]. When an array with a rigid baffle is used, the encoding process in (3) can be modified by using a radial function $b_n(\cdot)$ for a rigid baffle array, i.e.,

$$\alpha_{nm}(\bar{k}) = \frac{1}{b_n(\bar{k} R_{\mathrm{b}})} \int_{\Omega} X(\bar{k}, R_{\mathrm{b}}, \Omega) Y^{*}_{nm}(\Omega) d\Omega, \tag{6}$$

with

$$b_n(\bar{k} R_{\mathrm{b}}) = j_n(\bar{k} R_{\mathrm{b}}) - \frac{j'_n(\bar{k} R_{\mathrm{b}})}{h_n^{(1)'}(\bar{k} R_{\mathrm{b}})} h_n^{(1)}(\bar{k} R_{\mathrm{b}}), \tag{7}$$

where $h_n^{(1)}(\cdot)$ denotes the nth-order spherical Hankel function of the first kind, and $h_n^{(1)'}(\cdot)$ and $j'_n(\cdot)$ denote the derivatives of $h_n^{(1)}(\cdot)$ and $j_n(\cdot)$, respectively.

When the acoustic pressure on a sphere can be obtained by an appropriate sampling scheme, e.g., the quadrature method [29], the spatially discretized version of the encoding process in (6) can be represented in matrix form. Given that A microphones are used and the SH order is limited by N, an SH vector can be computed from an input vector with a frequency-dependent diagonal matrix $\mathbf{B}(\bar{k}) \in \mathbb{C}^{L \times L}$, where $L = (N+1)^2$, and a frequency-independent matrix $\mathbf{Y} \in \mathbb{C}^{A \times L}$:

$$\boldsymbol{\alpha}(\bar{k}) = \mathbf{B}(\bar{k})^{-1} \mathbf{Y}^{\dagger} \mathbf{x}(\bar{k}), \tag{8}$$

with

$$\mathbf{B}(\bar{k}) = \begin{pmatrix} b_0(\bar{k} R_{\mathrm{mic}}) & 0 & \dots & 0 \\ 0 & b_1(\bar{k} R_{\mathrm{mic}}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_N(\bar{k} R_{\mathrm{mic}}) \end{pmatrix}, \tag{9}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}(\Omega_1) & \mathbf{y}(\Omega_2) & \cdots & \mathbf{y}(\Omega_A) \end{bmatrix}^T, \tag{10}$$

where $\mathbf{y}(\Omega) \in \mathbb{C}^L$ is a vector containing $Y_{nm}(\Omega)$. The input vector and the encoded input vector are denoted as $\mathbf{x}(\bar{k}) \in \mathbb{C}^A$ and $\boldsymbol{\alpha}(\bar{k}) \in \mathbb{C}^L$, respectively. The diagonal part of $\mathbf{B}(\bar{k})$ and the SH functions in $\mathbf{y}(\Omega)$ are sorted with the SH index l such that $l = n^2 + n + m + 1$ holds. The radius of the spherical microphone array is denoted by R_{mic}.

The truncation order N is usually determined by adopting the criterion $N = \lceil e\bar{k}R/2 \rceil$ [34] or $N = \lceil \bar{k}R \rceil$ [35], where R is the radius of the region of interest, e.g., the radius of the spherical microphone array. When the number of microphones A is known, it is also important to limit the SH order to $(N+1)^2 \leq A$ to avoid spatial aliasing.

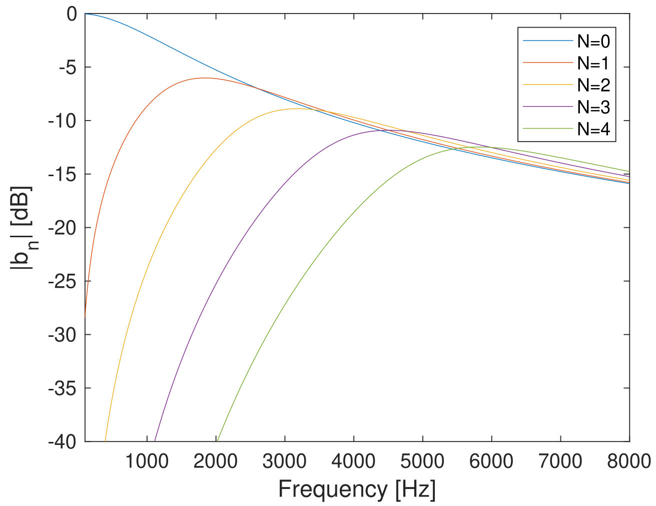

Fig. 1. Amplitude of radial function in dB scale for different SH orders when R_{b} is set to 0.042 m.

Fig. 1 shows the amplitude of the radial function b_n in dB scale for a rigid spherical microphone array with a radius of 0.042 m. It is clear from the figure that the response of the radial function for the high SH orders in the low-frequency regime, which corresponds to the evanescent region, decays drastically. This makes it difficult to fully exploit informative higher-order SH coefficients in the low frequency regime for source separation.

III. Multichannel Blind Source Separation

A. Local Gaussian Model

This section provides an introduction to the model underlying MNMF. The model assumes that an A-channel vector of a short-time Fourier transform (STFT) bin for ith source can be modeled as a multivariate complex Gaussian, i.e.,

$$\mathbf{s}^i_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}^i_{ft}\right), \tag{11}$$

where $\mathbf{s}^i_{ft} \in \mathbb{C}^A$ denotes the spatial image of the ith source in the STFT domain, $\mathbf{R}^i_{ft} = \mathbb{E}[\mathbf{s}^i_{ft} \mathbf{s}^{i\ H}_{ft}] \in \mathbb{C}^{A \times A}$ denotes the covariance matrix of the complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}$, f is the frequency bin index that corresponds to $\bar{k}$, and t is the time frame index.

The spatial image of a mixture of multiple sources $\mathbf{x}_{ft} \in \mathbb{C}^A$ is represented as a sum of complex Gaussians, i.e.,

$$\mathbf{x}_{ft} = \sum_i \mathbf{s}^i_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}_{ft}\right), \tag{12}$$

where $\mathbf{R}_{ft} \in \mathbb{C}^{A \times A}$ denotes the SCM. Assuming that the sources are mutually independent, the SCM of the mixture $\mathbf{R}_{ft}$ is given by the sum of the SCMs of all sources, i.e.,

$$\mathbf{R}_{ft} = \mathbb{E}\left[\mathbf{x}_{ft} \mathbf{x}^H_{ft}\right] = \sum_i \mathbf{R}^i_{ft}. \tag{13}$$

The log-likelihood of the spatial image $\mathbf{x}_{ft}$ for the model parameters $\boldsymbol{\varphi}$ is given by

$$\log \mathbb{P}(\mathbf{x}|\boldsymbol{\varphi}) = \sum_{ft} \log \mathcal{N}_{\mathbb{C}} \left(\mathbf{x}_{ft} | \mathbf{0}, \hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) \right), \tag{14}$$

where the SCM of the mixture is modeled by $\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi})$ and the parameter $\boldsymbol{\varphi}$ will be defined in the next section. The maximization of this likelihood can be interpreted as the minimization of the log-determinant divergence [36] between the empirical SCM, $\tilde{\mathbf{R}}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^H$, and the estimated SCM, $\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) \in \mathbb{C}^{A \times A}$:

$$\begin{aligned} C(\boldsymbol{\varphi}) &= \sum_{ft} D_{\mathrm{LD}} \left(\tilde{\mathbf{R}}_{ft} | \hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) \right) \\ &\equiv \sum_{ft} \operatorname{tr} \left(\tilde{\mathbf{R}}_{ft} \hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi})^{-1} \right) + \log \det \left(\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) \right), \end{aligned} \tag{15}$$

where $C(\boldsymbol{\varphi})$ can be seen as a cost function that we want to minimize with respect to the model parameters $\boldsymbol{\varphi}$. We denote the log-determinant divergence by D_{LD}.

B. Multichannel NMF

In the framework of MNMF proposed in [37] where $\boldsymbol{\varphi} = \{\mathbf{Q}_{fk}, w_{fk}, h_{kt}\}$, the SCM $\hat{\mathbf{R}}_{ft}$ is assumed to be a superposition of time-invariant SCMs $\mathbf{Q}_{fk} \in \mathbb{C}^{A \times A}$ coupled with a scale value that represents the power spectral density. The scale value is decomposed into a non-negative frequency weight w_{fk} and a non-negative activation h_{kt},

$$\hat{\mathbf{R}}_{ft}(\boldsymbol{\varphi}) = \sum_{k} \mathbf{Q}_{fk} w_{fk} h_{kt}, \tag{16}$$

where the NMF component index is denoted by k. Updating the model parameters requires a huge computational cost of order $O(A^3)$. Assuming that sound field reproduction up to the fourth order is desired, at least $(4+1)^2 = 25$ microphones are required to avoid spatial aliasing, which makes the method infeasible in practical situations.

IV. Proposed Method

A. Spherical Harmonic Transform

The STFT multichannel signals are converted into the SH domain, where the underlying probability model is based on the zero-mean complex Gaussian distribution,

$$\boldsymbol{\alpha}_{ft} = \mathbf{B}_f^{-1} \mathbf{Y}^{\dagger} \mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \mathbf{R}_{ft}^{\mathrm{SHD}} \right), \tag{17}$$

with

$$\mathbf{R}_{ft}^{\mathrm{SHD}} = \mathbf{B}_f^{-1} \mathbf{Y}^{\dagger} \mathbf{R}_{ft} (\mathbf{Y}^{\dagger})^H (\mathbf{B}_f^{-1})^H, \tag{18}$$

where $\mathbf{R}_{ft}^{\mathrm{SHD}} \in \mathbb{C}^{L \times L}$ denotes the SCM in the SH domain. Different from the prior works [15], [16], we apply the inverse radial function, $\mathbf{B}_f^{-1}$, to amplify the evanescent region as often done in the HOA workflow [28]. For source separation, this works as pre-processing to highlight the beneficial information buried in the evanescent region, i.e., $n > \lceil e\bar{k} R_{\mathrm{mic}}/2 \rceil$.

Log-determinant divergence similar to (15) can be constructed in the SH domain, i.e.,

$$C_{\mathrm{SHD}}(\boldsymbol{\varphi}) = \sum_{ft} D_{\mathrm{LD}} \left(\tilde{\mathbf{R}}_{ft}^{\mathrm{SHD}} | \hat{\mathbf{R}}_{ft}^{\mathrm{SHD}}(\boldsymbol{\varphi}) \right), \tag{19}$$

where $\tilde{\mathbf{R}}_{ft}^{\mathrm{SHD}} = \boldsymbol{\alpha}_{ft} \boldsymbol{\alpha}_{ft}^H$ denotes the empirical SCM in the SH domain and can be obtained in a similar way to (18), i.e.,

$$\tilde{\mathbf{R}}_{ft}^{\mathrm{SHD}} = \mathbf{B}_f^{-1} \mathbf{Y}^{\dagger} \tilde{\mathbf{R}}_{ft} (\mathbf{Y}^{\dagger})^H (\mathbf{B}_f^{-1})^H. \tag{20}$$

An SCM in the SH domain can be modeled with a frequency-dependent matrix $\mathbf{Q}_{fk}^{\mathrm{SHD}}$:

$$\hat{\mathbf{R}}_{ft}^{\mathrm{SHD}} = \sum_{k} \mathbf{Q}_{fk}^{\mathrm{SHD}} w_{fk} h_{kt}, \tag{21}$$

with

$$\mathbf{Q}_{fk}^{\mathrm{SHD}} = \mathbf{B}_f^{-1} \mathbf{Y}^{\dagger} \mathbf{Q}_{fk} (\mathbf{Y}^{\dagger})^H (\mathbf{B}_f^{-1})^H. \tag{22}$$

Given $L = (N+1)^2 \leq A$, the computational cost for updating the parameters can be reduced to $O(L^3)$, but the model is still not efficient enough to be used in practice.

B. Diagonal Extraction and Frequency-Independent Model

As the update of the SH-domain SCM, $\mathbf{Q}_{fk}^{\mathrm{SHD}}$, still requires a computational cost of order $O(L^3)$, we further simplify the cost function (21) to reduce the computational cost by taking into account only the diagonal part of the matrices and by eliminating the frequency dependence of $\mathbf{Q}_{fk}^{\mathrm{SHD}}$:

$$C_{\mathrm{SHD}}(\boldsymbol{\varphi}) \approx \sum_{lft} D_{\mathrm{LD}} \left(\gamma_{lft} | \hat{\gamma}_{lft} \right), \tag{23}$$

with

$$\hat{\gamma}_{lft} = \sum_{k} q_{lk} w_{fk} h_{kt}, \tag{24}$$

where γ_{lft} and q_{lk} denote the diagonal elements of $\tilde{\mathbf{R}}_{ft}^{\mathrm{SHD}}$ and $\mathbf{Q}_{fk}^{\mathrm{SHD}}$, respectively. With the diagonal extraction, the computational cost can be reduced to $O(L)$. The elimination of the frequency dependence also significantly reduces the computational cost. Although the frequency-independent model should have difficulty in approximating near-field sources, the masking scheme explained in the next section can alleviate the discrepancy between the observed signals and the model.

C. Evanescent-Region-Aware Masking for Near-Field Source

In the prior work [15], the authors did not employ the inverse radial function in the SH transform, and as a result, the system could not fully exploit the beneficial information for source separation hidden in the evanescent region, i.e., $n > \lceil e\bar{k} R_{\mathrm{mic}}/2 \rceil$. Furthermore, the Euclidean distance used in the cost function exacerbates this problem because the scale-variant nature of the distance makes the parameter estimation disregard TF bins that have a small amplitude. Even if the method incorporates the inverse radial function into the SH transform, the method based on a frequency-independent model is assumed to struggle with near-field sources that possess extremely high power for

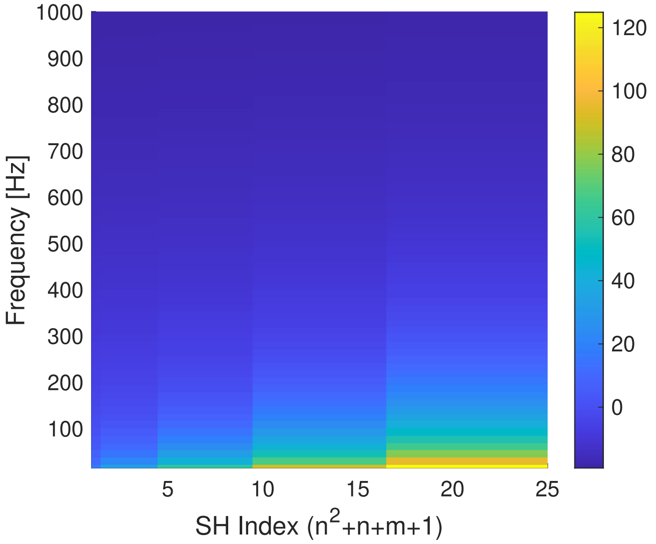

Fig. 2. $|\mathrm{i}\bar{k}h_n^{(1)}(\bar{k}R_{\mathrm{src}})|$ in dB scale where the radius of the source, R_{src}, is set to 0.5 m.

the higher-order SH coefficients in the low-frequency regime. To understand the concept more clearly, here we introduce a spherical wave represented in the SH domain [27], i.e.,

$$\alpha_{nm}(\bar{k}) = \mathrm{i}\bar{k}h_n^{(1)}(\bar{k}R_{\mathrm{src}})Y_{nm}^*(\Omega_{\mathrm{src}}), \tag{25}$$

where R_{src} and Ω_{src} denote the radius and the direction of the target source, respectively. There are wavenumber variables both inside and outside the spherical Hankel function in (25), which is not the case for a plane wave. Fig. 2 shows the frequency-dependent nature of a spherical wave where the radius of the source, R_{src}, is set to 0.5 m. While salient values can be observed when $n > \lceil e\bar{k}R_{\mathrm{src}}/2\rceil$, excluding this region makes the matrix flatter and more redundant with respect to frequencies. This can be explained by the large argument limits ($\bar{k} > 2n/(eR_{\mathrm{src}})$) of the spherical Hankel functions [27], i.e.,

$$h_n^{(1)}(\bar{k}R_{\mathrm{src}}) \approx (-\mathrm{i})^{n+1}\frac{e^{\mathrm{i}\bar{k}R_{\mathrm{src}}}}{\bar{k}R_{\mathrm{src}}}. \tag{26}$$

Given that R_{src} is fixed, the spherical Hankel function becomes close to the representation of a plane wave. The opposite case, i.e., fixing $\bar{k}$ while increasing R_{src}, can be more intuitively understood; a spherical wave from a far-field source can be approximated by a plane wave. By substituting (26) into (25), the amplitude of $\alpha_{nm}(\bar{k})$ no longer has a frequency dependence, thus making the matrix in Fig. 2 more redundant with respect to frequencies, which is suitable for a low-rank model such as NTF.

In this work, we focus on exploiting the characteristics of the spherical Hankel function, where the frequency dependence is strong only in the high-power region, i.e., $n > \lceil e\bar{k}R_{\mathrm{src}}/2\rceil$, and the other region can be approximated with a frequency-independent low-rank model. To this end, we devise a masking scheme that can eliminate the effect of the high-power region by applying a TF mask to the cost function. The masking scheme has another advantage that the noisy region amplified by the power compensation (6) can be excluded from the approximation of a mixed spectrogram tensor. Note that this noisy region should reside in the evanescent region, i.e., $n > \lceil e\bar{k}R_{\mathrm{mic}}/2\rceil$.

In the field of sound field reproduction, a continuous mask based on a trigonometric function was introduced in [38] to improve the perceptual quality of a reconstructed spherical wave. In contrast to [38], we apply a binary mask crafted by the method explained in the following section to improve the separation quality. The cost function can then be modified with a binary mask ξ_{lft} as follows:

$$C_{\mathrm{eSHD}}(\boldsymbol{\varphi}) = \sum_{lft}\xi_{lft}D_{\mathrm{LD}}\left(\gamma_{lft}|\hat{\gamma}_{lft}\right). \tag{27}$$

Again, this does not contradict our concept of modeling near-field sources because the frequency-dependent nature of the spherical Hankel function is particularly dominant in TF bins for $n > \lceil e\bar{k}R_{\mathrm{src}}/2\rceil$, and these TF bins are excluded from the NTF cost function by the binary mask, ξ_{lft}. Since it is often difficult to know R_{src} in practice, we discuss how to generate the mask in the next section.

D. Mask Generation

First, we can consider exploiting the radius of a spherical microphone array to determine the size of the evanescent region, i.e., $n > \lceil e\bar{k}R_{\mathrm{mic}}/2\rceil$, to mask. Using the radius of a spherical microphone array is straightforward and advantageous from the practical viewpoint because a mask can be generated prior to source separation as soon as the microphone setting including the array geometry is determined:

$$\xi_{lf}^{\mathrm{mic}} = \begin{cases} 1\,(n \leq \lceil e\bar{k}R_{\mathrm{mic}}/2\rceil) \\ 0\,(n > \lceil e\bar{k}R_{\mathrm{mic}}/2\rceil) \end{cases}. \tag{28}$$

We omit the time dependence from ξ_{lf}^{mic} as the values do not change with time.

However, information is lost in the range $\lceil e\bar{k}R_{\mathrm{mic}}/2\rceil < n \leq \lceil e\bar{k}R_{\mathrm{src}}/2\rceil$, which can still be beneficial for separating sources with the model (24). Obtaining R_{src} prior to source separation is challenging because the radii of target sources are generally not given to the system in advance. To identify the source having the minimum distance to the origin, we propose computing a threshold to generate a mask by multiplying the averaged spectrogram in the non-evanescent region, i.e., the masked region with ξ_{lf}^{mic}, by a constant value κ:

$$\xi_{lft} = \begin{cases} 1\,(\tilde{\gamma}_{lft} \leq \Xi_t) \\ 0\,(\tilde{\gamma}_{lft} > \Xi_t) \end{cases}, \tag{29}$$

with

$$\tilde{\gamma}_{lft} = 10^{\log_{10}\gamma_{lft} *_{lf} \varpi_{lf}^{\mathrm{Gauss}}}, \tag{30}$$

$$\Xi_t = \frac{\kappa}{LF}\sum_{lf}\gamma_{lft}\xi_{lf}^{\mathrm{mic}}, \tag{31}$$

where $\varpi_{lf}^{\mathrm{Gauss}}$ denotes the two-dimensional Gaussian kernel that can work as a smoothing function. The two-dimensional convolution is denoted by $*_{lf}$. The number of frequency bins is denoted by F.

With this mask, the model can efficiently exploit the redundancy of a mixed spectrogram tensor, and the cost function can exclude a low-SNR region in which the measurement noise is significantly amplified.

E. Derivation of Update Rules

To minimize the cost function in (27) with (24), we employ the majorization–minimization (MM) algorithm [39] to reduce the cost monotonically. The auxiliary function can be constructed by applying Jensen's inequality to the convex part of the cost function and the tangent line to the concave part:

$$C^{+}_{\text{eSHD}}(\boldsymbol{\varphi}, \eta_{lftk}, \psi_{lft}) = \sum_{lft} \xi_{lft} \Bigg(\sum_{k} \eta_{lftk}^{2} \frac{\gamma_{lft}}{q_{lk} w_{fk} h_{kt}} + \log u_{lft} + \frac{\gamma_{lft} - \psi_{lft}}{\psi_{lft}} \Bigg), \tag{32}$$

where η_{lftk} and ψ_{lft} are auxiliary variables that satisfy $\sum_k \eta_{lftk} = 1, \eta_{lftk} \geq 0$, and $\psi_{lft} \geq 0$. The equality of the auxiliary function and the cost function holds only when the auxiliary variables satisfy

$$\eta_{lftk} = \frac{q_{lk} w_{fk} h_{kt}}{\hat{\gamma}_{lft}}, \tag{33}$$

$$\psi_{lft} = \hat{\gamma}_{lft}. \tag{34}$$

The partial derivatives with respect to q_{lk}, w_{fk}, and h_{kt} are derived by minimizing the auxiliary function:

$$\frac{\partial C^{+}_{\text{eSHD}}}{\partial q_{lk}} = \sum_{ft} \xi_{lft} \left(-\eta_{lftk}^{2} \frac{\gamma_{lft}}{q_{lk}^{2} w_{fk} h_{kt}} + \frac{w_{fk} h_{kt}}{\psi_{lft}} \right), \tag{35}$$

$$\frac{\partial C^{+}_{\text{eSHD}}}{\partial w_{fk}} = \sum_{lt} \xi_{lft} \left(-\eta_{lftk}^{2} \frac{\gamma_{lft}}{q_{lk} w_{fk}^{2} h_{kt}} + \frac{q_{lk} h_{kt}}{\psi_{lft}} \right), \tag{36}$$

$$\frac{\partial C^{+}_{\text{eSHD}}}{\partial h_{kt}} = \sum_{lf} \xi_{lft} \left(-\eta_{lftk}^{2} \frac{\gamma_{lft}}{q_{lk} w_{fk} h_{kt}^{2}} + \frac{q_{lk} w_{fk}}{\psi_{lft}} \right). \tag{37}$$

The update rules can be derived by equating the derivative to zero and can be simplified by substituting (33) and (34):

$$q_{lk} \leftarrow q_{lk} \sqrt{\frac{\sum_{ft} \frac{\xi_{lft} \gamma_{lft}}{\hat{\gamma}_{lft}^{2}} w_{fk} h_{kt}}{\sum_{ft} \frac{\xi_{lft}}{\hat{\gamma}_{lft}} w_{fk} h_{kt}}}, \tag{38}$$

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_{lt} \frac{\xi_{lft} \gamma_{lft}}{\hat{\gamma}_{lft}^{2}} q_{lk} h_{kt}}{\sum_{lt} \frac{\xi_{lft}}{\hat{\gamma}_{lft}} q_{lk} h_{kt}}}, \tag{39}$$

$$h_{kt} \leftarrow h_{kt} \sqrt{\frac{\sum_{lf} \frac{\xi_{lft} \gamma_{lft}}{\hat{\gamma}_{lft}^{2}} q_{lk} w_{fk}}{\sum_{lf} \frac{\xi_{lft}}{\hat{\gamma}_{lft}} q_{lk} w_{fk}}}. \tag{40}$$

TABLE I
EXPERIMENTAL SETUP

Sampling rate	16 kHz
STFT frame size	1024
STFT hop size	512
Number of iterations	100
Number of initializations	5
Number of sources	4
Number of NMF components	24
Variation of angles	8
Clustering method	Mel NMF [42]

Algorithm 1: SH-Domain NTF With Masking Scheme.

```
Require: Mixture $\mathbf{x}_{ft}$
  Compute $\tilde{\mathbf{R}}_{ft} = \mathbf{x}_{ft}\mathbf{x}_{ft}^{H}$
  Apply spatial transform to $\tilde{\mathbf{R}}_{ft}$ with (20)
  Initialize $q_{lk}, w_{fk}, h_{kt}$ with randomized values
  Compute $\hat{\gamma}_{lft}$ with (24)
  Compute $\xi_{lft}$ with (28)–(31)
  for $\jmath = 0$ to MM iteration do
      $q_{lk} \leftarrow$ (38)
      Compute $\hat{\gamma}_{lft}$ with (24)
      $w_{fk} \leftarrow$ (39)
      Compute $\hat{\gamma}_{lft}$ with (24)
      $h_{kt} \leftarrow$ (40)
      Compute $\hat{\gamma}_{lft}$ with (24)
   end for
   Apply Wiener filtering with (41)
  Ensure: Estimates $\hat{\mathbf{s}}^{i}_{ft}$
```

F. Wiener Filtering in Spherical Harmonic Domain

Given the estimated model parameters, the STFT coefficients of each source can be recovered by a multichannel Wiener filter, i.e., a minimum mean squared error (MMSE) estimator [40]. The MMSE estimator in the SH domain is given by

$$\hat{\mathbf{s}}^{i}_{ft} = \mathbf{Y}\mathbf{B}_{k} \left(\sum_{k \in K_i} \mathbf{Q}_{k}^{\text{Diag}} w_{fk} h_{kt} \right) \left(\hat{\mathbf{R}}_{ft}^{\text{Diag}} \right)^{-1} \mathbf{B}_{k}^{-1} \mathbf{Y}^{\dagger} \mathbf{x}_{ft}, \tag{41}$$

where K_i denotes the set of NMF components that belong to the ith source and can be determined by a clustering method, such as LPC-based [41] or Mel-spectrum-based clustering [42]. The diagonal matrices having q_{lk} and $\hat{\gamma}_{lft}$ in the diagonal parts are denoted by $\mathbf{Q}_{k}^{\text{Diag}}$ and $\hat{\mathbf{R}}_{ft}^{\text{Diag}}$, respectively. The inverse of $\hat{\mathbf{R}}_{ft}^{\text{Diag}}$ can be efficiently calculated as the matrix is diagonal.

V. EVALUATION

A. Experimental Conditions

To evaluate the proposed method, several experiments were conducted in a music separation scenario. The experimental conditions in common in this section are listed in Table I. As evaluation metrics, we employed the source-to-distortion ratio (SDR) improvement, the source-to-interferences ratio (SIR) improvement, and the sources-to-artifacts ratio (SAR). The improvements for SDR and SIR were calculated by subtracting the values of a mixture from the values of each method. Note that we did not compute SAR improvements as the input SAR is infinity and, therefore, the improvement is not computable. Four-second-long sources of four types of musical instruments were obtained from the songKitamura dataset.[1] As a MIDI source, Garritan Personal Orchestra 4 was selected because it is considered more realistic than the other MIDI sources in the same dataset (See [43] for more information about the dataset).

B. Acoustic Simulation

A scenario where a spherical microphone array captures a sound field generated by near-field sound sources in a free field was achieved by an acoustic simulation. For microphones on a spherical microphone array, the same microphone positions as those in the em32 Eigenmike [2] were applied, i.e., 12 at the apexes of a regular icosahedron and 20 at those of a regular dodecahedron (in total $A = 32$). The radius of the array was set to 0.042 m, and the surface of the array was assumed to be covered with a rigid baffle, which was simulated by setting the reflection coefficient to 1. The azimuth and elevation angles of the four sources were chosen from the angles assigned to the microphone positions. The distance of each sound source was sampled from a normal distribution with a mean of 2 m and a standard deviation of 1 m. To avoid a collision between sources and the array, the minimum radius of sources was set to be at least 0.5 m larger than the array radius. The four instrument sound sources were permuted across the source positions so that the results did not depend on the TF characteristics of the four instrument sound sources. In total, we consider 160 samples in each experiment as we use five initializations for eight angles combined with four permuted source allocations.

The transfer function $G(r, \Omega | R_{\mathrm{src}}, \Omega_{\mathrm{src}}; \hbar)$ from the sound source to the microphone, including scattering on the surface of the spherical microphone array, was calculated in the SH domain as

$$G(r, \Omega | R_{\mathrm{src}}, \Omega_{\mathrm{src}}; \hbar) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \mathrm{i}\hbar h_n^{(1)}(\hbar R_{\mathrm{src}}) Y_{nm}^*(\Omega_{\mathrm{src}}) \times b_n(\hbar r) Y_{nm}(\Omega). \quad (42)$$

To simulate spatial aliasing from the high-order SH coefficients, the upper limit of the SH order was set to $N = 30$ considering the Nyquist frequency, 8 kHz, used in the evaluation; the power of high-order SH coefficients is sufficiently attenuated at $N = 30$ because $\lceil e\hbar R_{\mathrm{mic}}/2 \rceil = \lceil e \times 2\pi \times 8000/343 \times 0.042/2 \rceil = 9$. The spatial aliasing caused by microphones was simulated in [14] but not in [15]. Each microphone was subjected to measurement noise set to $\mathrm{SNR} = 40$ dB.

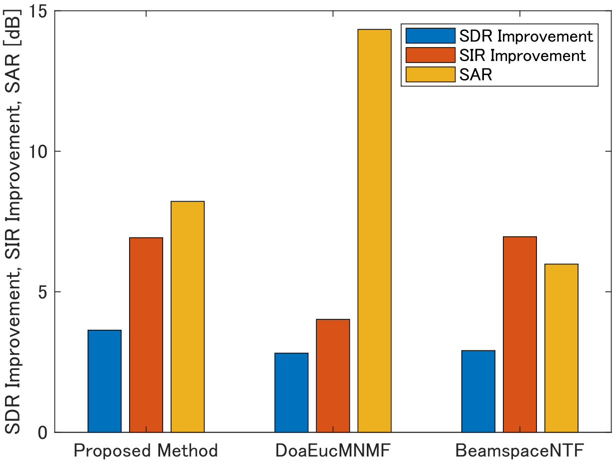

Fig. 3. Comparison of different methods in the simulated condition.

C. Comparison With Other Methods

For the proposed method, we generated a binary mask with (28)–(31) where the constant value κ was set to 2^{27}. The standard deviation of the Gaussian kernel was set to two. In addition to the proposed method, DOA-based Euclidean MNMF in the SH domain (DoaEucMNMF) [15] and Itakura–Saito NTF in the beamspace domain (BeamspaceNTF) [24] were evaluated. As in the original work, DoaEucMNMF was evaluated without power compensation of the inverse radial function because the Euclidean distance used as a cost function in [15] is scale-variant and sensitive to a high power; the method mostly approximates the low-SNR region boosted by the inverse radial function. BeamspaceNTF converts signals to the so-called beamspace domain, where the diagonalization of an SCM can be assumed except in the low-frequency regime. The number of DOA kernels, i.e., the beamspace dimension, was set to 32. For all the methods, the number of NMF components was set to 24. The number of iterations was set to 100 to guarantee the convergence of all the methods.

Fig. 3 shows the evaluation results in terms of the SDR improvement, SIR improvement, and SAR. The SDR improvement of the proposed method was higher than those of the other methods. For BeamspaceNTF, although the diagonal approximation worked as expected at high frequencies, the diagonal approximation was not achieved at low frequencies owing to the rank deficiency. DoaEucMNMF did not perform well in separating sources because the model cannot exploit beneficial information hidden in the evanescent region, although the SAR value was highest among the methods.

We also conducted an experiment using PEASS [44], [45], which is known to show a good agreement with the human perception of sound quality. Our proposed method also outperformed other methods in overall perceptual score (OPS) improvement: 13.50 for the proposed method, 6.75 for DoaEucMNMF, and 10.17 for BeamspaceNTF, reqpectively.

D. Validation of Masking Scheme

In Section V-C, it was confirmed that applying the binary mask ξ_{lft} was effective in improving the separation quality. In

[1] [Online]. Available: http://d-kitamura.net/dataset.htm

[2] [Online]. Available: https://mhacoustics.com/

TABLE II
SDR IMPROVEMENTS, SIR IMPROVEMENTS AND SARS WITH DIFFERENT MASKS

Masking Scheme	No mask	Proposed Method	Threshold $n = \lceil e\bar{k}R_{\mathrm{mic}}/2 \rceil$	Threshold $n = \lceil e\bar{k}R_{\mathrm{min}}/2 \rceil$
SDR Improvement	3.01	3.64	2.11	3.96
SIR Improvement	6.04	6.92	6.51	7.81
SAR	7.76	8.22	5.34	7.05

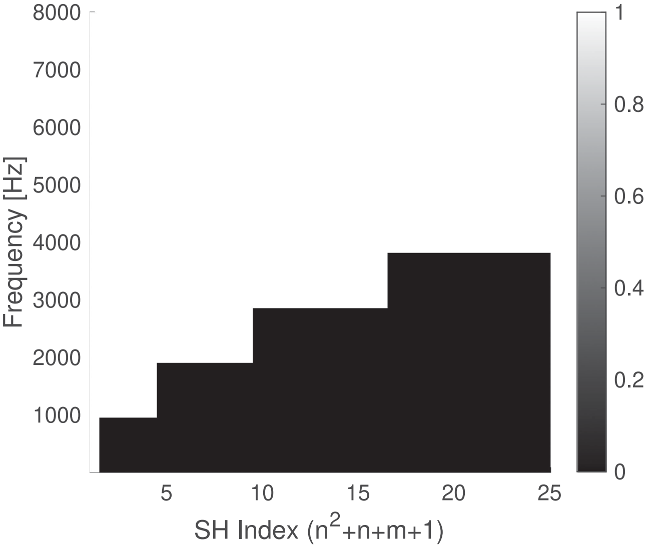

Fig. 4. Evanescent-region-aware mask generated with threshold value $n = \lceil e\bar{k}R_{\mathrm{mic}}/2 \rceil$ where R_{mic} is set to 0.042 m.

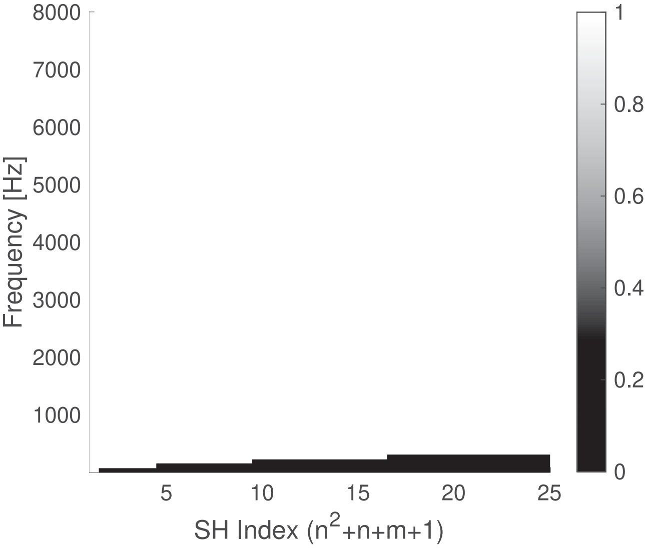

Fig. 5. Evanescent-region-aware mask generated with threshold value $n = \lceil e\bar{k}R_{\mathrm{src}}/2 \rceil$ where R_{src} is set to 0.5 m.

this experiment, several masks were generated by the following methods to confirm that the proposed mask generation schemes described in Section IV-D are valid:

1) Mask generated by setting all the values to one (no mask applied);
2) Mask generated with (28)–(31);
3) Mask generated with $n = \lceil e\bar{k}R_{\mathrm{mic}}/2 \rceil$ as a threshold described in (28); and
4) Mask generated with $n = \lceil e\bar{k}R_{\mathrm{min}}/2 \rceil$ as a threshold where the smallest radius of the four sources, R_{min}, was given in advance as an oracle.

The mask generated with $n = \lceil e\bar{k}R_{\mathrm{mic}}/2 \rceil$ as in Fig. 4 can be considered as a reasonable choice in practice because the threshold can be computed as soon as the array type is determined. However, the region to be masked is quite large when R_{mic} is small as shown in Fig. 4, and there are many informative TF bins with a high SNR in the region. In contrast, for the mask generated with $n = \lceil e\bar{k}R_{\mathrm{min}}/2 \rceil$, many TF bins with a high SNR can be exploited while the integration with source localization is essential. Fig. 5 shows a mask when $R_{\mathrm{src}} = 0.5$ m. The coverage of the mask is significantly smaller than that when applying a mask with $R_{\mathrm{mic}} = 0.042$ m.

Table II shows the SDR improvements of the above masking schemes. When a mask is not applied, the NTF model (24), which does not have a frequency-dependent q_{lk}, fails to cope with near-field sources. While the masking scheme that uses the microphone radius does not perform well owing to the lack of informative TF bins, the one that uses the distance of the closest source performs the best. The value yielded by knowing the oracle distance of the closest source can be regarded as the upper bound of the time-invariant masking scheme. Although we assumed that the proposed method with the time-variant masking scheme should have the potential to outperform the upper bound of the time-invariant masking scheme, the result did not meet our expectation. This is partly because the determination of κ is heuristic, and it may be possible to find the optimal value if we could conduct an exhaustive search for κ.

E. Effect of Spatial Aliasing and Measurement Noise

In [15], the SH order of the transfer function employed in the acoustic simulation was truncated at the fourth order. Therefore, the simulation did not take into account the influence of spatial aliasing that should occur in reality. An error analysis regarding spherical microphone designs was conducted in [7], [46], [47]. In this experiment, we evaluate the proposed method with two types of transfer functions computed by (42): a more realistic case ($N = 30$) and a less realistic case ($N = 4$). In addition, we evaluated the proposed method with or without adding measurement noise under two different SH-order scenarios ($N = 4, 30$). When measurement noise is not present in the input signals, a high SNR can be expected in the whole evanescent region. Therefore, it is considered that the application of the mask in the proposed method has a marginal effect on the separation performance. The measured noise values were set to $\mathrm{SNR} = 40$ dB and ∞.

TABLE III
EFFECT OF SPATIAL ALIASING AND MEASUREMENT NOISE

Metric	Noise level	Proposed Method ($N = 4$)	Proposed Method ($N = 30$)	No mask ($N = 30$)
SDR Improvement	SNR $= 40$	3.51	3.64	3.01
	SNR $= \infty$	4.75	5.23	5.21
SIR Improvement	SNR $= 40$	6.95	6.92	6.04
	SNR $= \infty$	8.40	8.60	8.62
SAR	SNR $= 40$	8.01	8.22	7.76
	SNR $= \infty$	8.03	8.61	8.59

TABLE IV
COMPUTATION TIME

Approach	time [s]
Proposed Method	30.94
DoaEucMNMF [15]	1584.27
BeamspaceNTF [24]	268.51

Table III shows the SDR improvements of the proposed methods with and without adding measurement noise under the two SH-order scenarios ($N = 4, 30$). In contrast to our assumption, the proposed method with $N = 30$ performs better than the one with $N = 4$ regardless of whether or not measurement noise exists in the input signals. As expected, when there is no measurement noise, it can be seen that the application of the mask to the evanescent region has a marginal effect. In contrast, when measurement noise is present, the improvement resulting from applying a mask to the evanescent region can be clearly observed.

F. Computation Time

The computation times for all the methods compared in Section V-C were measured to highlight the efficiency of the proposed method. The measurement was carried out by inserting MATLAB time commands in the space before inputting the STFT signals and after outputting the separated STFT signals. We ran the programs on a Xeon E5-2620 v4 CPU where each core has 2.1 GHz CPU capability. The results are listed in Table IV. By comparing the proposed method with DoaEucMNMF, the efficiency of the proposed method based on NTF can be clearly observed. We can expect that the difference in efficiency between the algorithms will be even more significant when the SH order, N, is high. For BeamspaceNTF, the speed can be improved if the dimension of the beamspace is reduced, which could, however, sacrifice the separation performance owing to the reduction of the spatial resolution.

G. Difference in Number of Microphones

In this experiment, we evaluated our proposed method with different number of microphones (12, 20, 32) in terms of separation performance and computation time. The experimental setup is the same as in Section V-D and Section V-F. No additional measurement noise was injected and, hence, the SNR is the same, independent of the number of microphones. From Table V, an increase of the SDR improvements can be observed with increasing the number of microphones, and we can conclude that more SH coefficients improve the separation performance while requiring more computational resources. We could also verify from the difference in computation time that the computational cost of our proposed method is approximately subject to $O(L)$.

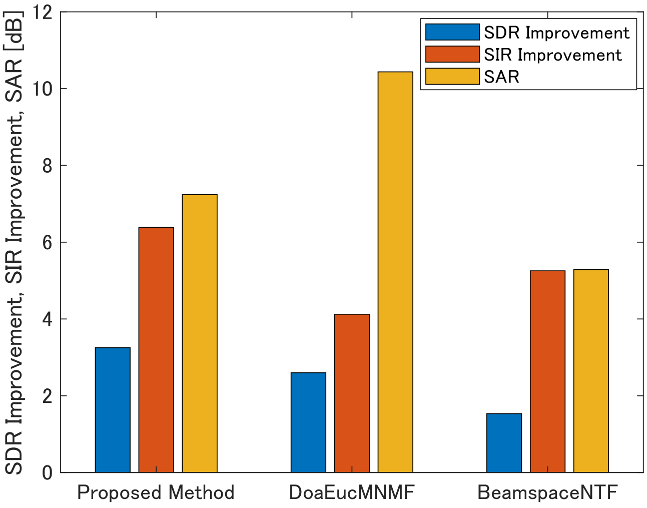

Fig. 6. Comparison of different methods in real environment.

H. Experiment Using Room Impulse Response

Finally, experiments were performed using the measured impulse responses. The METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset v1.0 [48] containing a large set of impulse responses recorded by the em32 Eigenmike was used. The room where the recording was conducted had a reverberation of $T60 = 1.12$ s, and its shape was rectangular ($6.5 \times 8.3 \times 2.9$ m) (see [2] for a detailed description of the dataset). Impulse responses from various source positions were chosen randomly from 240 candidates. To compare the proposed method with other methods, we chose DoaEucMNMF and BeamspaceNTF as in Section V-C. Fig. 6 shows the results of the SDR improvement, SIR improvement, and SAR. Similar to the experimental results in Section V-C, the proposed method showed the best separation performance in terms of SDR improvement and SIR improvement while DoaEucNMF showed the best result in terms of SAR. The results show the effectiveness of the proposed method in separating sources in real space.

VI. CONCLUSION

In this work, we devised an efficient NTF-based blind source separation method in the SH domain that can be used in practice.

TABLE V
SEPARATION PERFORMANCE AND COMPUTATION TIME FOR DIFFERENT NUMBER OF MICROPHONES

Number of mics.	SH order	SDR improvement [dB]	SIR improvement [dB]	SAR [dB]	time [s]
12	2	3.16	7.55	6.06	16.76
20	3	5.05	8.48	8.19	21.05
32	4	5.23	8.60	8.61	30.94

Our approach assumes not only far-field sources but also near-field sources by exploiting the characteristics of the spherical Hankel function, i.e., high power in the low-frequency regime. We also proposed a masking scheme to exclude a noisy region as well as the high-power region from the NTF cost function to make the model parameters less sensitive to these regions. In the evaluations, our model was confirmed to be effective in improving the separation quality in near-field scenarios, which makes the system easy to use for VR in practical situations where singers or speakers often perform close to microphones. The evaluations also showed that our masking scheme could reduce the computational cost and succeeded in exploiting redundancy for low-rank modeling.

REFERENCES

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio—The new standard for coding of immersive spatial audio," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 770–779, Aug. 2015.

[2] M. Wien, J. M. Boyce, T. Stockhammer, and W. Peng, "Standardization status of immersive video coding," *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 9, no. 1, pp. 5–17, Mar. 2019.

[3] M. Noisternig, T. Musil, A. Sontacchi, and R. Höldrich, "3D binaural sound reproduction using a virtual ambisonic approach," in *Proc. IEEE Int. Symp. Virtual Environ., Hum.-Comput. Interfaces, Meas. Syst.*, 2003, pp. 174–178.

[4] G. Enzner, M. W. Weinert, S. Abeling, J. Batke, and P. Jax, " Advanced system options for binaural rendering of ambisonic format," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 251–255.

[5] T. Magariyachi and Y. Mitsufuji, "Analytic error control methods for efficient rotation in dynamic binaural rendering of ambisonics," *J. Acoust. Soc. Amer.*, vol. 147, no. 1, pp. 218–230, 2020.

[6] J. Meyer and G. W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 1781–1784.

[7] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 1949–1952.

[8] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 713–716, Oct. 2005.

[9] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 702–714, Feb. 2007.

[10] J. Meyer and G. W. Elko, *Spherical Microphone Arrays for 3D Sound Recording*. Berlin, Germany: Springer, 2004.

[11] B. Rafaely, "Spatial sampling and beamforming for spherical microphone arrays," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, 2008, pp. 5–8.

[12] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Verlag Berlin Heidelberg: Springer, 2007.

[13] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio,Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011.

[14] N. Epain and C. T. Jin, "Independent component analysis using spherical microphone arrays," in *Acta Acustica United With Acustica*, vol. 98, no. 1, pp. 91–102, 2012.

[15] J. Nikunen and A. Politis, "Multichannel NMF for source separation with ambisonic signals," in *Proc. Int. Workshop Acoust. Sig. Enhancement*, 2018, pp. 251–255.

[16] M. Hafsati, N. Epain, R. Gribonval, and N. Bertin, "Sound source separation in the higher order ambisonics domain," in *Proc. Int. Conf. Digit. Audio Effects DAFx'19*, 2019, pp. 1–7.

[17] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 727–739, Mar. 2014.

[18] S. Moreau and J. Daniel, "Study of higher order ambisonic microphone," in *Proc. Congrès Français d'Acoustique (Joint Congr. CFA-DAGA'04)*, 2004.

[19] S. Bertet, J. Daniel, and S. Moreau, "3D sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone," in *Proc. Audio Eng. Soc. Conv.*, 2006.

[20] Y. Salaün *et al.*, "The flexible audio source separation toolbox version 2.0," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2014. [Online]. Available: https://hal.inria.fr/hal-00957412

[21] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Hoboken, NJ, USA: Wiley, 2009.

[22] Y. Mitsufuji, S. Koyama, and H. Saruwatari, "Multichannel blind source separation based on non-negative tensor factorization in wavenumber domain," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 56–60.

[23] 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, 2019.

[24] S. Lee, S. H. Park, and K. Sung, "Beamspace-domain multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Process. Lett.*, vol. 19, no. 1, pp. 43–46, Jan. 2012.

[25] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. Eur. Signal Process. Conf.*, 2016, pp. 1153–1157.

[26] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 81–85.

[27] E. G. Williams, *Fourier Acoust.: Sound Radiation and Nearfield Acoustical Holography*. New York, NY, USA: Academic Press, 1999.

[28] S. Moreau, J. Daniel, and S. Bertet, "3D sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone," in *Proc. Audio Eng. Soc. Conv.*, 2006, pp. 20–23.

[29] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin, Germany: Springer, 2015.

[30] I. Balmages and B. Rafaely, "Open-sphere designs for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 727–732, Feb. 2007.

[31] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 647–658, Mar. 2014.

[32] H. Chen, T. D. Abhayapala, and W. Zhang, "Theory and design of compact hybrid microphone arrays on two-dimensional planes for three-dimensional soundfield analysis," *J. Acoust. Soc.* America, vol. 138, no. 5, pp. 3081–3092, 2015.

[33] W.-H. Liao, Y. Mitsufuji, K. Osako, and K. Ohkuri, "Microphone array geometry for two dimensional broadband sound field recording," in *Proc. Audio Eng. Soc. Conv.*, 2018.

[34] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2542–2556, Jun. 2007.

[35] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, Sep. 2001.

[36] B. Kulis, M. A. Sustik, and I. S. Dhillon, "Low-rank kernel learning with bregman matrix divergences," *J. Mach. Learn. Res.*, vol. 10, pp. 341–376, 2009.

[37] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Multichannel extensions of non-negative matrix factorization with complex-valued data,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[38] J. Ahrens and S. Spors, “Spatial encoding and decoding of focused virtual sound sources,” in *Proc. Ambisonics Symp.*, 2009.

[39] D. R. Hunter and K. Lange, “A tutorial on MM algorithms,” *Amer. Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[40] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ, USA: Prentice Hall PTR, 1993.

[41] X. Guo, S. Uhlich, and Y. Mitsufuji, “ NMF-based blind source separation using a linear predictive coding error clustering criterion,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 261–265.

[42] M. Spiertz and V. Gnann, “Source-filter based clustering for monaural blind source separation,” in *Proc. Int. Conf. Digit. Audio Effects DAFx’09*, 2009.

[43] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, “Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 654–669, Apr. 2015.

[44] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

[45] E. Vincent, “Improved perceptual metrics for the evaluation of audio source separation,” in *Proc. Latent Variable Anal. Sig. Separation*, 2012, pp. 430–437.

[46] B. Rafaely, “Analysis and design of spherical microphone arrays,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[47] S. Brown and D. Sen, “Error analysis of spherical harmonic soundfield representations in terms of truncation and aliasing errors,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 360–364.

[48] O. Olgun and H. Hacihabiboglu, “METU SPARG eigenmike em32 acoustic impulse response dataset v0.1.0,” Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2635758

Yuki Mitsufuji (Member, IEEE) received the B.S. and M.S. degrees in information science from Keio University, Tokyo, Japan, in 2002 and 2004, respectively. He is currently working toward the Ph.D. degree with the University of Tokyo. He is a Deputy General Manager with Speech and Music Group, Sony Corporation, Tokyo, Japan. He has been leading teams that developed the sound design for the PlayStation game title called “Gran Turismo Sport,” and spatial audio solution called “Sonic Surf VR.” In 2004, he joined Audio Technology Development Group, Sony Corporation. From 2011 to 2012, he was a Visiting Researcher with Analysis/Synthesis Team, Institut de Rechereche et Coordination Acoustique/Musique (IRCAM), Paris, France. He was involved in the 3DTV content search project sponsored by European Project FP7, in research collaboration with IRCAM. He is a Reviewer with ICASSP, INTERSPEECH, etc. He currently became a General Chair of Signal Separation Evaluation Campaign where his team had scored the best results for three consecutive years. He has numerous granted patents for audio signal processing.

Norihiro Takamune received the B.E. degree in engineering, and the M.S. degree in information science and technology from the University of Tokyo, Japan, in 2012, and 2015, respectively. He is currently a Researcher with the University of Tokyo, Japan. His research interests include music information analysis, audio source separation and machine learning.

Shoichi Koyama (Member, IEEE) received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Japan, in 2007, 2009, and 2014, respectively. He joined Nippon Telegraph and Telephone Corporation in 2009 as a Researcher in acoustic signal processing. He moved to the University of Tokyo in 2014 and has been an Assistant Professor (Lecturer), since 2018. From 2016 to 2018, he was also a Visiting Researcher with Paris Diderot University (Paris7), Institut Langevin, Paris, France. His research interests include acoustic inverse problems, sound field analysis and synthesis, and spatial audio.

Hiroshi Saruwatari (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM IS Laboratory, Japan, in 1993, and Nara Institute of Science and Technology, Japan, in 2000. From 2014, he is currently a Professor with The University of Tokyo, Japan. His research interests include statistical speech signal processing, blind source separation (BSS), audio enhancement, and robot audition. He has successfully achieved his carrier, especially on BSS researches, and put his research into the world’s first commercially available Independent-Component-Analysis-based BSS microphone in 2007. He was the recipient of the Paper Awards from IEICE in 2001 and 2006, from TAF in 2004, 2009 and 2012, from IEEE-IROS2005 in 2006, and from APSIPA in 2013 and 2018. He received DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, The Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hattori-Hoko Award in 2018. He won the first prize in IEEE MLSP2007 BSS Competition. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ, including chair posts of international conferences and Associate Editor of journals.