STRUCTURED SPARSE SIGNAL MODELS AND DECOMPOSITION ALGORITHM FOR SUPER-RESOLUTION IN SOUND FIELD RECORDING AND REPRODUCTION

Shoichi Koyama, Naoki Murata, and Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT

A method for achieving super-resolution of sound field recording and reproduction is proposed. To obtain driving signals of loudspeakers for reproduction from received signals of microphones, sparse signal decomposition makes it possible to reduce spatial aliasing artifacts when the number of microphones is less than that of loudspeakers. For more accurate and robust signal decomposition, we propose three types of group sparse signal model based on the physical properties of a sound field. In addition, a decomposition algorithm is derived to address these signal models as an extension of M-FOCUSS. In the simulation experiments, the accuracy of the sparse decomposition was significantly improved compared with that of M-FOCUSS. Furthermore, the accuracy of sound field reproduction using our proposed method was higher than that using current methods, especially at frequencies above the spatial Nyquist frequency.

Index Terms— sound field reproduction, sparse signal representation, super-resolution, wave field synthesis, wave field reconstruction filter

1. INTRODUCTION

To achieve high-fidelity audio systems, physical reproduction of a sound field may be one of the promising techniques. In practical recording and reproduction systems, sound pressures at multiple positions in the desired sound field are obtained with microphones, and then, they are reproduced with loudspeakers in a target area. Therefore, a method for obtaining driving signals of loudspeakers from the received signals of microphones is necessary. We define this type of signal transformation as sound-pressure-to-driving-signal (SP-DS) conversion. We focus on the SP-DS conversion problem when the array configuration of the microphones and loudspeakers are planar or linear.

Wave field synthesis (WFS) [1] is a well-known sound field synthesis method based on Kirchhoff-Helmholtz or Rayleigh integrals. WFS for a planar or linear loudspeaker array is based on the Rayleigh integral of the first kind [2]. Because driving signals of WFS must be equivalent to the distribution of the sound pressure gradient of a desired sound field, WFS cannot be directly applied for SP-DS conversion.

On the other hand, the *wave field reconstruction (WFR) filtering* method [3] makes SP-DS conversion possible by decomposing the received sound pressure distribution into spatial Fourier basis functions that correspond to uniformly sampled plane waves. Although stable and efficient signal conversion can be achieved by using this representation, artifacts originating from spatial aliasing notably occur, depending on the microphone and loudspeaker intervals. Under the significant effect of the spatial aliasing artifacts, listeners may be unable to clearly localize the reproduced sound images. Furthermore, frequency characteristics of the reproduced sound are greatly affected, i.e., *coloration effect* [4].

To reduce the spatial aliasing artifacts, we have proposed an SP-DS conversion method based on a sparse sound field representation [5]. This method makes it possible to improve the reproduction accuracy at frequencies above the spatial Nyquist frequency when the number of microphones is smaller than that of loudspeakers; this feature can be regarded as a super-resolution of sound field recording and reproduction. For more accurate and robust decomposition and higher reproduction accuracy, prior information on the structure of the recording sound field may be useful. We propose three types of signal model for group sparse sound field representation that have advantages for super-resolution SP-DS conversion. In addition, we propose a group sparse decomposition algorithm by extending the M-FOCUSS algorithm [6] to address these signal models.

In a prior work, Ahrens and Spors [7] proposed a method of reducing spatial aliasing artifacts in sound field synthesis. In this method, the desired sound field is assumed to be known and is synthesized in a limited region; therefore, the recording step is not considered. Wabnitz *et al.* [8] proposed an upscaling method for the Ambisonics order, on the basis of a sparse plane-wave decomposition in a spherical array case. However, the plane-wave decomposition of the received sound pressure distribution can rarely be sparse in our planar or linear array case. In [9], a method based on MAP estimation was proposed. Optimal basis functions representing a sound field are obtained using prior knowledge of the primary source locations. Therefore, this method requires these locations in the SP-DS conversion. Our sparse representation-based algorithm was initially proposed in [5]. The method proposed in this paper is aimed at improving that algorithm using the structured sparse signal models.

2. GENERATIVE MODEL OF SOUND FIELD AND ITS SPARSE DECOMPOSITION

First, we briefly revisit the generative model of a sound field proposed in [5]. As shown in Fig. 1, a sound field in the recording area is divided into two regions, internal and external regions of a closed surface. The internal region is denoted as Ω. Components approximated as monopole sources are assumed to exist only inside Ω. When the sound pressure of the temporal frequency ω at the position $\mathbf{r}$ is denoted as $p(\mathbf{r})$, $p(\mathbf{r})$ can be represented as the sum of inhomogeneous and homogeneous terms, $p_{\mathrm{i}}(\mathbf{r})$ and $p_{\mathrm{h}}(\mathbf{r})$, as

$$\begin{aligned} p(\mathbf{r}) &= p_{\mathrm{i}}(\mathbf{r}) + p_{\mathrm{h}}(\mathbf{r}) \\ &= \int_{\mathbf{r}' \in \Omega} Q(\mathbf{r}') G(\mathbf{r}|\mathbf{r}') d\mathbf{r}' + p_{\mathrm{h}}(\mathbf{r}), \end{aligned} \tag{1}$$

where $Q(\mathbf{r})$ is the distribution of the monopole source components inside Ω, and $G(\mathbf{r}|\mathbf{r}')$ is three-dimensional free-field Green's func-

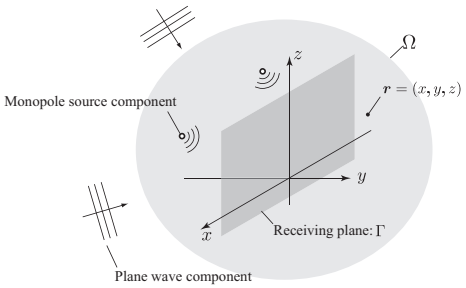


Fig. 1. Sound field in recording area modeled by sum of monopole source and plane wave components. Sound pressure is obtained on receiving plane Γ.

tion. The argument of the temporal frequency ω is omitted for notational simplicity. The homogeneous term $p_{\mathrm{h}}(\mathbf{r})$ can be represented as the sum of plane waves [10]. We assume that the sound pressure distribution on the receiving plane Γ is obtained using a microphone array.

When the region Ω is discretized as a set of grid points, (1) can be represented in the discrete form as

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{h}, \tag{2}$$

where $\mathbf{y} \in \mathbb{C}^M$ is the received signals of the microphones, $\mathbf{x} \in \mathbb{C}^N$ is the distribution of the monopole components at the grid points, $\mathbf{h} \in \mathbb{C}^M$ is the homogeneous term of the received signals, $\mathbf{D} \in \mathbb{C}^{M\times N}$ is the dictionary matrix of the monopole components, which has Green's function between the grid points and the microphones in each element, and M and N are, respectively, the numbers of microphones and grid points. Here, $N \gg M$ is assumed. Since the monopole source components may exist only at a few locations in Ω, only a few element of $\mathbf{x}$ may have nonzero values. Therefore, the sparse decomposition algorithm [11] can be applied to decompose $\mathbf{y}$ into $\mathbf{x}$ and $\mathbf{h}$. We applied the M-FOCUSS algorithm [6] to achieve this decomposition in [5].

The driving signals of the loudspeakers correspond to the sound pressure gradient on Γ when the loudspeakers are also aligned on a plane [2]. The decomposed components $\mathbf{x}$ and $\mathbf{h}$ are separately converted [2, 3, 5], and then the driving signals of the loudspeakers are obtained as their sum. Since more precise interpolation can be achieved by using the basis functions that depend on the monopole source components, spatial aliasing artifacts are reduced when there are more loudspeakers than microphones [5].

3. STRUCTURED SPARSITY BASED ON PHYSICAL PROPERTIES

For more precise SP-DS conversion, the decomposition of (2) must be more accurate and robust. Prior information on the structure of the sound field, i.e., the structure of the solution vector $\mathbf{x}$, may be useful for this purpose. We describe three types of group sparse signal model on the basis of the physical properties of the sound field.

Model 1: multiple time frames

When multiple time frames of $\mathbf{y}$ are available, each $\mathbf{x}$ may have the same sparsity pattern. This model is already used in [5] and its sparse decomposition is referred to as the multiple measurement vectors (MMV) problem [6].

We denote the index of the time frame as $l \in \{1, \cdots, L\}$, the signals of each l as $\mathbf{y}_l \in \mathbb{C}^M$, $\mathbf{x}_l \in \mathbb{C}^N$, and $\mathbf{h}_l \in \mathbb{C}^M$. By concatenating them in vectors, we can represent (2) as

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_L\end{bmatrix} = \begin{bmatrix}\mathbf{D} & & \mathbf{0}\\ & \mathbf{D} & \\ & \ddots & \\ \mathbf{0} & & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_L\end{bmatrix} + \begin{bmatrix}\mathbf{h}_1\\ \mathbf{h}_2\\ \vdots\\ \mathbf{h}_L\end{bmatrix}. \tag{3}$$

Each $\mathbf{x}_l$ is assumed to have nonzero values at the same positions.

In the context of the MMV problem, (3) is generally represented in a matrix form. Several sparse decomposition algorithms for this representation have been proposed [6, 12–14].

Model 2: temporal frequencies

Many kinds of acoustic source signals have a broad frequency band. Therefore, each $\mathbf{x}$ of multiple frequency bins may have the same sparsity pattern. Similar to model 1, using the index of the frequency bin $l \in \{1, \cdots, L\}$, we denote the signals of each l as $\mathbf{y}_l \in \mathbb{C}^M$, $\mathbf{x}_l \in \mathbb{C}^N$, and $\mathbf{h}_l \in \mathbb{C}^M$. Since Green's function depends on temporal frequency, the dictionary matrix of each l is denoted as $\mathbf{D}_l \in \mathbb{C}^{M\times N}$. Therefore, (2) can be represented as

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_L\end{bmatrix} = \begin{bmatrix}\mathbf{D}_1 & & \mathbf{0}\\ & \mathbf{D}_2 & \\ & \ddots & \\ \mathbf{0} & & \mathbf{D}_L\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_L\end{bmatrix} + \begin{bmatrix}\mathbf{h}_1\\ \mathbf{h}_2\\ \vdots\\ \mathbf{h}_L\end{bmatrix} \tag{4}$$

Again, each $\mathbf{x}_l$ is assumed to have nonzero values at the same positions. Note that the dictionary matrices in (4) are different in each group whereas those in (3) are the same.

Model 3: image sources and multipole components

Signals obtained in an ordinary room have reflections from walls in addition to direct sound. This phenomenon leads to the presence of monopole source components at the reflective image source locations [15]. As another example, multipole source components, such as a dipole and a quadrupole, may exist at the same location as the monopole source components [10]. These properties can be represented by the same group sparse signal model.

Using the index of the image sources $l \in \{1, \cdots, L\}$, we denote the signal of each l as $\mathbf{x}_l \in \mathbb{C}^N$. Green's function between the l-th image source location and the microphones is denoted as $\mathbf{D}_l \in \mathbb{C}^{M\times N}$. Therefore, (2) can be represented as

$$\mathbf{y} = \begin{bmatrix}\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_L\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_L\end{bmatrix} + \mathbf{h}. \tag{5}$$

Again, each $\mathbf{x}_l$ is assumed to have nonzero values at the same positions. Note that the length of $\mathbf{y}$ is degenerated to M. Therefore, the structure of the dictionary matrix is different from that in (3) and (4). In the case of image source components, room geometry must be known to design $\mathbf{D}_l$. In the case of multipole components, $\mathbf{D}_l$ becomes Green's function of each multipole.

Combinatorial model

Models 1, 2, and 3 can be combined. For example, in the case of combination of two groups, each $\mathbf{x}_l$ in (5) is replaced by the solution vector in (3) to combine models 1 and 3. The dictionary matrix must be designed accordingly. To combine I groups, the sets of indexes of the groups are denoted as $\mathcal{G}_1, \cdots, \mathcal{G}_I$, and the index of each group is denoted as $l_i \in \{1, \cdots, |\mathcal{G}_i|\}$. The signal vectors and dictionary matrix are denoted as $\tilde{\mathbf{x}} \in \mathbb{C}^{N|\mathcal{G}_1|\cdots|\mathcal{G}_I|}$, $\tilde{\mathbf{y}}$, $\tilde{\mathbf{h}}$, and $\tilde{\mathbf{D}}$, respectively. These can be related as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{D}}\tilde{\mathbf{x}} + \tilde{\mathbf{h}}. \tag{6}$$

Algorithm 1 Proposed group sparse decomposition algorithm

Initialize $\tilde{\mathbf{x}}^{(0)}, t = 1$
while loop $\neq 0$ **do**
  $\mathbf{w}^{(t)} \leftarrow [p^{-1/2}\|\tilde{\mathbf{x}}[1]^{(t-1)}\|_2^{1-p/2}, \cdots, p^{-1/2}\|\tilde{\mathbf{x}}[N]^{(t-1)}\|_2^{1-p/2}]$
  $\tilde{\mathbf{W}}^{(t)} \leftarrow \mathrm{diag}\left(\mathbf{w}^{(t)}, \cdots, \mathbf{w}^{(t)}\right)$
  $\mathbf{A}^{(t)} \leftarrow \tilde{\mathbf{D}}\tilde{\mathbf{W}}^{(t)}$
  $\tilde{\mathbf{x}}^{(t)} \leftarrow \tilde{\mathbf{W}}^{(t)}\mathbf{A}^{(t)H}(\mathbf{A}^{(t)}\mathbf{A}^{(t)H} + \lambda\mathbf{I})^{-1}\tilde{\mathbf{y}}$
  $t \leftarrow t + 1$
  if stopping condition is satisfied **then**
    loop $= 0$
  end if
end while

Each group is nested in the solution vector $\tilde{\mathbf{x}}$. The sizes of $\tilde{\mathbf{y}}$, $\tilde{\mathbf{h}}$, and $\tilde{\mathbf{D}}$ depend on the types of combined models.

In the context of the sound source localization [16–18], several works using combinatorial models 1 and 2 can be found. Model 3 plays an important role in sound field recording and reproduction because each basis function must be a solution of the wave equation to enable conversion from the decomposed signals to the driving signals of the loudspeakers. Helwani *et al.* [19] used model 2 and 3 for multichannel adaptive filtering.

4. GROUP SPARSE DECOMPOSITION ALGORITHM

To solve group sparse decomposition (6), we derive an extended algorithm of M-FOCUSS [6, 14]. We define a groupwise diversity measure of $\tilde{\mathbf{x}}$ as

$$J_{p,q}(\tilde{\mathbf{x}}) = \sum_{n=1}^{N} \|\tilde{\mathbf{x}}[n]\|_q^p = \sum_{n=1}^{N}\left(\sum_{l_1\in\mathcal{G}_1}\cdots\sum_{l_I\in\mathcal{G}_I}|x_{n,l_1,\cdots,l_I}|^q\right)^{p/q}, \quad (7)$$

where $0 \le p \le 1$ and $q \ge 1$, $\tilde{\mathbf{x}}[n]$ is the n-th element of each group, and $x_{n,l_1,\cdots,l_I}$ is the n-th element of the group $l_1, \cdots, l_I$. The optimization criteria can be described as

$$\min_{\tilde{\mathbf{x}}} \frac{1}{2}\|\tilde{\mathbf{y}} - \tilde{\mathbf{D}}\tilde{\mathbf{x}}\|_2^2 + \lambda J_{p,q}(\tilde{\mathbf{x}}), \quad (8)$$

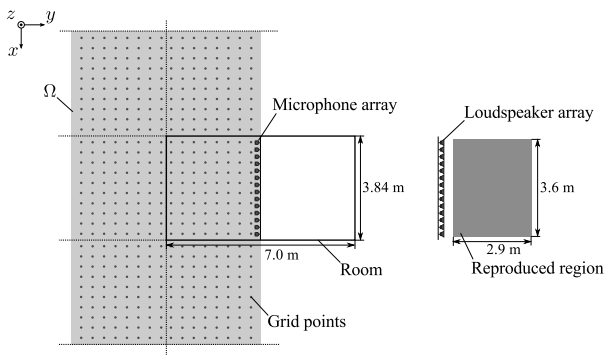
where λ is a parameter that balances the approximation error and the sparsity-inducing penalty $J_{p,q}(\tilde{\mathbf{x}})$.

Similar to M-FOCUSS, the case where $0 < p \le 1$ and $q = 2$ is addressed. The partial derivative of $J_{p,2}(\tilde{\mathbf{x}})$ with respect to the entry $x_{n',l'_1,\cdots,l'_I}$ is derived as

$$\begin{aligned}\frac{\partial J_{p,2}}{\partial x^*_{n',l'_1,\cdots,l'_I}} &= \frac{\partial}{\partial x^*_{n',l'_1,\cdots,l'_I}}\sum_{n=1}^{N}\left(\sum_{l_1\in\mathcal{G}_1}\cdots\sum_{l_I\in\mathcal{G}_I}|x_{n,l_1,\cdots,l_I}|^2\right)^{p/2}\\ &= p\left(\sum_{l_1\in\mathcal{G}_1}\cdots\sum_{l_I\in\mathcal{G}_I}\left|x_{n',l_1,\cdots,l_I}\right|^2\right)^{p/2-1}\cdot x_{n',l'_1,\cdots,l'_I}\\ &= p\left\|\tilde{\mathbf{x}}[n']\right\|_2^{p-2}x_{n',l'_1,\cdots,l'_I}. \end{aligned} \quad (9)$$

We define a vector $\mathbf{p} \in \mathbb{R}^N$ as

$$\mathbf{p} = \left[p\|\tilde{\mathbf{x}}[1]\|_2^{p-2}, \cdots, p\|\tilde{\mathbf{x}}[N]\|_2^{p-2}\right], \quad (10)$$



(a) Recording area (b) Target area

Fig. 2. Simulation setup

and a diagonal matrix $\tilde{\mathbf{P}} \in \mathbb{R}^{N|\mathcal{G}_1|\cdots|\mathcal{G}_I|\times N|\mathcal{G}_1|\cdots|\mathcal{G}_I|}$ as

$$\tilde{\mathbf{P}} = \mathrm{diag}(\mathbf{p}, \cdots, \mathbf{p}). \quad (11)$$

The gradient of the objective function (8) can be derived as

$$-\mathbf{D}^H(\tilde{\mathbf{y}} - \tilde{\mathbf{D}}\tilde{\mathbf{x}}) + \lambda\tilde{\mathbf{P}}\tilde{\mathbf{x}}. \quad (12)$$

By defining a vector,

$$\mathbf{w} = [p^{-1/2}\|\tilde{\mathbf{x}}[1]\|_2^{1-p/2}, \cdots, p^{-1/2}\|\tilde{\mathbf{x}}[N]\|_2^{1-p/2}], \quad (13)$$

and a diagonal matrix $\tilde{\mathbf{W}}$ accordingly, i.e., $\tilde{\mathbf{W}}^{-2} = \tilde{\mathbf{P}}$, the necessary optimality condition can be written as

$$\left((\tilde{\mathbf{D}}\tilde{\mathbf{W}})^H(\tilde{\mathbf{D}}\tilde{\mathbf{W}}) + \lambda\mathbf{I}\right)\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{x}} = (\tilde{\mathbf{D}}\tilde{\mathbf{W}})^H\tilde{\mathbf{y}}. \quad (14)$$

Finally, the following iterative scheme can be derived:

$$\tilde{\mathbf{x}}^{(t+1)} = \tilde{\mathbf{W}}^{(t)}\left((\tilde{\mathbf{D}}\tilde{\mathbf{W}}^{(t)})^H(\tilde{\mathbf{D}}\tilde{\mathbf{W}}^{(t)}) + \lambda\mathbf{I}\right)^{-1}(\tilde{\mathbf{D}}\tilde{\mathbf{W}}^{(t)})^H\tilde{\mathbf{y}}, \quad (15)$$

where $(\cdot)^{(t)}$ denotes the iteration index.

The proposed decomposition algorithm is summarized as Algorithm 1. As in the method presented in [5], the decomposed signals are separately converted into the driving signals of the loudspeakers.

5. EXPERIMENTS

Numerical simulations were conducted to evaluate the proposed method. First, sparse decomposition performances of the proposed decomposition algorithm and M-FOCUSS are compared. Second, we demonstrate a super-resolution of sound field recording and reproduction using the proposed method. Although the proposed method was derived for the case of planar arrays, we assumed that these arrays are linear in the experiments. The proposed method can be straightforwardly extended to the linear case.

5.1. Sparse decomposition performance

We focused on a signal model that combines models 1 and 3 in the experiments. Specifically, a group sparsity derived from multiple time frames and reflective image sources was considered. As shown in Fig. 2a, a half-space of a room was set as the recording area. The room size was $3.84\times7.0\times3.0$ m^3. The origin of the coordinate system was set at the center of the room. A linear microphone array was set along the x-axis with its center at the origin. The number of microphones was 32 and they were set at intervals of 12 cm. The directivity of the microphones was assumed to be omnidirectional. The room reverberation was simulated by the image method [15].

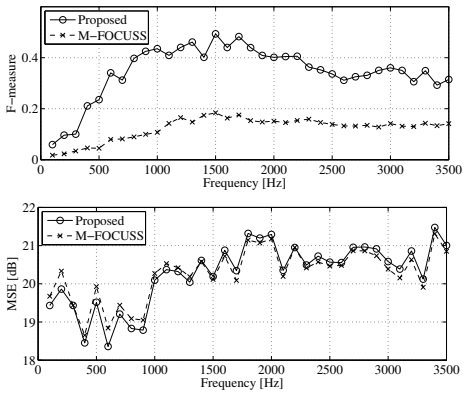


Fig. 3. Results of sparse decomposition performance. F_{msr} and MSE were averaged over 100 trials at each frequency.

We assumed that only the three walls at $y = \pm 1.92$ m and $x = -3.5$ m reflect sound waves, and their reflection coefficients were set as 0.4.

The two-dimensional region Ω was set to be 11.5×7.0 m^2 in size on the x-y-plane at $z = 0$ (shaded region in Fig. 2a); therefore, Ω included five image source areas. The number of grid points was 38 $(x) \times 17$ (y) inside the room, i.e., the direct source area, and they were set at intervals of 0.1 m for x and 0.2 m for y. The center of the grid points was at $(0.0, -1.6, 0.0)$ m. In the image source areas, these grid points were aligned at the corresponding image source locations. The total number of grid points was 114 $(x) \times 34$ (y).

A single sound source location was randomly chosen from the grid points in the direct source area. The source was assumed to have monopole characteristics. The source signal was a single-frequency sinusoidal wave. The amplitude of the source signal was generated by a complex Gaussian distribution with a mean of 1.0 and a variance of 0.5 at each time frame.

In both the proposed method and M-FOCUSS, the parameter p in the penalty term $J_{p,2}(\tilde{\mathbf{x}})$ was set as $p = 0.8$, and λ in (8) was set as 1.0×10^{-3}. The number of time frames of the observed signal was 16. The maximum iteration number of both methods was 100.

To evaluate the performance of sparse decomposition, we defined F-measure (F_{msr}) and mean square error (MSE) [14]. An operator $\mathrm{supp}(\cdot)$ extracts a set of indexes such that the amplitude of each element of the solution vector $\tilde{\mathbf{x}}$ is larger than a threshold value μ, as

$$\mathrm{supp}\,(\tilde{\mathbf{x}}) = \{n \in \{1, \cdots, N|\mathcal{G}_1||\mathcal{G}_2|\} \mid |x_n| > \mu\}, \qquad (16)$$

where μ is a threshold value that was set as 0.32. F_{msr} is defined as

$$F_{\mathrm{msr}} = 2\frac{|\mathrm{supp}\,(\tilde{\mathbf{x}}_{\mathrm{est}}) \cap \mathrm{supp}\,(\tilde{\mathbf{x}}_{\mathrm{true}})|}{|\mathrm{supp}\,(\tilde{\mathbf{x}}_{\mathrm{est}})| + |\mathrm{supp}\,(\tilde{\mathbf{x}}_{\mathrm{true}})|}, \qquad (17)$$

where $\tilde{\mathbf{x}}_{\mathrm{est}}$ and $\tilde{\mathbf{x}}_{\mathrm{true}}$ are the estimated and true solution vectors, respectively. Therefore, F_{msr} is equal to 1 when the set of activated indexes of these vectors are exactly the same. MSE is defined as the squared ℓ_2-norm of the error of the solution vector.

$$\mathrm{MSE} = 10\log_{10}\|\tilde{\mathbf{x}}_{\mathrm{true}} - \tilde{\mathbf{x}}_{\mathrm{est}}\|_2^2 \qquad (18)$$

These values were averaged over 100 trials.

Figure 3 shows plots of the results of F_{msr} and MSE when the frequency of the source signal was in the range from 100 Hz to 3500 Hz. Although MSE was almost the same in the two methods, F_{msr} was significantly improved when using the proposed method.

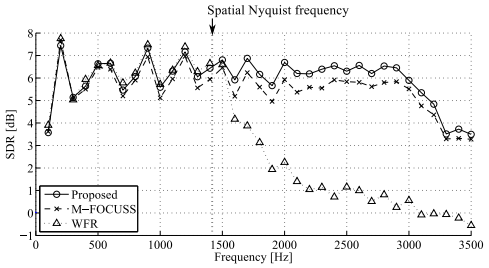


Fig. 4. Relationship between frequency and SDR

Model 3 makes it possible to accurately detect small amplitudes of reflective image sources. By introducing model 2, the variation in these values at each frequency may be reduced.

5.2. Reproduction performance

To evaluate the reproduction accuracy, the sound field captured as described in the previous section was reproduced using a linear loudspeaker array in the free field (Fig. 2b). In addition to the proposed method (Proposed) and the method proposed in [5] (M-FOCUSS), the WFR filtering method (WFR) [3] was also compared.

The linear loudspeaker array was located along the x-axis, as shown in Fig. 2b. The number of loudspeakers was 64 and they were set at intervals of 6 cm. To eliminate an artifact of faster amplitude decay [3], the loudspeakers were assumed to have line source characteristics. The sound pressure distribution were simulated in a 3.6×2.9 m^2 region at intervals of 1.5 cm on the x-y-plane at $z = 0$. The center of the simulated region was at $(0.0, 1.95, 0.0)$ m. The amplitudes were normalized using the average squared amplitude in the simulated region. In this experiment, the single sound source location was fixed at $(-0.65, -1.2, 0.0)$ m in the recording area. The general reproduction accuracy was evaluated using the signal-to-distortion ratio (SDR) [3, 5].

Figure 4 shows the relationship between SDRs and the frequency of the source signal. The spatial Nyquist frequency determined from the intervals between the microphones is indicated by the dashed line. The SDRs of Proposed and M-FOCUSS were significantly higher than that of WFR at frequencies above the spatial Nyquist frequency. Moreover, the SDR of Proposed was higher than that of M-FOCUSS. By improving F_{msr}, we also improved the reproduction accuracy, especially at frequencies above the spatial Nyquist frequency.

6. CONCLUSION

An SP-DS conversion method for super-resolution of sound field recording and reproduction was proposed. Three types of group sparse signal model were represented on the basis of the physical properties of the sound field. In addition, a sparse decomposition algorithm was derived to address these signal models as an extension of M-FOCUSS. Using these group sparse signal models, we significantly improved F-measure in the decomposition stage compared with the case of using M-FOCUSS. Furthermore, the reproduction accuracy was higher than that of current methods, especially at frequencies above the spatial Nyquist frequency.

7. REFERENCES

[1] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2764–2778, 1993.

[2] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *Proc. 124th AES Conv.*, Amsterdam, Oct. 2008.

[3] 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, 2013.

[4] D. de Vries, *Wave Field Synthesis*, AES Monograph. Audio Eng. Soc., 2009.

[5] 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. (ICASSP)*, Florence, May 2014, pp. 4476–4480.

[6] 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, 2005.

[7] J. Ahrens and S. Spors, "An analytical approach to local sound field synthesis using linear arrays of loudspeakers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, May 2011.

[8] 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. (WASPAA)*, New Paltz, 2011, pp. 1–4.

[9] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "MAP estimation of driving signals of loudspeakers for sound field reproduction from pressure measurements," in *Proc. IEEE Int. Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, 2013.

[10] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, New York, 1999.

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

[12] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. Part I: greedy pursuit," *Signal Process.*, vol. 86, pp. 572–588, 2006.

[13] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, 2007.

[14] A. Rakotomamonjy, "Surveying and computing simultaneous sparse approximation (or group-lasso) algorithms," *Signal Process.*, vol. 91, pp. 1505–1526, 2011.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[16] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[17] J. Le Roux, P. T. Boufounos, K. Kang, and J. R. Hershey, "Source localization in reverberant environments using sparse optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 4310–4314.

[18] 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. (ICASSP)*, Florence, May 2014, pp. 1453–1457.

[19] K. Helwani, H. Buchner, and S. Spors, "Multichannel adaptive filtering with sparseness constraints," in *Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Sep. 2012.