Sparse Representation of a Spatial Sound Field in a Reverberant Environment

Shoichi Koyama, *Member, IEEE*, and Laurent Daudet, *Senior Member, IEEE*

***Abstract*—This paper investigates sound-field modeling in a realistic reverberant setting. Starting from a few point-like microphone measurements, the goal is to estimate the direct source field within a whole three-dimensional (3-D) space around these microphones. Previous sparse sound field decompositions assumed only a spatial sparsity of the source distribution, but could generally not handle reverberation. We here add an explicit model of the reverberant sound field, that has two components: the first component sparse in the plane-wave domain, the other component low-rank as a multiplication of transfer functions and source signals. We derive the corresponding decomposition algorithm based on the alternating direction method of multipliers. We furthermore provide empirical rules for tuning the two parameters to be set in the algorithm. Numerical and experimental results indicate that the decomposition and reconstruction performances are significantly improved, in the case of reverberant environments.**

***Index Terms*—Sound field decomposition, sparse representation, sound field recording, source identification, reverberation.**

I. Introduction

THE estimation of a space-varying acoustic field, inside a target region, by interpolating from a discrete set of point measurements by microphones, is referred to as *sound field estimation*, which is a fundamental inverse problem in acoustics. Such estimation can be applied to various acoustic measurement tasks, e.g., visualization of an acoustic field [1]–[3], identification of sound sources [4]–[7], and capturing a sound field for reproduction using loudspeakers or headphones [8]–[11].

When the region to be estimated does not include any sources, the sound field can be reconstructed by representing it as a linear combination of element solutions of the homogeneous Helmholtz equation, such as plane waves or harmonic functions, i.e., *sound field decomposition*. This decomposition-based strategy is well-known as a theoretical foundation of Fourier acoustics and near-field acoustic holography [12]. It is also possible to apply the equivalent source method [13], [14], which is based on the representation by using Green's functions on the fictitious boundary enclosing the target region. On the contrary, the sound field estimation inside a region including sources is an ill-posed problem. This requires some assumptions on the sources because the target sound field is now governed by the inhomogeneous Helmholtz equation that contains an arbitrary function of the source distribution. Even for the estimation of the homogeneous field, the sound field decomposition only by plane waves or harmonic functions suffers from errors originating from the discrete sampling of the target field by microphones, which is referred to as spatial aliasing artifacts.

Owing to the recent development of sparse decomposition algorithms in the context of compressed sensing [15], the sparse representation of acoustic fields has been proved to be an effective regularization framework in several applications, such as acoustic holography [16], source localization [17], [18], estimation of room transfer functions [19], and sound field recording [20], [21]. In these models, it is assumed that the sound field can be well approximated in a low-dimensional subspace, that can be estimated from a small number of incoherent measurements. In practice, such an assumption improves the spatial resolution on the acoustic measurement and estimation.

Sparse sound field decompositions are typically based on the spatial sparsity of the sound sources. In the far-field, the corresponding sound field is approximated by a small number of plane waves. This assumption approximately holds when the region of interest does not include any sources and the array aperture of microphones is relatively small. Besides, based on the Vekua's theory [22], any homogeneous sound field in a bounded convex region can be well approximated by a limited number of plane waves with some guarantees on the approximation quality.

In this study, we focus on sound field estimation in the near-field, where the region of interest includes sound sources and the array aperture of microphones is relatively large, but also in a possibly reverberant setting, which is a typical scenario in sound field recording for reproduction [9], [10]. In [23], the near-field sources are assumed to be sparsely distributed inside the predefined *source region*. The fundamental solutions inside the source region, i.e., Green's functions, are used as the basis functions for decomposition, in order to approximately represent the solution of the inhomogeneous Helmholtz equation. By using sparse decomposition algorithms, the target sound field is decomposed into a sum of its particular and homogeneous solutions, which can be regarded as the direct and reverberant components, respectively. For many of the applications discussed

Manuscript received June 29, 2018; revised December 14, 2018 and February 11, 2019; accepted February 16, 2019. Date of publication February 22, 2019; date of current version April 11, 2019. This work was supported in part by JSPS KAKENHI under Grant JP15H05312 and in part by the LABEX WIFI (Laboratory of Excellence within the French Program "Investments for the Future") under reference ANR- 10-IDEX-0001-02 PSL*. The guest editor coordinating the review of this paper and approving it for publication was Prof. Walter Kellermann. *(Corresponding author: Shoichi Koyama.)*

S. Koyama was with the Institut Langevin, Paris Diderot University, Paris 75005, France. He is now with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan (e-mail: koyama.shoichi@ieee.org).

L. Daudet is with the Institut Langevin, ESPCI Paris, PSL University, Paris Diderot University, CNRS UMR 7587, Paris 75005, France (e-mail: laurent.daudet@espci.fr).

Digital Object Identifier 10.1109/JSTSP.2019.2901127

above, an accurate estimation of the direct sound field is needed. In weakly reverberating environments, or in the case of a diffuse sound field, the reverberant components can be assumed spatially uncorrelated, and hence easily separated. However, this assumption is not valid in a highly reverberant environment, when strong early reflections exist outside the source region, and this may significantly impact the decomposition accuracy. To overcome this problem, the reverberant component has to be explicitly modeled.

Here, we introduce a general model for sound field decomposition in a reverberant environment, whose preliminary results were presented in [24]. It models the sound field as the sum of two components: the direct source field and the reverberant field. The direct source component is assumed to follow a spatially sparse source distribution, as in [23]. We model the reverberant component as the sum of sparse in the plane-wave domain and low-rank matrices. A convex relaxation of the problem can be derived, that brings an efficient decomposition algorithm based on the alternating direction method of multipliers (ADMM) [25], [26].

This paper is organized as follows: The sound field model and its sparse decomposition in current methods are given in Sect. II. In Sect. III, the proposed method for decomposing a reverberant sound field is described. Numerical simulations and experimental results are reported in Sect. IV. Finally, Sect. V concludes this paper.

A. Notation

Italic letters denote scalars, lower-case boldface letters denote vectors, and upper-case boldface letters denote tensors of two or more orders including matrices. Real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively.

Subscripts of scalars, vectors, and tensors stand for their indexes. For example, x_{ij} is the (i,j)th entry of matrix $\mathbf{X}$, and $\mathbf{x}_{ij}$ and $\mathbf{X}_i$ are the (i,j)th vector and ith matrix extracted from a third-order tensor $\mathbf{X}$, respectively.

The complex conjugate, conjugate transpose, and inverse are denoted by superscripts $(\cdot)^*$, $(\cdot)^{\mathsf{H}}$, and $(\cdot)^{-1}$, respectively. The absolute value of a scalar x is denoted as $|x|$. The ℓ_p-norm of a vector $\mathbf{x}$ is denoted as $\|\mathbf{x}\|_p$. The Frobenius norm of a matrix $\mathbf{X}$ is denoted as $\|\mathbf{X}\|_F$. Superscript $(\cdot)^{(i)}$ stands for the ith iteration.

II. Sound Field Model and Its Sparse Decomposition

A. Representation of Inhomogeneous Sound Field

We firstly formulate our sound field model, which is also described in [21], [23]. As shown in Fig. 1, we assume that a sound field consists of near-field sound sources inside a source region Ω. Sound pressures are measured inside Ω and/or its peripheral area using multiple microphones. By denoting the sound pressure of the angular frequency ω at position $\mathbf{r}$ as $u(\mathbf{r},\omega)$, $u(\mathbf{r},\omega)$ satisfies the following inhomogeneous Helmholtz equation:

$$\left(\nabla^2 + k^2\right) u(\mathbf{r},\omega) = -Q(\mathbf{r},\omega), \tag{1}$$

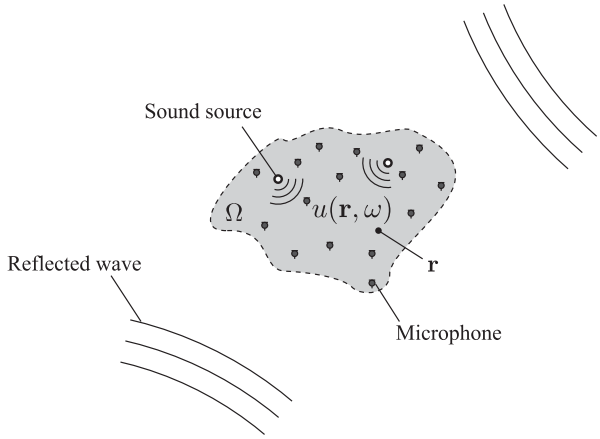


Fig. 1. Sound field inside a region Ω including sources. The direct sound field for continuous $\mathbf{r}$ is to be estimated from the measurements by microphones.

with unknown boundary condition on the room surface. Here, $Q(\mathbf{r},\omega)$ is the distribution of the sources inside Ω and $k = \omega/c$ is the wave number when the sound velocity is c. Note that the boundary condition is imposed on an unknown room surface that is independent of Ω. We hereafter omit ω for notational simplicity. The solution $u(\mathbf{r})$ of (1) becomes the sum of the particular and homogeneous solutions, $u_{\mathrm{P}}(\mathbf{r})$ and $u_{\mathrm{H}}(\mathbf{r})$, respectively,

$$u(\mathbf{r}) = u_{\mathrm{P}}(\mathbf{r}) + u_{\mathrm{H}}(\mathbf{r}). \tag{2}$$

The particular solution $u_{\mathrm{P}}(\mathbf{r})$ can be obtained by convolution of $Q(\mathbf{r})$ with the three-dimensional free-field Green's function $G(\mathbf{r}|\mathbf{r}')$ inside Ω as

$$u_{\mathrm{P}}(\mathbf{r}) = \int_{\mathbf{r}'\in\Omega} Q(\mathbf{r}')G(\mathbf{r}|\mathbf{r}')d\mathbf{r}'. \tag{3}$$

Here, the free-field Green's function $G(\mathbf{r}|\mathbf{r}')$ is defined as

$$G(\mathbf{r}|\mathbf{r}') = \frac{e^{jk\|\mathbf{r}-\mathbf{r}'\|_2}}{4\pi\|\mathbf{r}-\mathbf{r}'\|_2}, \tag{4}$$

where $G(\mathbf{r}|\mathbf{r}')$ corresponds to the transfer function of the monopole between $\mathbf{r}$ and $\mathbf{r}'$. The homogeneous solution $u_{\mathrm{H}}(\mathbf{r})$ is determined by $u_{\mathrm{P}}(\mathbf{r})$ and $u(\mathbf{r})$ satisfying the boundary condition on the room surface, which can be represented as a linear combination of plane-wave or harmonic functions [12]. Finally, our sound field model can be represented as

$$u(\mathbf{r}) = \int_{\mathbf{r}'\in\Omega} Q(\mathbf{r}')G(\mathbf{r}|\mathbf{r}')d\mathbf{r}' + u_{\mathrm{H}}(\mathbf{r}). \tag{5}$$

It is assumed that M microphones are placed at $\mathbf{r}_m$ ($m \in \{1,\ldots,M\}$). Our objective is to estimate $u_{\mathrm{P}}(\mathbf{r})$ for continuous $\mathbf{r}$ as well as $Q(\mathbf{r})$ from the discrete set of pressure measurements $u(\mathbf{r}_m)$.

B. Sparse Sound Field Decomposition for Estimation of Inhomogeneous Field

We here assume that the source distribution is spatially sparse, to solve (5). First, we discretize the region Ω into a set of small regions Ω_n ($n \in \{1,\ldots,N\}$). The representative point of Ω_n is

defined as a grid point and its location is denoted as $\mathbf{r}_n$. Then, $u_{\mathrm{P}}(\mathbf{r})$ is approximated as

$$u_{\mathrm{P}}(\mathbf{r}) = \sum_{n=1}^{N} \int_{\mathbf{r}' \in \Omega_n} Q(\mathbf{r}') G(\mathbf{r}|\mathbf{r}') \approx \sum_{n=1}^{N} G(\mathbf{r}|\mathbf{r}_n) \int_{\mathbf{r}' \in \Omega_n} Q(\mathbf{r}') d\mathbf{r}'. \quad (6)$$

The signals received by the microphones are converted into the time-frequency domain, for example, by using short-time Fourier transform (STFT). We denote the signals at tth time frame ($t \in \{1, \ldots, T\}$) and fth frequency bin ($f \in \{1, \ldots, F\}$) as $\mathbf{y}_{t,f} \in \mathbb{C}^M$, whose elements consist of $u(\mathbf{r}_m)$. The dictionary matrix consisting of $G(\mathbf{r}_m|\mathbf{r}_n)$, the vector of the source distribution inside Ω_n, and the vector of the homogeneous term $u_{\mathrm{H}}(\mathbf{r}_m)$ are also defined as $\mathbf{D}_f \in \mathbb{C}^{M \times N}$, $\mathbf{x}_{t,f} \in \mathbb{C}^N$, and $\mathbf{z}_{t,f} \in \mathbb{C}^M$, respectively. Thus, (5) is represented in a matrix form as

$$\mathbf{y}_{t,f} = \mathbf{D}_f \mathbf{x}_{t,f} + \mathbf{z}_{t,f}. \quad (7)$$

We also define third-order tensors $\mathbf{Y} \in \mathbb{C}^{M \times T \times F}$, $\mathbf{X} \in \mathbb{C}^{N \times T \times F}$, and $\mathbf{Z} \in \mathbb{C}^{M \times T \times F}$ by collecting $\mathbf{y}_{t,f}$, $\mathbf{x}_{t,f}$, and $\mathbf{z}_{t,f}$ for all t and f. Therefore, $\mathbf{X}$ needs to be estimated from $\mathbf{Y}$ given $\mathbf{D}_f$. Note that (7) is an underdetermined linear equation because N is assumed to be much larger than M.

Under the assumption that the source distribution is spatially sparse, $\mathbf{x}_{t,f}$ will have few non-zero elements. Such a solution can be obtained by solving the following optimization problem [27]:

$$\underset{\mathbf{x}_{t,f}}{\text{minimize}} \frac{1}{2} \|\mathbf{y}_{t,f} - \mathbf{D}_f \mathbf{x}_{t,f}\|_2^2 + \lambda \|\mathbf{x}_{t,f}\|_p^p, \quad (8)$$

where $\|\cdot\|_p$ represents the ℓ_p-norm and $0 < p \leq 1$. A number of algorithms to solve (8) have been investigated in the last decades, particularly in the context of compressed sensing [15]. By setting $p = 1$, (8) becomes a convex problem and various convex optimization algorithms can be applied [27]. When $p < 1$, iteratively reweighted least-squares algorithms are typically used [18], [28]. We here focus on the case of $p = 1$. In addition, it can be assumed that the indexes of the non-zero values of $\mathbf{x}_{t,f}$ are consistent for t and f when the sources are static and the source signals are broadband. This group-sparse assumption can be incorporated to increase the robustness of the sparse decomposition by using $\ell_{1,2}$-norm penalty term as [23], [29]

$$\underset{\mathbf{X}}{\text{minimize}} \frac{1}{2} \sum_{t=1}^{T} \sum_{f=1}^{F} \|\mathbf{y}_{t,f} - \mathbf{D}_f \mathbf{x}_{t,f}\|_2^2 + \lambda \|\mathbf{X}\|_{1,2}, \quad (9)$$

where λ is the constant balancing parameter and the $\ell_{1,2}$-norm $\|\cdot\|_{1,2}$ is defined as

$$\|\mathbf{X}\|_{1,2} = \sum_{n=1}^{N} \sqrt{\sum_{t=1}^{T} \sum_{f=1}^{F} |x_{n,t,f}|^2}. \quad (10)$$

Note that the $\ell_{1,2}$-norm is defined for the third-order tensor by generalizing its standard definition. Again, various convex optimization algorithms can be applied to solve (9).

Algorithm 1: Accelerated Proximal Gradient Method for (9).

Initialize $\mathbf{\Phi}^{(1)}$ ($=\mathbf{X}^{(1)}$) and set $i = 1$ and $s^{(1)} = 1$.
while (13) and (14) are satisfied or i reaches predefined maximum value **do**
  $\mathbf{X}^{(i+1)} \leftarrow \mathcal{T}^{1,2}_{\lambda/\eta_D}\left(\boldsymbol{\phi}^{(i)}_{t,f} - \frac{1}{\eta_D}\mathbf{D}_f^{\mathsf{H}}(\mathbf{D}_f \boldsymbol{\phi}^{(i)}_{t,f} - \mathbf{y}_{t,f})\right)$
  $s^{(i+1)} \leftarrow \left(1 + \sqrt{1 + 4(s^{(i)})^2}\right)/2$
  $\mathbf{\Phi}^{(i+1)} \leftarrow \mathbf{X}^{(i+1)} + \frac{s^{(i)}-1}{s^{(i+1)}}(\mathbf{X}^{(i+1)} - \mathbf{X}^{(i)})$
  $i \leftarrow i + 1$
end while

For example, in the proximal gradient method [30], the solution of (9) is obtained by iteratively updating $\mathbf{X}$ as

$$\mathbf{X}^{(i+1)} = \arg\min_{\mathbf{X}} \lambda \|\mathbf{X}\|_{1,2} + \frac{\eta_D}{2} \sum_{t=1}^{T} \sum_{f=1}^{F} \|\mathbf{x}_{t,f} - \left(\mathbf{x}^{(i)}_{t,f} - \frac{1}{\eta_D}\mathbf{D}_f^{\mathsf{H}}\left(\mathbf{D}_f \mathbf{x}^{(i)}_{t,f} - \mathbf{y}_{t,f}\right)\right)\|_2^2 = \mathcal{T}^{1,2}_{\lambda/\eta_D}\left(\mathbf{x}^{(i)}_{t,f} - \frac{1}{\eta_D}\mathbf{D}_f^{\mathsf{H}}\left(\mathbf{D}_f \mathbf{x}^{(i)}_{t,f} - \mathbf{y}_{t,f}\right)\right), \quad (11)$$

where $\mathcal{T}^{1,2}_{\alpha}(\cdot)$ is the soft-thresholding operator defined as

$$\left\{\mathcal{T}^{1,2}_{\alpha}(\mathbf{A})\right\}_{n,t,f} = \max\left(1 - \frac{\alpha}{\|\mathbf{A}_n\|_F}, 0\right) a_{n,t,f}. \quad (12)$$

The update rule (11) is obtained by linearization of the first term of (9) around $\mathbf{x}^{(i)}_{t,f}$ and applying proximal operator for the $\ell_{1,2}$-norm [31]. Here, η_D is a constant parameter that should satisfy $\eta_D > \sigma^2_{\max}(\mathbf{D}_f)$, where $\sigma^2_{\max}(\cdot)$ represents the maximum eigenvalue. In addition, the convergence rate can be improved by acceleration using $s^{(i)}$ in the fifth line of Algorithm 1 [30], [32]. The stopping rule of this algorithm is obtained based on Karush–Kuhn–Tucker (KKT) condition [26] as

$$\sum_{f=1}^{F} \|\mathbf{D}_f \mathbf{X}_f - \mathbf{Y}_f\|_F \Big/ \|\mathbf{Y}\|_F \leq \xi_1 \quad (13)$$

$$\eta_D \left\|\mathbf{X}^{(i+1)} - \mathbf{X}^{(i)}\right\|_F \Big/ \lambda \|\mathbf{Y}\|_F \leq \xi_2, \quad (14)$$

where ξ_1 and ξ_2 are the sufficiently small constants. The maximum number of iterations can also be set.

III. Sparse Sound Field Decomposition in Reverberant Environment

In (9), the reverberant component $\mathbf{z}_{t,f}$ is treated as a small residual, assuming a complex Gaussian distribution. However, this does not hold in a reverberant environment, which deteriorates the estimate of the direct component $\mathbf{x}_{t,f}$. For a more accurate and robust decomposition, one needs an explicit model of the reverberant component - such a model is the main contribution of this study. Here, we assume that the reverberant component $\mathbf{z}_{t,f}$ is the sum of two components: the first part is sparse in the plane-wave domain and the second part is low-rank.

A. Reverberation Modeling

Referring to Vekua's theory [22], any homogeneous sound field in a bounded convex region can be well approximated by a limited number of plane waves, which is successfully applied in various contexts [16], [19]. The homogeneous term $u_{\mathrm{H}}(\mathbf{r})$ is represented by the linear combination of L plane wave functions as

$$u_{\mathrm{H}}(\mathbf{r}) \approx \sum_{l=1}^{L} \varphi_l e^{j\mathbf{k}_l^{\mathsf{T}}\mathbf{r}}, \tag{15}$$

where l ($\in \{1, \ldots, L\}$) is the index of plane wave, $\mathbf{k}_l$ is the wave vector of the lth plane wave, and φ_l is its weight coefficient. When L is sufficiently large, most elements of φ_l can be approximated as zero. We denote the dictionary matrix of the plane wave functions $e^{j\mathbf{k}_l^{\mathsf{T}}\mathbf{r}_m}$ and the vector of φ_l as $\mathbf{W}_f \in \mathbb{C}^{M\times L}$ and $\mathbf{u}_{t,f} \in \mathbb{C}^{L}$, respectively. Then, $\mathbf{z}_{t,f}$ is represented as

$$\mathbf{z}_{t,f} = \mathbf{W}_f \mathbf{u}_{t,f}, \tag{16}$$

where the limited number of elements of $\mathbf{u}_{t,f}$ will have nonzero values. When the sound field is static within T time frames, same plane waves can be used to represent $\mathbf{z}_{t,f}$ for all T time frames and their weights only change. It is also assumed that the same subset of plane wave directions can be used for all F frequency bins because the set of plane waves for approximating a sound field at high frequencies is sufficient to approximate that at low frequencies. Thus, by denoting $\mathbf{U} \in \mathbb{C}^{M\times L\times F}$ consisting of $\mathbf{u}_{t,f}$, such a solution can be obtained by solving the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{U}}{\text{minimize}}\ \|\mathbf{X}\|_{1,2} + \mu\|\mathbf{U}\|_{1,2} \\ &\text{subject to}\ \ \mathbf{y}_{t,f} = \mathbf{D}_f\mathbf{x}_{t,f} + \mathbf{W}_f\mathbf{u}_{t,f}, \end{aligned} \tag{17}$$

where μ is the balancing parameter.

In a different perspective, the reverberant component can be approximated as the multiplication of the source signals and their transfer functions excluding the direct path [33]. When the source signal and transfer function excluding the direct path of the jth source ($j \in \{1, \ldots, J\}$) are defined as s_j and $h_j(\mathbf{r})$, respectively, $u_{\mathrm{H}}(\mathbf{r})$ is represented as

$$u_{\mathrm{H}}(\mathbf{r}) \approx \sum_{j=1}^{J} h_j(\mathbf{r}) s_j. \tag{18}$$

We denote the transfer function matrix consisting of $h_j(\mathbf{r})$ and the vector of the source signal s_j as $\mathbf{H}_f \in \mathbb{C}^{M\times J}$ and $\mathbf{s}_{t,f} \in \mathbb{C}^{J}$, respectively. Then, $\mathbf{z}_{t,f}$ is represented as

$$\mathbf{z}_{t,f} = \mathbf{H}_f \mathbf{s}_{t,f}. \tag{19}$$

We assume that the transfer function $h_j(\mathbf{r})$ is static within T time frames. Then, by using $\mathbf{S}_f \in \mathbb{C}^{J\times T}$ consisting of $\mathbf{s}_{t,f}$, $\mathbf{Z}_f = [\mathbf{z}_{1,f}, \ldots, \mathbf{z}_{T,f}]$ is represented as

$$\mathbf{Z}_f = \mathbf{H}_f \mathbf{S}_f. \tag{20}$$

Since the spatial sparsity of the source distribution $Q(\mathbf{r})$ is assumed, J will be sufficiently small compared to the number of microphones. Therefore, the rank of $\mathbf{Z}_f$ will approximately correspond to the number of sources J, which leads to the low-rank spatial covariance matrix $\mathbf{Z}_f\mathbf{Z}_f^{\mathsf{H}}$ as typically used in the literature of array processing [34]. By newly defining $\mathbf{V} \in \mathbb{C}^{M\times T\times F}$ consisting of $\mathbf{v}_{t,f} = \mathbf{H}_f\mathbf{s}_{t,f}$, such a solution can be obtained by solving the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{V}}{\text{minimize}}\ \|\mathbf{X}\|_{1,2} + \nu\sum_{f=1}^{F}\|\mathbf{V}_f\|_{*} \\ &\text{subject to}\ \ \mathbf{y}_{t,f} = \mathbf{D}_f\mathbf{x}_{t,f} + \mathbf{v}_{t,f}, \end{aligned} \tag{21}$$

where $\|\cdot\|_{*}$ represents the nuclear norm, which is the tightest convex lower bound of the rank function [35], and ν is the balancing parameter. We use the same ν for all the frequency bins to simplify the problem. This low-rank assumption will hold when both the number of sources J and the length of the transfer function $h_j(\mathbf{r})$ are not very large.

In the first model, although the set of plane waves can well approximate a homogeneous sound field, the optimal number of plane waves for the estimation of $\mathbf{X}$ with separating $\mathbf{U}$ depends on the shape and size of the target region Ω as well as the angular frequency ω [22]. Moreover, it is empirically known that the plane-wave approximation of the homogeneous field is sensitive to the setting of the balancing parameter μ. In the second model, although the low-rank matrix is more flexible and can represent any static transfer functions within the time frame, the separation of $\mathbf{X}$ and $\mathbf{V}$ is not a trivial task because the direct component also has a low-rank structure. Besides, the constant balancing parameter ν for all the frequency bins limits its flexibility. To compensate for their drawbacks, we propose a hybrid model of $\mathbf{U}$ and $\mathbf{V}$. Therefore, $\mathbf{z}_{t,f}$ is represented as

$$\mathbf{z}_{t,f} = \mathbf{W}_f\mathbf{u}_{t,f} + \mathbf{v}_{t,f}. \tag{22}$$

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{U},\mathbf{V}}{\text{minimize}}\ \|\mathbf{X}\|_{1,2} + \mu\|\mathbf{U}\|_{1,2} + \nu\sum_{f=1}^{F}\|\mathbf{V}_f\|_{*} \\ &\text{subject to}\ \ \mathbf{y}_{t,f} = \mathbf{D}_f\mathbf{x}_{t,f} + \mathbf{W}_f\mathbf{u}_{t,f} + \mathbf{v}_{t,f}. \end{aligned} \tag{23}$$

This proposed model is intended that the sparse plane waves $\mathbf{U}$ and low-rank matrices $\mathbf{V}$ produce a complementary effect to represent the reverberant component to estimate $\mathbf{X}$ with separating $\mathbf{Z}$ from $\mathbf{Y}$.

B. ADMM for Sparse Sound Field Decomposition

Since the optimization problem (23) is convex, there are several choices to solve (23). We use ADMM because of its computational efficiency and its flexibility.

First, we define the augumented Lagrangian function $\mathcal{L}$ as

$$\begin{aligned} \mathcal{L}(\mathbf{X},\mathbf{U},\mathbf{V},\boldsymbol{\Theta}) &= \|\mathbf{X}\|_{1,2} + \mu\|\mathbf{U}\|_{1,2} + \nu\sum_{f=1}^{F}\|\mathbf{V}_f\|_{*} \\ &+ \sum_{t=1}^{T}\sum_{f=1}^{F}\langle\boldsymbol{\theta}_{t,f}, \mathbf{D}_f\mathbf{x}_{t,f} + \mathbf{W}_f\mathbf{u}_{t,f} + \mathbf{v}_{t,f} - \mathbf{y}_{t,f}\rangle \\ &+ \frac{1}{2\rho}\sum_{t=1}^{T}\sum_{f=1}^{F}\|\mathbf{D}_f\mathbf{x}_{t,f} + \mathbf{W}_f\mathbf{u}_{t,f} + \mathbf{v}_{t,f} - \mathbf{y}_{t,f}\|_2^2, \end{aligned} \tag{24}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product, $\mathbf{\Theta} \in \mathbb{C}^{M \times T \times F}$ is the Lagrangian multiplier, and $\rho > 0$ is a constant parameter. In ADMM, each variable is alternately updated, starting with arbitrary initial values as

$$\begin{cases} \mathbf{X}^{(i+1)} = \arg\min_{\mathbf{X}} \mathcal{L}\left(\mathbf{X}, \mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{\Theta}^{(i)}\right) \\ \mathbf{U}^{(i+1)} = \arg\min_{\mathbf{U}} \mathcal{L}\left(\mathbf{X}^{(i+1)}, \mathbf{U}, \mathbf{V}^{(i)}, \mathbf{\Theta}^{(i)}\right) \\ \mathbf{V}^{(i+1)} = \arg\min_{\mathbf{V}} \mathcal{L}\left(\mathbf{X}^{(i+1)}, \mathbf{U}^{(i+1)}, \mathbf{V}, \mathbf{\Theta}^{(i)}\right) \\ \boldsymbol{\theta}_{t,f}^{(i+1)} = \boldsymbol{\theta}_{t,f}^{(i)} \\ \quad + \left(\mathbf{D}_f \mathbf{x}_{t,f}^{(i+1)} + \mathbf{W}_f \mathbf{u}_{t,f}^{(i+1)} + \mathbf{v}_{t,f}^{(i+1)} - \mathbf{y}_{t,f}\right)/\rho \end{cases},$$

where (i) is the iteration index. The Lagrangian function for each update is minimized for one variable while fixing the other variables, which can be efficiently solved by using proximal gradient operators [31].

The update of $\mathbf{X}$ can be derived as

$$\begin{aligned} \mathbf{X}^{(i+1)} = & \arg\min_{\mathbf{X}} \|\mathbf{X}\|_{1,2} + \frac{\eta_D}{2\rho} \sum_{t=1}^{T} \sum_{f=1}^{F} \left\| \mathbf{x}_{t,f} - \mathbf{x}_{t,f}^{(i)} \right. \\ & - \frac{\rho}{\eta_D} \mathbf{D}_f^{\mathsf{H}} \left(\boldsymbol{\theta}_{t,f}^{(i)} + \left(\mathbf{D}_f \mathbf{x}_{t,f}^{(i)} + \mathbf{W}_f \mathbf{u}_{t,f}^{(i)} \right. \right. \\ & \left. \left. + \mathbf{v}_{t,f}^{(i)} - \mathbf{y}_{t,f} \right)/\rho \right) \Big\|_2^2 \\ = & \mathcal{T}_{\rho/\eta_D}^{1,2} \left(\mathbf{x}_{t,f}^{(i)} - \frac{\rho}{\eta_D} \mathbf{D}_f^{\mathsf{H}} \left(\boldsymbol{\theta}_{t,f}^{(i)} \right. \right. \\ & \left. \left. + \left(\mathbf{D}_f \mathbf{x}_{t,f}^{(i)} + \mathbf{W}_f \mathbf{u}_{t,f}^{(i)} + \mathbf{v}_{t,f}^{(i)} - \mathbf{y}_{t,f} \right)/\rho \right) \right), \end{aligned} \tag{25}$$

where $\mathcal{T}_{\alpha}^{1,2}(\cdot)$ is defined in (12). Again, this update rule is obtained by linearization of the second term of the first line of (25) at $\mathbf{X}^{(i)}$ with a positive constant parameter η_D. Then, the proximal operator for the $\ell_{1,2}$-norm is applied.

The update rule of $\mathbf{U}$ is obtained by a similar procedure to that of $\mathbf{X}$.

$$\begin{aligned} \mathbf{U}^{(i+1)} = & \arg\min_{\mathbf{U}} \mu \|\mathbf{U}\|_{1,2} + \frac{\eta_W}{2\rho} \sum_{t=1}^{T} \sum_{f=1}^{F} \left\| \mathbf{u}_{t,f} - \mathbf{u}_{t,f}^{(i)} \right. \\ & - \frac{\rho}{\eta_W} \mathbf{W}_f^{\mathsf{H}} \left(\boldsymbol{\theta}_{t,f}^{(i)} + \left(\mathbf{D}_f \mathbf{x}_{t,f}^{(i+1)} + \mathbf{W}_f \mathbf{u}_{t,f}^{(i)} \right. \right. \\ & \left. \left. + \mathbf{v}_{t,f}^{(i)} - \mathbf{y}_{t,f} \right)/\rho \right) \Big\|_2^2 \\ = & \mathcal{T}_{\mu\rho/\eta_W}^{1,2} \left(\mathbf{u}_{t,f}^{(i)} - \frac{\rho}{\eta_W} \mathbf{W}_f^{\mathsf{H}} \left(\boldsymbol{\theta}_{t,f}^{(i)} \right. \right. \\ & \left. \left. + \left(\mathbf{D}_f \mathbf{x}_{t,f}^{(i+1)} + \mathbf{W}_f \mathbf{u}_{t,f}^{(i)} + \mathbf{v}_{t,f}^{(i)} - \mathbf{y}_{t,f} \right)/\rho \right) \right) \end{aligned} \tag{26}$$

The linearization at $\mathbf{Z}^{(i)}$ with a positive constant parameter η_W and the proximal operator for the $\ell_{1,2}$-norm are applied.

Algorithm 2: Proposed Algorithm for Solving (23) Based on ADMM.

Initialize $\mathbf{X}^{(1)}, = \mathbf{U}^{(1)}$, $\mathbf{V}^{(1)}$, and $\mathbf{\Theta}^{(1)}$. Set $i = 1$.
while (30) and (31) are satisfied or i reaches predefined maximum value **do**
  Update $\mathbf{X}^{(i)}$ by calculating (25)
  Update $\mathbf{U}^{(i)}$ by calculating (26)
  Update $\mathbf{V}^{(i)}$ by calculating (27)
  Update $\mathbf{\Theta}^{(i)}$ by

$$\boldsymbol{\theta}_{t,f}^{(i+1)} \leftarrow \boldsymbol{\theta}_{t,f}^{(i)} + \left(\mathbf{D}_f \mathbf{x}_{t,f}^{(i+1)} + \mathbf{W}_f \mathbf{u}_{t,f}^{(i+1)} + \mathbf{v}_{t,f}^{(i+1)} - \mathbf{y}_{t,f} \right)/\rho$$

  $i \leftarrow i + 1$
end while

For the update of $\mathbf{V}_f$, the proximal operator for the nuclear norm can be directly applied as

$$\begin{aligned} \mathbf{V}_f^{(i+1)} = & \arg\min_{\mathbf{V}_f} \nu \|\mathbf{V}_f\|_* + \frac{1}{2\rho} \left\| \mathbf{D}_f \mathbf{X}_f^{(i+1)} + \mathbf{W}_f \mathbf{U}_f^{(i+1)} \right. \\ & \left. + \mathbf{V}_f - \mathbf{Y}_f + \rho \mathbf{\Theta}_f^{(i)} \right\|_2^2 \\ = & \mathcal{T}_{\nu\rho}^{*} \left(\mathbf{Y}_f - \mathbf{D}_f \mathbf{X}_f^{(i+1)} - \mathbf{W}_f \mathbf{U}_f^{(i+1)} - \rho \mathbf{\Theta}_f^{(i)} \right), \end{aligned} \tag{27}$$

where $\mathcal{T}_{\alpha}^{*}(\cdot)$ is defined as

$$\mathcal{T}_{\alpha}^{*}(\mathbf{A}) = \bar{\mathbf{U}} \max\left(\mathbf{\Sigma} - \alpha \mathbf{I}, 0\right) \bar{\mathbf{V}}^{\mathsf{H}}. \tag{28}$$

Here, $\bar{\mathbf{U}}$, $\mathbf{\Sigma}$, and $\bar{\mathbf{V}}$ are obtained by the singular value decomposition of $\mathbf{A}$ as

$$\mathbf{A} = \bar{\mathbf{U}} \mathbf{\Sigma} \bar{\mathbf{V}}^{\mathsf{H}}. \tag{29}$$

The proposed algorithm for solving (23) is summarized in Algorithm 2. The stopping rule of this algorithm can be obtained based on KKT condition [26] as

$$\sum_{f=1}^{F} \|\mathbf{D}_f \mathbf{X}_f + \mathbf{W}_f \mathbf{U}_f + \mathbf{V}_f - \mathbf{Y}_f\|_F \Big/ \|\mathbf{Y}\|_F \leq \epsilon_1 \tag{30}$$

$$\begin{aligned} & \max\left(\sqrt{\eta_D} \left\| \mathbf{X}^{(i+1)} - \mathbf{X}^{(i)} \right\|_F, \sqrt{\eta_W} \left\| \mathbf{U}^{(i+1)} - \mathbf{U}^{(i)} \right\|_F, \right. \\ & \left. \left\| \mathbf{V}^{(i+1)} - \mathbf{V}^{(i)} \right\|_F \right) \Big/ \rho \|\mathbf{Y}\|_F \leq \epsilon_2, \end{aligned} \tag{31}$$

where ϵ_1 and ϵ_2 are the sufficiently small constants. The maximum number of iterations can also be set. The optimization problem (23) imposes the equality constraint to represent the reverberation component without modeling the Gaussian noise in contrast to (9). To include the Gaussian noise model, an additional variable for the noise and its update step are required, which also requires an additional parameter to be controlled. Nevertheless, in Sect. IV, we experimentally validate that the proposed algorithm performs well without including the noise variable even when small Gaussian noise is added.

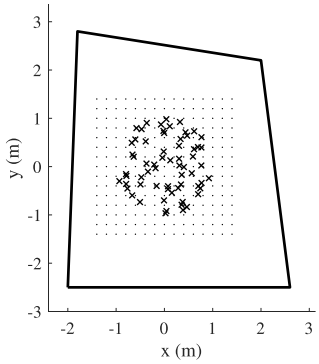


Fig. 2. Setup for 2D numerical experiments. Bold line, crosses, and dots represent room geometry, microphone positions, and grid point positions, respectively. 64 microphones are randomly placed in the circular region.

IV. Experiments

Simulation and practical experiments are conducted to evaluate the proposed method. For comparison, we also evaluate the methods based on the simple sparse decomposition (9), sparse decomposition in the monopole and plane-wave dictionaries (17), and sparse in the monopole dictionary and low-rank decomposition (21), which are denoted as **S**, **SS**, and **SL**, respectively. The proposed method is denoted as **SSL**.

First, numerical simulation results in 2D are shown. Single frequency and broadband cases are evaluated in Sect. IV-A and IV-B, respectively. Second, experimental results in a practical environment are shown in Sect. IV-C.

A. Simulation in 2D for Single Frequency Case

The experimental setup is shown in Fig. 2. The bold line, crosses, and dots represent the room geometry, microphone positions, and grid point positions for the dictionary, respectively. The dictionary matrix $\mathbf{D}_f$ is constructed by using the grid points regularly aligned on a square region of 3.0×3.0 m^2 with its center at the origin. The intervals and number of grid points are set as 0.2 m and 15×15, respectively. 64 microphones are randomly placed inside a circular region of 1.0 m radius with its center at the origin. The number of time frames used for decomposition, i.e., T, is set at 100. The 2D reverberant sound field is simulated by a finite element method [36]. By using the estimated $\mathbf{X}$, the sound field of the particular solution $u_{\mathrm{P}}(\mathbf{r})$ is reconstructed inside a square region of 3.0×3.0 m^2 with its center at the origin. The reconstruction region is regularly discretized at intervals of 0.05 m and the pressure $u_{\mathrm{P}}(\mathbf{r})$ is calculated at each point.

256 uniformly sampled plane waves from 0 to 2π rad are used as $\mathbf{W}$. The constant parameters η_D and η_W are determined as $\eta_D = 1.02\sigma^2_{\max}(\mathbf{D}_f)$ and $\eta_W = 1.02\sigma^2_{\max}(\mathbf{W}_f)$, respectively. The parameter ρ is adaptively changed at each iteration as in [26]. Initial values are set as matrices of all ones. For stopping rules, ϵ_1 and ϵ_2 in (30), (31) and ξ_1 and ξ_2 in (13), (14) are set to 1.0×10^{-4}. The maximum number of iterations is 400.

For evaluation, we define three types of performance measure: signal-to-distortion ratio for decomposition (SDRD), F-measure (F_{msr}), and signal-to-distortion ratio for reconstruction (SDRR).

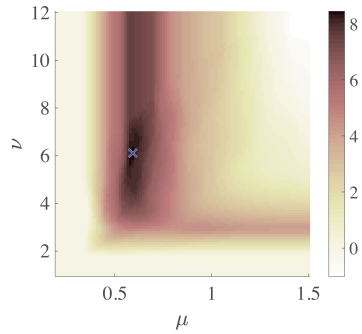


Fig. 3. Signal-to-distortion ratio for reconstruction (SDRR) of the proposed **SSL** model, with respect to balancing parameters μ and ν. Blue cross mark represents the values for the highest SDRR 8.44 dB.

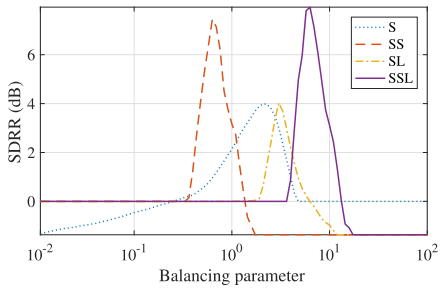


Fig. 4. Signal-to-distortion ratio for reconstruction (SDRR) as a function of balancing parameter (λ in **S**, μ in **SS**, ν in **SL**, and γ in **SSL**) when β in (36) is set at 0.10. Highest SDRDs for **S**, **SS**, **SL**, and **SSL** are 3.99, 7.45, 4.03, 7.95 dB, respectively.

The SDRD evaluates the accuracy of the decomposition result of $\mathbf{X}$ as

$$\mathrm{SDRD} = 10\log_{10}\frac{\|\mathbf{X}_{\mathrm{true}}\|_F^2}{\|\hat{\mathbf{X}} - \mathbf{X}_{\mathrm{true}}\|_F^2}, \tag{32}$$

where $\mathbf{X}_{\mathrm{true}}$ and $\hat{\mathbf{X}}$ are the true and estimated $\mathbf{X}$, respectively. Note that $\mathbf{X}_{\mathrm{true}}$ can be defined only when the point sources are located on the grid points. The operator $\mathrm{supp}(\cdot)$ is defined to extract a set of indexes such that the activated grid point is larger than the threshold value ζ as

$$\mathrm{supp}(\mathbf{X}) = \left\{ n \in \{1, \ldots, N\} \mid \|\mathbf{X}_n\|_F^2 > \zeta \right\}. \tag{33}$$

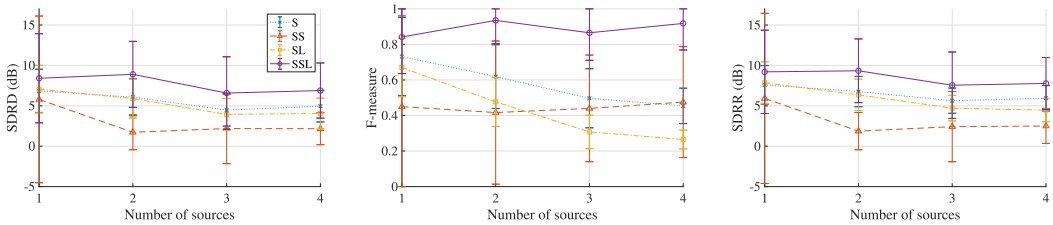
We here set ζ by using $\mathbf{X}_{\mathrm{true}}$ as $\zeta = \min(\|\mathbf{X}_{\mathrm{true},n'}\|_F^2) \times 10^{-2}$, where n' denotes the index set of the true activations. Then, F_{msr} is defined as

$$F_{\mathrm{msr}} = 2\frac{|\mathrm{supp}(\hat{\mathbf{X}}) \cap \mathrm{supp}(\mathbf{X}_{\mathrm{true}})|}{|\mathrm{supp}(\hat{\mathbf{X}})| + |\mathrm{supp}(\mathbf{X}_{\mathrm{true}})|}. \tag{34}$$

Therefore, F_{msr} equals to 1 when the activated indexes of these matrices are exactly the same. The SDRR is defined to evaluate the reconstruction accuracy of $u_{\mathrm{P}}(\mathbf{r})$ as

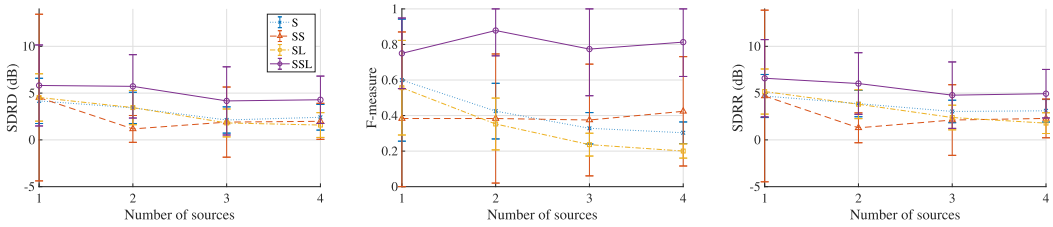
$$\mathrm{SDRR} = 10\log_{10}\frac{\iint |u_{\mathrm{P,true}}(\mathbf{r},\omega)|^2 d\mathbf{r}d\omega}{\iint |\hat{u}_{\mathrm{P}}(\mathbf{r},\omega) - u_{\mathrm{P,true}}(\mathbf{r},\omega)|^2 d\mathbf{r}d\omega}, \tag{35}$$

where $u_{\mathrm{P,true}}(\mathbf{r},\omega)$ and $\hat{u}_{\mathrm{P}}(\mathbf{r},\omega)$ are true and estimated pressure distribution of the particular solution $u_{\mathrm{P}}(\mathbf{r},\omega)$ at the frequency



(a) SDRD (b) F_{msr} (c) SDRR

Fig. 5. Signal-to-distortion ratio for decomposition (SDRD), F-measure (F_{msr}), and signal-to-distortion ratio for reconstruction (SDRR), as a function of the number of sources for absorption ratio 0.50 at frequency of 1 kHz (SNR 30 dB). Error bar denotes standard deviation.



(a) SDRD (b) F_{msr} (c) SDRR

Fig. 6. Signal-to-distortion ratio for decomposition (SDRD), F-measure (F_{msr}), and signal-to-distortion ratio for reconstruction (SDRR), as a function of the number of sources for absorption ratio 0.30 at frequency of 1 kHz (SNR 30 dB).

ω. Note that these performance measures are calculated for a single frequency in this section.

First, we investigate the effect of the balancing parameters to the performance. The SDRR of **SSL** is plotted with respect to μ and ν when a point source is located at (0.0, 0.4) m and the frequency is 1.0 kHz in Fig. 3. The absorption ratio on the room boundary is set at 0.30 and SNR is 30 dB. When μ is set at an excessively large value, $\mathbf{V}$ tends to be zero and $\mathbf{Z}$ is represented only by $\mathbf{U}$, whose performance gets closer to **SS**. Similarly, when ν is set at an excessively large value, $\mathbf{U}$ tends to be zero and $\mathbf{Z}$ is represented only by $\mathbf{V}$, whose performance gets closer to **SL**. The highest SDRR is achieved around the point of balancing these two values, which is denoted by the blue cross. The highest SDRR is 8.44 dB and its balancing parameters are $\mu = 0.59$ and $\nu = 6.11$. Obviously, finding the best balancing parameters is not a trivial task for **SSL**. In the following experiments, we fixed the ratio between μ and ν as

$$\mu = \beta\gamma, \quad \nu = (1-\beta)\gamma. \tag{36}$$

Fig. 4 shows the relationship between SDRR and the balancing parameter of each method (λ in **S**, μ in **SS**, ν in **SL**, and γ in **SSL**) when β in (36) is set at 0.10. The highest SDRRs of **S**, **SS**, **SL**, and **SSL** are 3.99, 7.45, 4.03, and 7.95 dB, respectively. Although the parameter choice by (36) is an empirical rule and the best result of **SSL** cannot be obtained, exhaustive search of two parameters can be avoided by limiting the search range.

We investigate three parameters of the absorption ratio on the room boundary, 0.50, 0.30, and 0.10. The balancing parameters are set so that the highest SDRRs are achieved when a point source is located at the origin. The parameter β in (36) is set at 0.10. The frequency is set at 1.0 kHz. The amplitude of each source is generated by the complex Gaussian distribution. Gaussian noise is also added so that signal-to-noise ratio (SNR) becomes 30 dB. We randomly choose locations of the point sources from the grid points and the performance measures are averaged over 20 trials. The relationships between three performance measures and the number of sources are plotted in Figs. 5, 6, and 7. The error bar denotes the standard deviation of the performance measures. As the absorption ratio decreases, the performance measures decrease for all the methods. The highest performance measures are achieved by **SSL** at the absorption ratio of 0.5 and 0.3. They get closer to those of **SS** as the absorption ratio decreases, and those of **SS** and **SSL** are similar at the absorption ratio 0.1. The standard deviation of **SS** is large because its performance depends on the source position. It is reduced in **SSL**, especially when the number of sources is 1. To investigate the case that there are more sources than 4, SDRR is plotted with respect to the number of sources up to 12 in Fig. 8. The average SDRRs for the absorption ratio of 0.30 are only shown. The reconstruction accuracy of all the methods decreases as the number of sources increases; however, **SSL** exhibits the highest SDRR even for the case of 12 sources.

To analyze the decomposition and reconstruction results in detail, the case that two point sources are located at (0.0, 0.4) m and $(-0.2, -1.2)$ m for 0.30 absorption ratio is investigated as an example. Figs. 9 and 10 are the distributions of $\mathbf{X}$ and $u_{\mathrm{P}}(\mathbf{r})$, respectively. The performance measures of **S**, **SS**, **SL**, and **SSL** are 3.24, 2.88, 2.84, and 8.49 for SDRD, 0.25, 0.67, 0.25, and 1.00 for F_{msr}, and 3.77, 3.23, 3.36, and 8.95 dB for SDRR. In **SS**, one source is activated but the other one is not. On the other hand, two point sources are accurately activated in **SSL**. The power of weight coefficients of the plane waves $\mathbf{U}_f$ for **SS** and **SSL** are plotted in Fig. 11. The distribution of plane waves for **SSL** is more sparse than that of **SS**. Fig. 12 shows the normalized singular value of $\mathbf{V}_f$ for **SL** and **SSL**. The number of singular values larger than 10^{-3} is 7 for **SL** and 1 for **SSL**, which is different from the number of sources 2. The reverberant component is not well separated in **SL** with the matrix $\mathbf{V}_f$ of

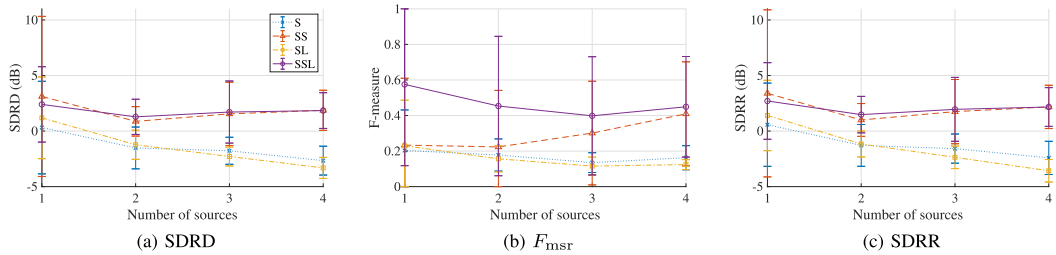


Fig. 7. Signal-to-distortion ratio for decomposition (SDRD), F-measure (F_{msr}), and signal-to-distortion ratio for reconstruction (SDRR), as a function of the number of sources for absorption ratio 0.10 at frequency of 1 kHz (SNR 30 dB).

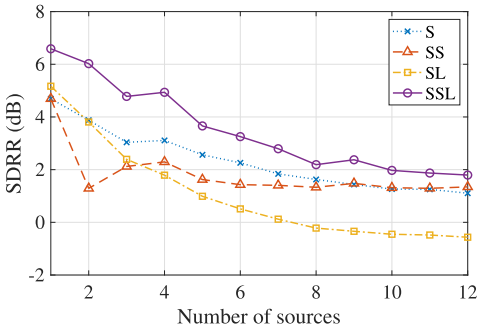


Fig. 8. Signal-to-distortion ratio for reconstruction (SDRR) as a function of the number of sources from 1 to 12 for absorption ratio 0.30 at frequency of 1 kHz (SNR 30 dB).

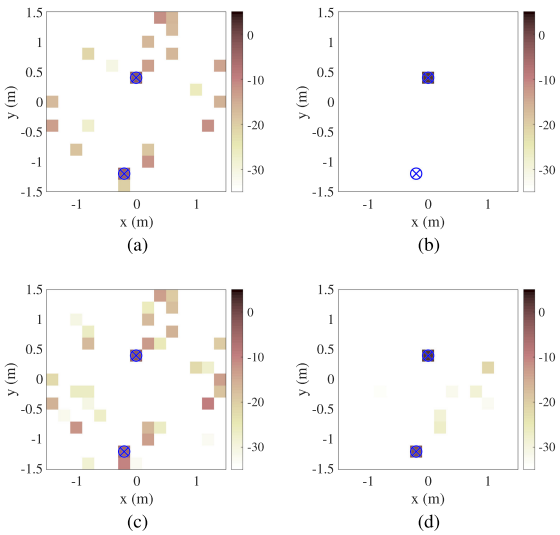


Fig. 9. Decomposition result for the different sound field models, 0.30 absorption ratio, single frequency at 1 kHz. Blue markers indicate the source locations. SDRD for **S**, **SS**, **SL**, and **SSL** are 4.05, 2.88, 4.55, and 9.06 dB, respectively. Their F_{msr} are 0.80, 0.67, 0.40, 1.00, respectively. (a) **S**; (b) **SS**; (c) **SL**; (d) **SSL**.

rank 7. On the other hand, in **SSL**, the reverberant component that the sparse plane waves cannot approximate seems to be represented by the rank 1 matrix $\mathbf{V}_f$. The accuracy of the reverberant component (SDRD defined for $\mathbf{Z}$) of **S**, **SS**, **SL**, **SSL** is 6.36, 7.22, 6.96, and 12.94 dB.

Figs. 13 and 14 shows the results when the balancing parameters are optimized for this specific case. The SDRR of **S**, **SS**, **SL**, and **SSL** are 4.15, 9.13, 3.68, 9.15 dB. In this case, the distribution of the plane waves of **SS** and **SSL** are very close as shown in Fig. 13. In Fig. 14, the singular values larger than 10^{-3} for **SL** is reduced although that for **SSL** is the same as the result in Fig. 12. The accuracy of the reverberant component of **S**, **SS**, **SL**, and **SSL** are 6.93, 12.93, 6.93, and 13.01 dB. The performance of **SS** and **SSL** gets closer when the balancing parameters are optimized for each scenario. When the balancing parameters optimized for the single point source at the origin is used, the performance degradation of **SS** is significant compared to **SSL**.

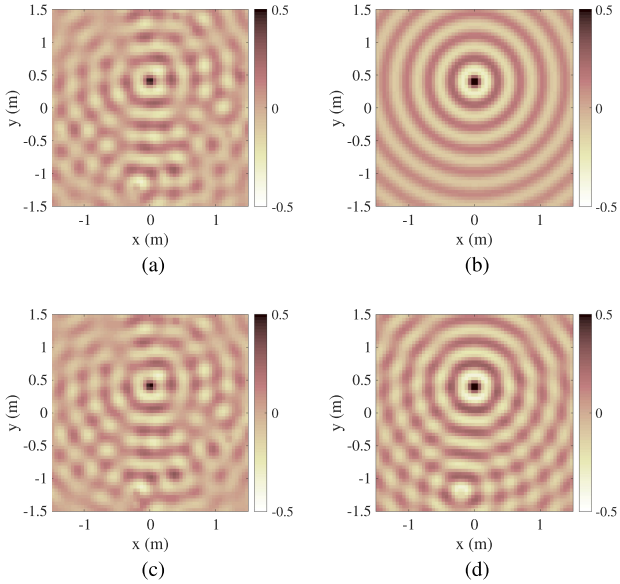


Fig. 10. Reconstructed sound field for the different sound field models, 0.30 absorption ratio, single frequency at 1 kHz. SDRR for **S**, **SS**, **SL**, and **SSL** are 3.38, 3.95, 3.37, and 8.52 dB, respectively. (a) **S**; (b) **SS**; (c) **SL**; (d) **SSL**.

The relationship between SDRR and number of sources for the case of 20 dB SNR is shown in Figs. 15. The SDRRs are slightly decreased from the SNR 30 dB case. Although the noiseless case is assumed in the proposed algorithm (23), **SSL** still performs well when a small noise is added.

B. Simulation in 2D for Broadband Case

In the broadband case, the sampling frequency is set at 4.0 kHz. The source signal is generated by the Gaussian distribution. An STFT is used to convert it to the time-frequency

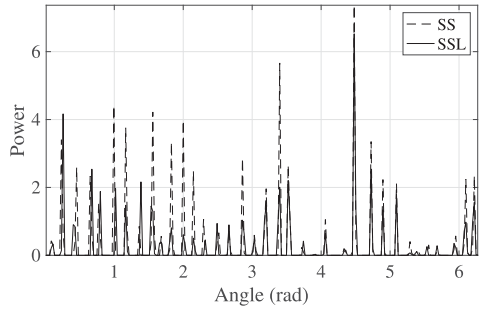


Fig. 11. Power of weight coefficients of plane waves $\mathbf{U}_f$ for **SS** and **SSL** when two point sources are located at $(0.0, 0.4)$ m and $(-0.2, -1.2)$ m and the absorption ratio is 0.30 at the frequency of 1 kHz (SNR 30 dB).

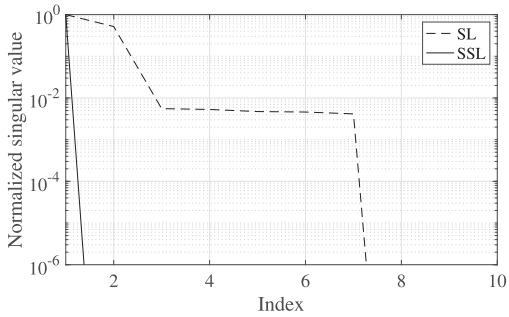


Fig. 12. Normalized singular value of $\mathbf{V}_f$ for **SL** and **SSL** when two point sources are located at $(0.0, 0.4)$ m and $(-0.2, -1.2)$ m and the absorption ratio is 0.30 at the frequency of 1 kHz (SNR 30 dB).

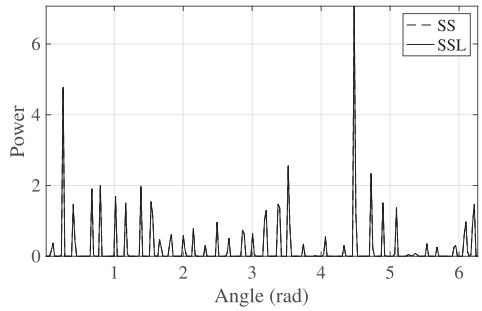


Fig. 13. Power of weight coefficients of plane waves $\mathbf{U}_f$ for **SS** and **SSL** when two point sources are located at $(0.0, 0.4)$ m and $(-0.2, -1.2)$ m and the absorption ratio is 0.30 at the frequency of 1 kHz (SNR 30 dB). The balancing parameters are optimized for this specific case.

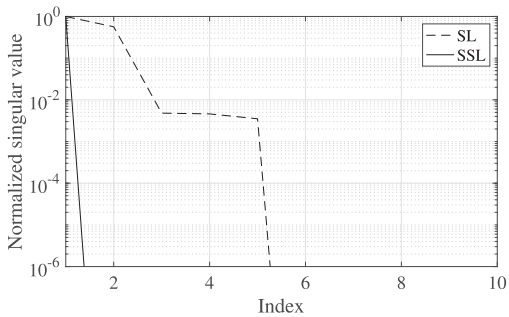


Fig. 14. Normalized singular value of $\mathbf{V}_f$ for **SL** and **SSL** when two point sources are located at $(0.0, 0.4)$ m and $(-0.2, -1.2)$ m and the absorption ratio is 0.30 at the frequency of 1 kHz (SNR 30 dB). The balancing parameters are optimized for this specific case.

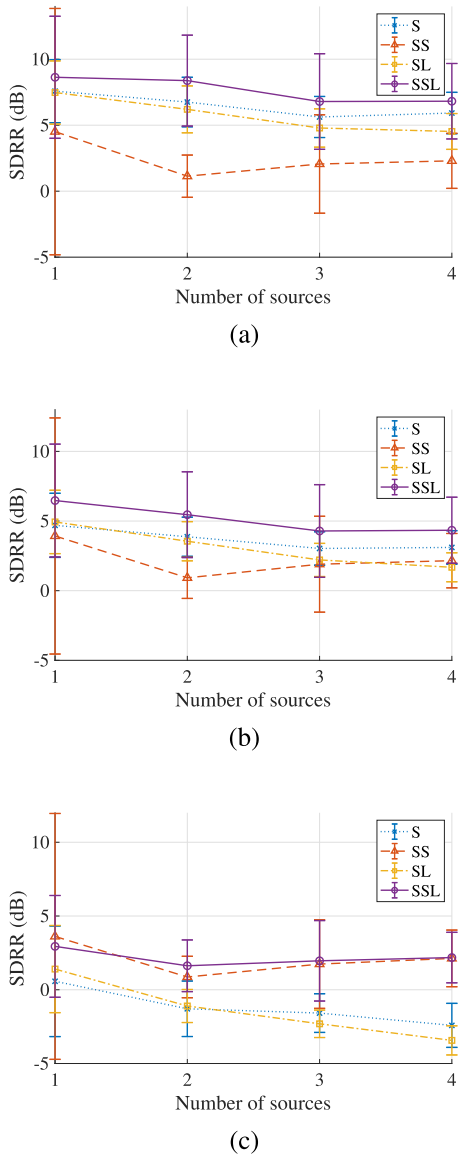
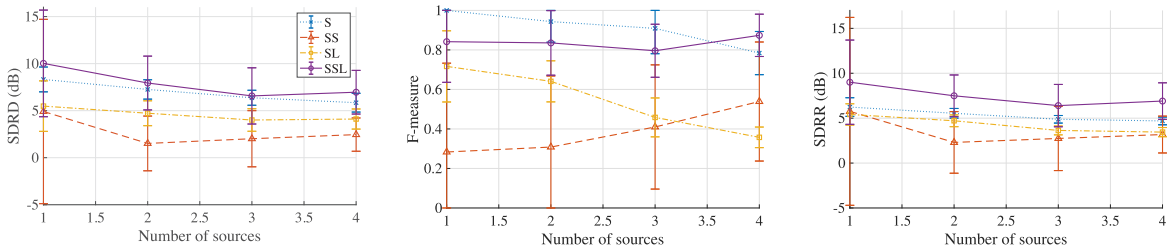


Fig. 15. Signal-to-distortion ratio for reconstruction (SDRR), as a function of the number of sources for absorption ratio a) 0.50; b) 0.30; c) 0.10, at frequency of 1 kHz (SNR 20 dB). Error bar denotes standard deviation.

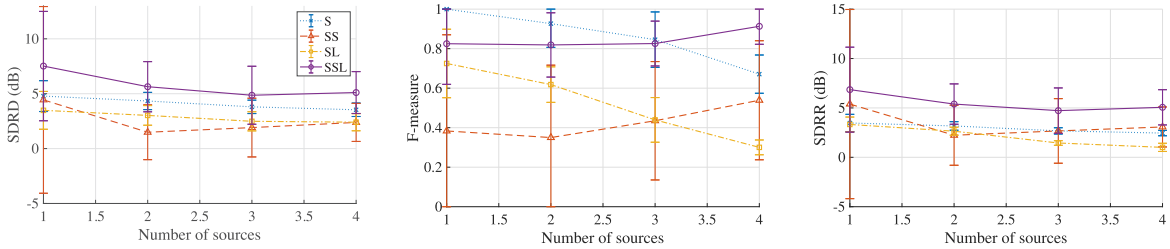
domain with a frame length of 64 samples and a shift length of 32 samples. The other settings are the same as those in the single frequency case.

The three performance measures for the absorption ratio of 0.50, 0.30, and 0.10 are plotted with respect to the number of sources in Figs. 16, 17, and 18, respectively. The SNR is set at 30 dB. Again, the balancing parameters are set so that the highest SDRRs are achieved when a point source is located at the origin. The parameter β in (36) is set at 0.30. The results show the similar tendency of those in the single frequency case. The average SDRR with respect to the number of sources up to 12 is also plotted in Fig. 19 for the case of the absorption ratio of 0.30. The highest SDRR is achieved by **SSL** even for the case of 12 sources. The difference of the reconstruction accuracy between **SSL** and **S** is larger than that of the single-frequency case.



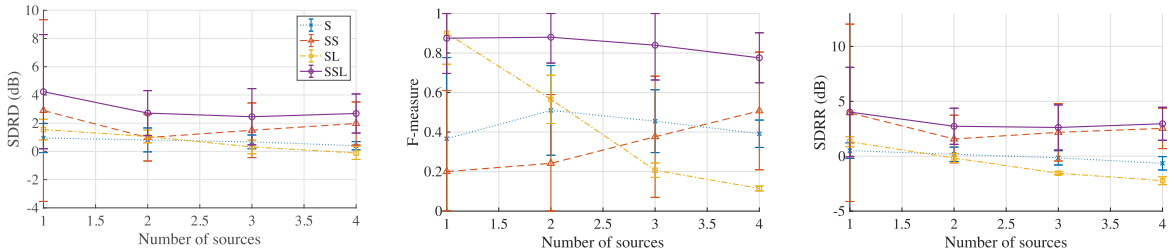
(a) SDRD (b) F_{msr} (c) SDRR

Fig. 16. Signal-to-distortion ratio for decomposition (SDRD), F-measure (F_{msr}), and signal-to-distortion ratio for reconstruction (SDRR), as a function of the number of sources for absorption ratio 0.50 in broadband case (SNR 30 dB). Error bar denotes standard deviation.



(a) SDRD (b) F_{msr} (c) SDRR

Fig. 17. Signal-to-distortion ratio for decomposition (SDRD), F-measure (F_{msr}), and signal-to-distortion ratio for reconstruction (SDRR), as a function of the number of sources for absorption ratio 0.30 in broadband case (SNR 30 dB).



(a) SDRD (b) F_{msr} (c) SDRR

Fig. 18. Signal-to-distortion ratio for decomposition (SDRD), F-measure (F_{msr}), and signal-to-distortion ratio for reconstruction (SDRR), as a function of the number of sources for absorption ratio 0.10 in broadband case (SNR 30 dB).

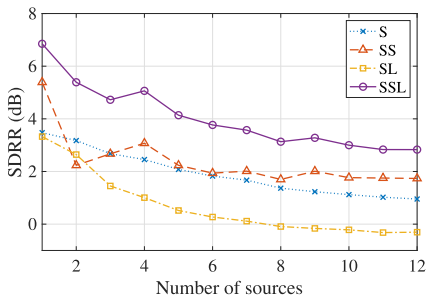


Fig. 19. Signal-to-distortion ratio for reconstruction (SDRR) as a function of the number of sources from 1 to 12 for absorption ratio 0.30 in broadband case (SNR 30 dB).

The decomposition and reconstruction results for 0.30 absorption ratio are plotted in Figs. 20 and 21. Again, two point sources are located at $(0.0, 0.4)$ m and $(-0.2, -1.2)$ m. The performance measures of **S**, **SS**, **SL**, and **SSL** are 4.05, 2.88, 4.55, and 9.06 for SDRD, 0.80, 0.67, 0.40, and 1.00 for F_{msr}, and 3.38, 3.95, 3.37, and 8.52 dB for SDRR. Numerical tests with SNR 20 dB gave similar results.

C. Experiments in Practical Environment

We demonstrate the experimental results in a practical environment. A linear microphone array with 32 elements is set as shown in Fig. 22. This array geometry is the most commonly used in telecommunication and live-broadcasting systems. The interval between microphones is 0.06 m. The reverberation time (T_{60}) of this room is 603 ms. The grid points are set inside a 2D square region of 2.5×2.5 m^2 centered at $(0.0, -2.0, 0.0)$ m. The intervals of the grid points are 0.10 m in the x direction and 0.20 m in the y direction. Thus, the number of grid points is 25×13. Two loudspeakers are set at $(-0.5, 1.0, 0.0)$ m and $(0.5, 2.0, 0.0)$ m as sound sources. The sampling frequency is 8.0 kHz. The STFT parameters are 128 samples for the frame

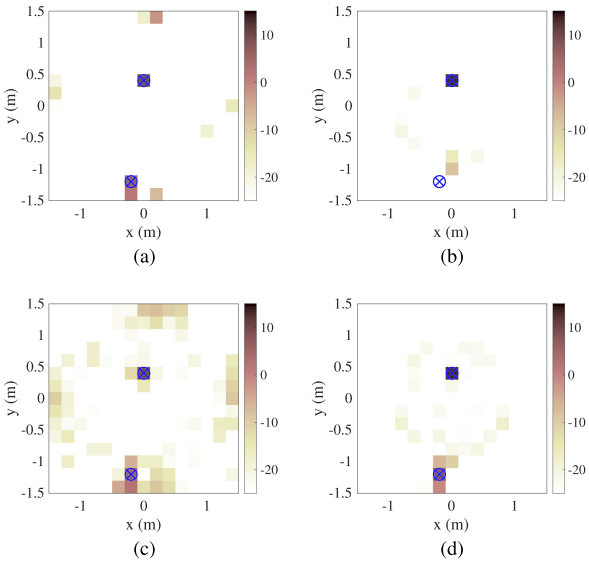


Fig. 20. Decomposition result for the different sound field models, 0.30 absorption ratio in broadband case. SDRD for **S**, **SS**, **SL**, and **SSL** are 4.05, 2.88, 4.55, 9.06, respectively. Those F_{msr} are 0.80, 0.67, 0.40, 1.00, respectively. (a) **S**; (b) **SS**; (c) **SL**; (d) **SSL**.

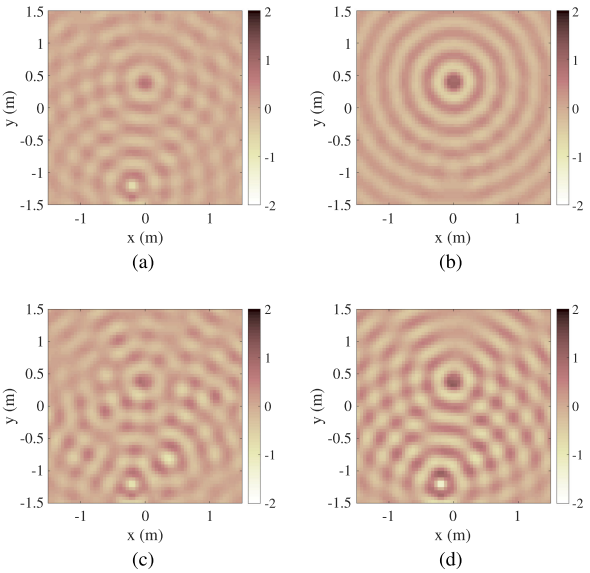


Fig. 21. Reconstructed sound field at 1000 Hz for the different sound field models, 0.30 absorption ratio in broadband case broadband. SDRR for **S**, **SS**, **SL**, and **SSL** are 3.38, 3.95, 3.37, and 8.52, respectively. (a) **S**; (b) **SS**; (c) **SL**; (d) **SSL**.

length and 64 samples for the shift length. The source signals are speech signals taken from the RWCP-SP99 [37] database. The other settings are the same as those in the simulation experiments.

Since true sound field distribution cannot be defined in this case, we evaluate each method by using F_{msr} with a small modification. The estimated $\mathbf{X}$ is normalized by the smaller amplitude at the grid points of the true source locations. Then, F_{msr} defined in (34) is calculated by setting ζ at 0.1. The

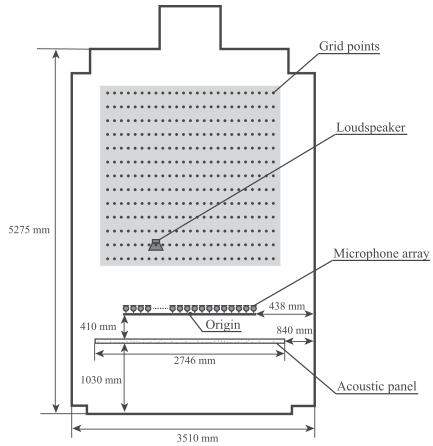


Fig. 22. Experimental setup in a realistic environment. A linear microphone array is used, with 32 elements. Reverberation time (T_{60}) is 603 ms.

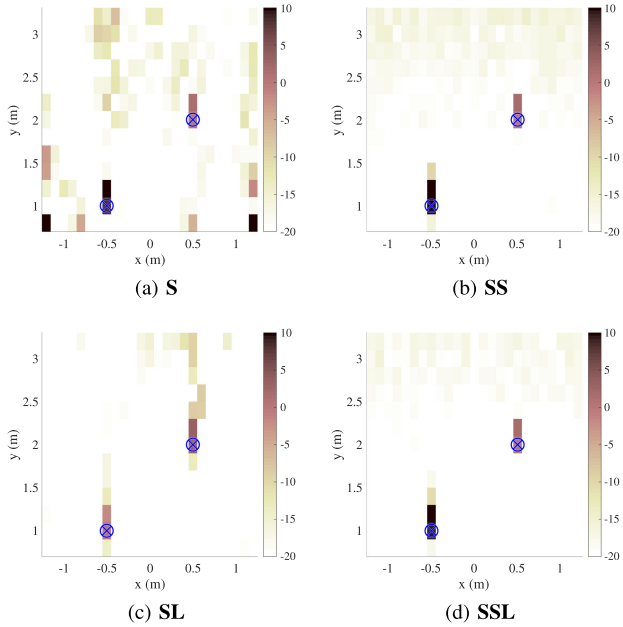


Fig. 23. Decomposition result in practical environment. F_{msr} of **S**, **SS**, **SL**, **SSL** are 0.24, 0.57, 0.36, and 0.67, respectively.

balancing parameters of each method are set so that F_{msr} becomes the largest. The parameter β in (36) is set at 0.40. Fig. 23 shows the decomposition results. F_{msr} of **S**, **SS**, **SL**, and **SSL** are 0.24, 0.57, 0.36, and 0.67, respectively. In **S**, many false activations can be found. They are significantly reduced in **SL** and **SS**, but small activations still remain on the grids in no source region. The sources are more sparsely identified in **SSL** and the largest F_{msr} is achieved.

V. Conclusion

This paper proposes a general model for sparse sound field decomposition method, in a reverberant environment. Numerical and experimental results indicate that this decomposition

enables an accurate sound field estimation, inside a region including sources. This study demonstrates that, in a reverberant environment, the reverberating component of the field cannot be treated as a residual, but must be explicitly treated. Here, we have shown that this reverberating component can be modeled as the sum of a small number of plane waves, and a low-rank part.

Although we have here derived an efficient ADMM-based algorithm for the corresponding decomposition, there are still a number of open issues. By making the model richer, one is able to model sound fields in a wider range of practical cases, but this comes at the cost of extra hyperparameters such as the balancing coefficients between components. In this study, we have derived empirical rules to set these parameters, that provide near-optimal results in the tested configurations, but the robustness of this approach is still to be verified. Furthermore, a more complex model increases the computational complexity, limiting, for instance, the number of spatial grid points in the model. Future work will also focus on algorithmic implementations that leverage specific computing architectures, found for instance in highly-parallel GPU boards. Near real-time decompositions might allow rich augmented / virtual reality environments, where only the parameters of the sound field, and not the individual microphone signals, are remotely transmitted.

REFERENCES

[1] J. D. Maynard, E. G. Williams, and Y. Lee, "Nearfield acoustic holography: I. theory of generalized holography and the development of NAH," *J. Acoust. Soc. Am.*, vol. 78, no. 4, pp. 1395–1413, 1985.

[2] Z. Wang and S. F. Wu, "Helmholtz equation-least-squares method for reconstructing the acoustic pressure field," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2020–2032, 1997.

[3] E. Fermandez-Grande, "Sound field reconstruction using a spherical microphone array," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. 1168–1178, 2016.

[4] E. G. Williams, J. D. Maynard, and E. Skudrzyk, "Sound source reconstructions using a microphone array," *J. Acoust. Soc. Am.*, vol. 68, no. 1, pp. 340–344, 1980.

[5] P. A. Nelson and S. H. Yoon, "Estimation of acoustic source strength by inverse methods: Part I, conditioning of the inverse problems," *J. Sound Vib.*, vol. 233, no. 4, pp. 639–664, 2000.

[6] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3094–3103, 2005.

[7] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2724–2736, 2006.

[8] M. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, 2005.

[9] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "Analytical approach to wave field reconstruction filtering in spatio-temporal frequency domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 685–696, Apr. 2013.

[10] 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.

[11] N. Ueno, S. Koyama, and H. Saruwatari, "Sound field recording using distributed microphones based on harmonic analysis of infinite order," *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 135–139, Jan. 2018.

[12] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. London, U.K.: Academic Press, 1999.

[13] G. H. Koopmann, L. Song, and J. B. Fhanline, "A method for computing acoustic fields based on the principle of wave superposition," *J. Acoust. Soc. Am.*, vol. 86, no. 5, pp. 2433–2438, 1989.

[14] M. E. Johnson, S. J. Elliot, K. H. Baek, and J. Garcia-Bonito, "An equivalent source technique for calculating the sound field inside an enclosure containing scattering objects," *J. Acoust. Soc. Am.*, vol. 104, no. 3, pp. 1221–1231, 1998.

[15] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[16] G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin, and R. Gribonval, "Near-field acoustic holography using sparsity and compressive sampling principles," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1521–1534, 2012.

[17] A. Asaei, H. Bourlard, M. Taghizadeh, and V. Cevher, "Model-based sparse component analysis for reverberant speech localization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, May 2014, pp. 1453–1457.

[18] N. Murata, S. Koyama, N. Takamune, and H. Saruwatari, "Sparse representation using multidimensional mixed-norm penalty with application to sound field decomposition," *IEEE Trans. Signal Process.*, vol. 66, no. 12, pp. 3327–3338, Jun. 2018.

[19] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 11, pp. 2301–2312, Nov. 2013.

[20] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling ambisonics sound scenes using compressed sensing techniques," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, 2011, pp. 1–4.

[21] S. Koyama, S. Shimauchi, and H. Ohmuro, "Sparse sound field representation in recording and reproduction for reducing spatial aliasing artifacts," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, May 2014, pp. 4443–4447.

[22] A. Moiola, R. Hiptmair, and I. Perugia, "Plane wave approximation of homogeneous Helmholtz solutions," *Z. Angew. Math. Phys.*, vol. 62, pp. 809–837, 2011.

[23] S. Koyama, N. Murata, and H. Saruwatari, "Sparse sound field decomposition for super-resolution in recording and reproduction," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3780–3895, 2018.

[24] S. Koyama and L. Daudet, "Comparison of reverberation models for sparse sound field decomposition," in *Proc. IEEE Int. Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, Oct. 2017, pp. 214–218.

[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *J. Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[26] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, 2011, pp. 612–620.

[27] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, NY, USA: Springer, 2010.

[28] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A reweighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[29] S. F. Cotter, D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[30] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[31] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[32] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Math. Doklady*, vol. 27, pp. 372–376, 1983.

[33] S. Koyama and H. Saruwatari, "Sound field decomposition in reverberant environment using sparse and low-rank signal models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, Mar. 2016, pp. 345–349.

[34] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York, NY, USA: Wiley, 2002.

[35] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions to linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[36] F. Hecht, "New development in freefem++," *J. Numer. Math.*, vol. 20, no. 3–4, pp. 251–265, 2012.

[37] RWCP Speech Resources Consortium, "Japanese speech database (RWCP-SP99)," Accessed on: Feb. 10, 2016. [Online]. Available: http://research.nii.ac.jp/src/RWCP-SP99.html

Shoichi Kyoama (M'10) received the B.E., M.S., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 2007, 2009, and 2014, respectively. In 2009, he was a Researcher in acoustic signal processing with the Nippon Telegraph and Telephone Corporation. He moved to the University of Tokyo in 2014 and since 2018, he has been an Assistant Professor (Lecturer). From 2016 to 2018, he was also a Visiting Researcher and JSPS overseas research fellow with Paris Diderot University (Paris7), Institut Langevin, Paris, France. His research interests include acoustic inverse problems, sound field analysis and synthesis, and spatial audio.

He is a member of the Acoustical Society of America, the Audio Engineering Society, the Institute of Electronics, Information and Communication Engineers, and the Acoustical Society of Japan. He was the recipient of Itakura Prize Innovative Young Researcher Award by ASJ in 2015, and the Research Award by Funai Foundation for Information Technology in 2018.

Laurent Daudet (M'04–SM'10) received the Graduate degree from Ecole Normale Supérieure, Paris, France, and the Ph.D. degree in applied mathematics from Aix-Marseille University, Marseille, France. He is a Professor in physics with Paris Diderot University, Paris, France, currently on leave to become cofounder and CTO at LightOn, a startup developing optical co-processors for machine learning. He held various academic and honorary positions: Fellow of the Institut Universitaire de France, Visiting Scholar with Stanford University, USA, Visiting Senior Lecturer with Queen Mary University of London, U.K., Visiting Professor with the National Institute for Informatics in Tokyo, Japan. He has been a consultant to various small and large companies, and is a co-inventor in several patents. He has authored or coauthored nearly 200 scientific publications, in journal or international conferences, mostly on signal processing techniques for the physics of waves.