SPATIO-TEMPORAL SPARSE SOUND FIELD DECOMPOSITION CONSIDERING ACOUSTIC SOURCE SIGNAL CHARACTERISTICS

Naoki Murata, Shoichi Koyama, Norihiro Takamune, Hiroshi Saruwatari

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

ABSTRACT

We propose a sound field decomposition method that takes into consideration spatio-temporal sparsity. It has been proved that sparse representation of a sound field is effective in reducing errors originating from spatial aliasing artifacts compared with conventional plane wave decomposition. In most current methods of sparse sound field decomposition, the spatial sparsity of the sound source distribution is only assumed. However, it is known that the temporal structure of the source signal to be decomposed can also be sparse in the time-frequency domain. We formulate an objective function for sparse sound field decomposition by using the $\ell_{p,q}$-norm to simultaneously induce sparsity in the space and time domains. An optimization algorithm on the auxiliary function method is derived to solve it. Numerical simulations of acoustic holography indicate that the reconstruction accuracy can be improved by controlling the parameter of temporal sparsity. We also demonstrate that a statistical measure of the source signals can be used as an indicator to determine a nearly optimal parameter.

Index Terms— Sound field decomposition, sparse signal representation, spatio-temporal sparsity, auxiliary function method

1. INTRODUCTION

Various applications of acoustic signal processing, such as sound field analysis, visualization, and reproduction, are founded on sound field decomposition. The objective of sound field decomposition is to represent a sound field as a linear combination of fundamental solutions of the Helmholtz equation using signals received by multiple microphones. This representation makes it possible to estimate the entire sound field from pressure measurements; therefore, it can be used for various applications including the acoustic inverse problem.

Plane-wave decomposition has been commonly used for sound field decomposition because of its computational efficiency. This method corresponds to spatial Fourier analysis of the sound field [1]. Acoustic holography is used to estimate the pressure or velocity distribution in the inverse direction of sound propagation. The method based on spatial Fourier analysis is referred to as near-field acoustic holography (NAH) [1–3]. Sound field recording and reproduction have also been achieved with this representation [4–6], which is applied to high-fidelity audio systems. Even though these methods make it possible for efficient and stable signal processing to be carried out using the fast Fourier transform (FFT), a critical issue arises from artifacts originating from spatial aliasing. This artifacts cause significant deterioration of the decomposition accuracy above the spatial Nyquist frequency, which depends on the interelement spacing in the microphone array. For example, in sound field recording and reproduction, listeners are unable to localize the reproduced sound images and the frequency characteristics of the reproduced sound are adversely affected.

In recent years, the sparse representation of a sound field has been proved to be effective in reducing spatial aliasing artifacts [7,8]. Sparse sound field decomposition is generally based on the assumption that a distribution of sound sources is spatially sparse. NAH based on a sparse representation enables sound field imaging with high resolution even at high frequencies [7]. The use of sparse sound field decomposition for recording and reproduction makes it possible to reproduce the sound field above the spatial Nyquist frequency [8–10].

As discussed above, most current methods of sparse sound field decomposition are only based on the spatial sparsity of the sound sources. However, it is known that sparsity also appears in the temporal structure of source signals to be decomposed in the time-frequency domain, which is typically used in blind source separation problems [11, 12]. We propose a sound field decomposition method that takes into consideration the sparsity in the space and time domains. We derive an objective function incorporating the $\ell_{p,q}$-norm ($0 < p \leq q \leq 2$) of a matrix of the source signals as a penalty term [13, 14]. The auxiliary function method [15–17] is applied to obtain an optimization algorithm. Obviously, the sparsity in the time domain is less strong than that in the space domain; therefore, we also discuss a method of adjusting a parameter to control the temporal sparsity using prior information on statistical measure of the source signals. Numerical simulations are conducted to evaluate the proposed method for the acoustic holography problem.

There have been few previous works on sparse sound field decomposition using prior information on the temporal structure of source signals. We previously proposed a sparse sound field decomposition method incorporating a complex non-negative matrix factorization (NMF) model [18]. This method assumed that a large data set of source signals to be decomposed is available for training because it is inherently necessary to train all the possible spectrum structures of the source signals. For the multitask learning problem, Rakotomamonjy et al. proposed an algorithm for $\ell_{p,q}$-norm minimization [14]; however, our proposed algorithm can treat a wider range of the penalty parameter, which is discussed in Sec. 3.2.

2. SPARSE SOUND FIELD DECOMPOSITION

2.1. Generative model of sound field

First, we briefly revisit the generative model of a sound field proposed in [8]. As shown in Fig. 1, the sound field is divided into internal and external regions of a closed surface. The internal region is denoted as Ω. We assume that the sound field consists of monopole-source and plane-wave components and that the monopole components exist only inside Ω. The sound pressure distribution is obtained by placing microphones on the receiving plane Γ. By denoting the position vector and frequency as $\mathbf{r}$ and ω, respectively, the spatial distribution of the monopole components inside Ω is represented as $Q(\mathbf{r}, \omega)$ ($\mathbf{r} \in \Omega$). Therefore, the sound pressure at $\mathbf{r}$, $p(\mathbf{r}, \omega)$, can be

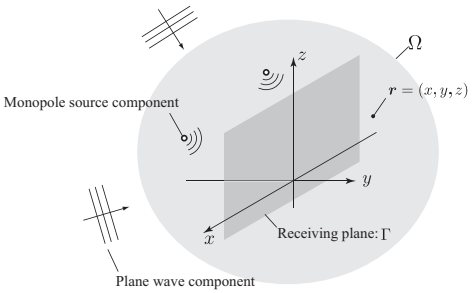

Fig. 1. Generative model of sound field

represented as the sum of inhomogeneous and homogeneous terms, $p_{\mathrm{i}}(\mathbf{r},\omega)$ and $p_{\mathrm{h}}(\mathbf{r},\omega)$, respectively, as

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

where $G(\mathbf{r}|\mathbf{r}',\omega)$ is the three-dimensional free-field Green's function that corresponds to the transfer function of the monopole sources:

$$G(\mathbf{r}|\mathbf{r}',\omega) = \frac{\exp(-j\frac{\omega}{c}\|\mathbf{r}-\mathbf{r}'\|_2)}{4\pi\|\mathbf{r}-\mathbf{r}'\|_2}. \tag{2}$$

Here, c is the velocity of sound. Hereafter, the temporal frequency ω is omitted for notational simplicity. The objective of the sound field decomposition is to decompose the sound field into $Q(\mathbf{r})$ and $p_{\mathrm{h}}(\mathbf{r})$ as in (1), by using the pressure measurements $p(\mathbf{r})$ ($\mathbf{r}\in\Gamma$).

2.2. Sound field decomposition based on spatial sparsity of source distribution

By discretizing Ω, (1) can be treated as the sparse representation problem. First, the region Ω is discretized as a set of grid points and their number is denoted as N. M microphones are assumed to be arranged on Γ. We assume $N \gg M$ since the grid points should entirely and densely cover the region Ω. Then, the discrete form of (1) can be represented as

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{z}, \tag{3}$$

where $\mathbf{y}\in\mathbb{C}^M$ comprises the signals received by the microphones, $\mathbf{x}\in\mathbb{C}^N$ is the distribution of the monopole components at the grid points, $\mathbf{z}\in\mathbb{C}^M$ is the homogeneous term of the received signals, i.e., the plane-wave components, and $\mathbf{D}\in\mathbb{C}^{M\times N}$ is the dictionary matrix of the monopole components whose elements consist of the Green's function (2) between the grid points and microphones. Therefore, the sound field decomposition problem becomes the estimation of $\mathbf{x}$ and $\mathbf{z}$ when $\mathbf{y}$ and $\mathbf{D}$ are given. We assume that $\mathbf{D}\mathbf{x}$ is the dominant component of $\mathbf{y}$ and $\mathbf{z}$ is the residual. Although the linear equation (3) is an underdetermined problem, a small number of elements in $\mathbf{x}$ will have nonzero values because the number of monopole components inside Ω should be sufficiently smaller than the number of grid points. Therefore, the sound field decomposition can be achieved by obtaining a sparse solution of (3).

Although (3) deals with a model in a single time frame and a single frequency bin, several group-sparse models based on physical properties can be introduced for more accurate and robust decomposition [9]. We here assume that the observations of multiple time frames are available and that the source locations are static during the observations, which is typically referred to as the multiple measurement vector (MMV) problem or simultaneous sparse approximation problem [8, 13, 19–21]. By denoting the signals at each time frame t ($t\in\{1,\cdots,T\}$) as $\mathbf{y}_t$, $\mathbf{x}_t$, and $\mathbf{z}_t$, the matrices $\mathbf{Y}\in\mathbb{C}^{M\times T}$, $\mathbf{X}\in\mathbb{C}^{N\times T}$, and $\mathbf{Z}\in\mathbb{C}^{M\times T}$ whose columns respectively consist of $\mathbf{y}_t$, $\mathbf{x}_t$, and $\mathbf{z}_t$ can be defined. Then, the signal model (3) can be represented in matrix form as

$$\mathbf{Y} = \mathbf{D}\mathbf{X} + \mathbf{Z}. \tag{4}$$

Under the assumption of static source locations, each column of $\mathbf{X}$ has nonzero values at the same positions; therefore, $\mathbf{X}$ can be assumed to be sparse in terms of its rows. The row-sparse solution of $\mathbf{X}$ can be obtained by solving the following optimization problem [13]:

$$\underset{\mathbf{X}}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \lambda J(\mathbf{X}) \right\}, \tag{5}$$

where $J(\mathbf{X})$ is a penalty term for inducing the row-sparsity of $\mathbf{X}$ and λ is a parameter that balances the approximation error and the penalty term. Generally, $J(\mathbf{X})$ is defined as follows:

$$J(\mathbf{X}) = \sum_n \left(\sum_t |x_{n,t}|^2 \right)^{p/2}, \tag{6}$$

where $x_{n,t}$ is the (n,t)th element of $\mathbf{X}$. In our previous studies [8,9], an algorithm called M-FOCUSS [19] or its extension [9] was applied to achieve sparse sound field decomposition based on (4).

3. SPATIO-TEMPORAL SPARSE SOUND FIELD DECOMPOSITION

In (5), sparsity is only imposed on the spatial source distribution; however, the temporal structure is not assumed to be sparse. By using the $\ell_{p,2}$-norm penalty term (6), each activated row, i.e., the time sequence of the source signal, is estimated in a least-square-error sense. However, it is known that many acoustic source signals are sparse in their time sequences in the time-frequency domain, and this fact is typically exploited in blind source separation problems [11, 12]. Therefore, by taking into consideration the spatio-temporal sparsity of $\mathbf{X}$ in the sound field decomposition, it will be possible to increase the decomposition accuracy. Indeed, the temporal structure, i.e., the rows of $\mathbf{X}$, is less sparse than the spatial structure, i.e., the columns of $\mathbf{X}$. Therefore, it will be useful if the column-sparsity of $\mathbf{X}$ can be controlled to obtain an estimate of $\mathbf{X}$.

We previously proposed a sound field decomposition method using time-frequency spectrum patterns trained in advance [18], which was derived by incorporating the complex NMF model [16] into the monopole components. This method is useful when a large data set of source signals is available in advance because all the possible spectrum patterns are necessary for training in principle; however, this will be difficult to achieve in some situations. Therefore, a method using approximate information on the temporal structure of the source signals is required rather than detailed spectrum patterns trained in advance.

3.1. Sound field decomposition based on $\ell_{p,q}$-norm minimization

To achieve spatio-temporal sparse sound field decomposition, we consider the following $\ell_{p,q}$-norm minimization problem:

$$\underset{\mathbf{X}}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \lambda J_{p,q}(\mathbf{X}) \right\}, \tag{7}$$

Algorithm 1 Proposed spatio-temporal sparse sound field decomposition algorithm

Initialize $\mathbf{X}^{(0)}$, $l = 0$
while loop $\neq 0$ **do**
  $\xi_n^{(l)} = \sum_t |x_{n,t}^{(l)}|^q$
  $\eta_{n,t}^{(l)} = |x_{n,t}^{(l)}|$
  for $t = 1$ to T **do**
    $\mathbf{W}_t^{(l)} \leftarrow \mathrm{diag}\left(p^{-1/2}(\xi_n^{(l)})^{1/2-p/2q}(\eta_{n,t}^{(l)})^{1-q/2}\right)$
    $\mathbf{A}_t^{(l)} \leftarrow \mathbf{D}\mathbf{W}_t^{(l)}$
    $\mathbf{x}_t^{(l+1)} \leftarrow \mathbf{W}_t^{(l)}(\mathbf{A}_t^{(l)})^{\mathsf{H}}\left(\mathbf{A}_t^{(l)}(\mathbf{A}_t^{(l)})^{\mathsf{H}} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}_t$
  end for
  $l \leftarrow l + 1$
  if stopping condition is satisfied **then**
    loop $= 0$
  end if
end while

where

$$J_{p,q}(\mathbf{X}) = \sum_n \left(\sum_t |x_{n,t}|^q\right)^{p/q}. \tag{8}$$

Here, p and q ($p,\ q > 0$) denote parameters used to control the sparsity in the space and time domains, i.e., in the columns and rows, respectively. When $q = 2$, the penalty term (8) corresponds to that for the MMV problem (6). By setting $q < 2$, this penalty term induces sparsity in the rows of $\mathbf{X}$ as well as in the columns. Then, a spatio-temporal sparse solution of $\mathbf{X}$ can be obtained by solving (7). The spatial sparsity can be assumed to be very strong because it corresponds to the spatial distribution of the monopole components. Therefore, a smaller p will be preferable as long as a local minimum due to non-convexity is avoidable. On the other hand, the temporal sparsity will not be so strong compared with the spatial sparsity. Therefore, it will be useful to adjust q according to prior information on the sparsity of the source signal to be decomposed. A method for choosing q is discussed in Sec. 4.

3.2. Optimization algorithm based on auxiliary function method

We derive an optimization algorithm based on the auxiliary function method [15–17] that gives stable and fast update rules. First, we develop an auxiliary function of the objective function. The penalty term in (7) is concave with respect to $\sum_t |x_{n,t}|^q$ for $0 < p/q \leq 1$. Since a concave function lies below its tangent line, the following inequality can be obtained:

$$\begin{aligned} J_{p,q}(\mathbf{X}) &= \sum_n \left(\sum_t |x_{n,t}|^q\right)^{p/q} \\ &\leq \sum_n \frac{p}{q}\xi_n^{p/q-1}\left(\sum_t |x_{n,t}|^q - \xi_n\right) + \xi_n^{p/q}, \end{aligned} \tag{9}$$

where $\xi_n \geq 0$ is an auxiliary variable that corresponds to the tangent point. The equality holds for $\xi_n = \sum_t |x_{n,t}|^q$. Next, for $q < 2$, (9) can be bounded as follows by using a quadratic function whose axis is on the line $x_{n,t} = 0$:

$$\begin{aligned} &\sum_n \frac{p}{q}\xi_n^{p/q-1}\left(\sum_t |x_{n,t}|^q - \xi_n\right) + \xi_n^{p/q} \\ &\leq \sum_n \frac{p}{q}\xi_n^{p/q-1}\left(\sum_t \frac{q}{2}\eta_{n,t}^{q-2}|x_{n,t}|^2 + \left(1 - \frac{q}{2}\right)\eta_{n,t}^q - \xi_n\right) \\ &\quad + \xi_n^{p/q} \\ &= J_{p,q}^+(\mathbf{X}, \mathbf{\Theta}), \end{aligned} \tag{10}$$

where $\eta_{n,t} \geq 0$ is an auxiliary variable and the equality holds for $\eta_{n,t} = |x_{n,t}|$. We hereafter denote the set of auxiliary variables, i.e., $\{\xi_n\}$ and $\{\eta_{n,t}\}$, as $\mathbf{\Theta}$. Therefore, the auxiliary function for the objective function in (7) can be obtained as

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda J_{p,q}^+(\mathbf{X}, \mathbf{\Theta}). \tag{11}$$

On the basis of the principle of the auxiliary function method, the objective function in (7) monotonically decreases upon iteratively minimizing (11) with respect to $\mathbf{X}$ and $\mathbf{\Theta}$. In addition, this algorithm corresponds to M-FOCUSS upon setting $q = 2$. The proposed algorithm for spatio-temporal sparse sound field decomposition is summarized in Algorithm 1.

In the context of the multitask learning problem, Rakotomamonjy et al. applied an algorithm for $\ell_{p,q}$-norm minimization [13, 14]. This algorithm is a type of iteratively reweighted least-squares algorithm; therefore, it is similar to the proposed algorithm. In this algorithm, however, no direct solutions are given for (7); instead, the relaxation problem of (7) is employed by raising (8) to the power of $2/p$. As a result, Rakotomamonjy's algorithm is derived for $q \geq 1$ because the relaxation problem is equivalent to the original problem only when the penalty term is convex, i.e., $p, q \geq 1$. On the other hand, our proposed algorithm can be applied for $0 < p \leq q \leq 2$, even when $q < 1$, owing to the direct formulation of the relaxation problem using the auxiliary function method.

4. EXPERIMENT

Numerical simulations were conducted to evaluate the proposed method for the acoustic holography problem under the free-field assumption. In Cartesian coordinates, omni directional microphones were linearly aligned along the x-axis with its center at the origin. The number of microphones was 32 and the interval between them was 0.12 m. The region Ω was set as a rectangular region of 4.0×3.0 m^2 on the x-y plane centered at (0.0, –1.5, 0.0) m. This region was discretized into grid points with intervals of 0.1 m along both the x- and y-axes. A single point source was located at (-7.4×10^{-1}, -7.2×10^{-1}, 0.0) m. The source signal was a single-frequency sinusoidal wave but its complex amplitude was extracted from speech. First, the short-time Fourier transform (STFT) was performed to obtain time-frequency spectrograms of the speech signals. The sampling frequency of the speech signals was 16 kHz and a square-root Hanning window of 32 ms length with a 16 ms overlap was used in the STFT. Since the duration of the speech signal ranged from 2.7 to 7.7 s, the number of time frames was between 167 and 480. Next, we chose the frequency bin between 1.8 and 2.2 kHz with the highest power, then its amplitude sequence was applied to the simulated source signal of a sinusoidal wave at 2.0 kHz. The speech signals were extracted from a Japanese speech database (RWCP-SP99) [22], and included three female (Speakers F #1–3) and three male (Speakers M #1–3) utterances. The number of utterances per speaker was 10; therefore, we used 60 speech signals

in total. Gaussian noise was added to the signals received by the microphones so that the signal-to-noise ratio was 20 dB.

We evaluated the efficacy of adjusting the parameter for controlling the temporal sparsity q in the estimation of the pressure distribution on the line $y = -0.1$ m, which was defined as the reconstruction line. The length of the reconstruction line was 4.0 m and its center was at (0, –0.10) m. By discretizing the reconstruction line, 401 evaluation points at intervals of 0.01 m were obtained. For quantitative evaluation, we define the signal-to-distortion ratio (SDR) as

$$\mathrm{SDR} = 10 \log \frac{\sum_{i,t} |P_{\mathrm{true}}(i,t)|^2}{\sum_{i,t} |P_{\mathrm{true}}(i,t) - P_{\mathrm{est}}(i,t)|^2}, \qquad (12)$$

where $P_{\mathrm{true}}(i,t)$ and $P_{\mathrm{est}}(i,t)$ are the true pressure and the pressure estimated using the decomposition result in the time-frequency domain, respectively. Here, i and t respectively denote the indexes of the evaluation points and time frames.

p was here fixed at 1.0 and q was changed from 1.0 to 2.0 at intervals of 0.05. Note that the proposed algorithm corresponds to M-FOCUSS for $q = 2$, which is the method used in [8]. The balancing parameter λ was chosen by using the golden section search method [23] so that the highest SDR was obtained at each q. Figure 2 shows the relationship between q and the SDR for three types of utterance of four speakers (12 utterances in total). Each line style and color represents a speaker and utterance, respectively. The result when the source signal was generated by the Gaussian distribution is also shown (Gaussian). In all the results for the speech signal, including the utterances and speakers not shown in Fig. 2, the highest SDR was achieved at $q < 2$. On the other hand, $q = 2$ gave the best results for the signals generated by the Gaussian distribution. The maximum improvement in the SDR of the speech signal was 0.13 dB and the corresponding value of q was 1.6 (the blue line of Speaker M #1 in Fig. 2). Therefore, controlling the parameter of the temporal sparsity q is effective for increasing the reconstruction accuracy when the source signal to be decomposed is speech.

Although it is not a trivial task to optimize q in practice, it will be useful if rough information such as a statistical measure of the source signal can be used as an indicator to choose a nearly optimal q. It is difficult to derive an analytical relationship between q and a statistical measure; therefore, we experimentally investigated the relationship between the optimal q and the kurtosis of the speech signal, which is a statistical measure of super-Gaussianity of data. Figure 3 is a plot of the inverse of the kurtosis obtained from the amplitude of the speech and the parameter q giving the highest SDR. We obtained the kurtosis from a sequence of absolute values of the source signal. All the results for utterances and speakers are shown and each symbol indicates a speaker. Since it is possible to generate artificial time sequences with various kurtoses by using the gamma distribution, the results for source signals generated by the gamma distribution are also shown (Gamma). By setting the shape parameter of the gamma distribution, we obtained time sequences that have kurtosis from 10 to 120 at intervals of 10. Note that the shape parameter κ corresponds to a kurtosis of $6/\kappa$ [24]. The length of the artificial signals was 200 frames and three time sequences were generated at each kurtosis. One can observe an approximately linear relationship between the inverse of the kurtosis of the source signal and the optimal q. When the inverse of the kurtosis was large, i.e., the super-Gaussianity was weak, the optimal q became large. In contrast, when the inverse of the kurtosis was small, i.e., the super-Gaussianity was strong, the optimal q became small. This result suggests that the kurtosis of the source signal to be decomposed can be used as an indicator to determine a nearly optimal q.

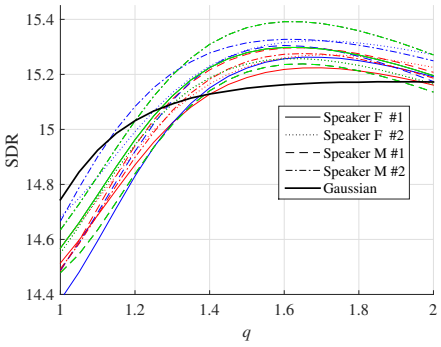

Fig. 2. Relationship between q and SDR

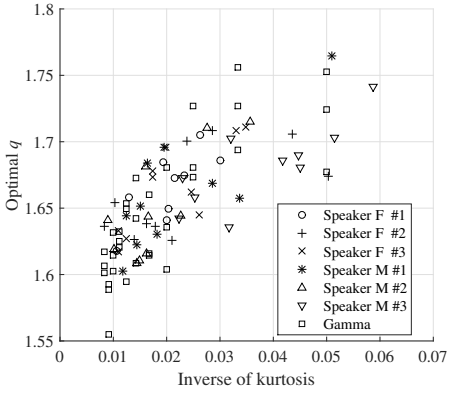

Fig. 3. Relationship between inverse of kurtosis of source signal and optimal q

5. CONCLUSION

A sound field decomposition method that takes into consideration spatio-temporal sparsity was proposed. The spatial sparsity of the sound source distribution has only been considered in sparse sound field decomposition methods even though it is known that sparsity also appears in the temporal structure of source signals to be decomposed in the time-frequency domain. We formulated an objective function for sparse sound field decomposition by using the $\ell_{p,q}$-norm to simultaneously induce sparsity in the space and time domains. The proposed algorithm was based on the auxiliary function method. In numerical simulations of acoustic holography, the reconstruction accuracy was improved when q was set at a smaller value than 2, which means that it is effective to control the parameter of the temporal sparsity. In addition, we demonstrated that the inverse of the kurtosis of the source signal can be used as an indicator to determine a nearly optimal q. A future work will be to develop a method for sound field decomposition based on sparsity in the time-frequency and space domains.

6. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP15H05312 and SECOM Science and Technology Foundation.

7. REFERENCES

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

[2] J. 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.

[3] J. Hald, "Basic theory and properties of statistically optimized near-field acoustical holography," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2105–2120, 2009.

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

[5] 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.

[6] S. Koyama, K. Furuya, Y. Hiwasaki, Y. Haneda, and Y. Suzuki, "Wave field reconstruction filtering in cylindrical harmonic domain for with-height recording and reproduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1546–1557, 2014.

[7] 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.

[8] 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. 4443–4447.

[9] S. Koyama, N. Murata, and H. Saruwatari, "Structured sparse signal models and decomposition algorithm for super-resolution in sound field recording and reproduction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brisbane, Apr. 2015, pp. 619–623.

[10] 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. (ICASSP)*, Shanghai, Mar. 2016, pp. 345–349.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[12] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, Oct. 2007, pp. 147–150.

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

[14] A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu, "$\ell_p - \ell_q$ penalty for sparse linear and sparse multiple kernel multitask learning," *IEEE Trans. Neural Netw.*, vol. 22, no. 8, pp. 1307–1320, 2011.

[15] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization,," in *Proc. Advances in Neural Inf. Process. Systems (NIPS)*, Vancouver, Dec. 2001, pp. 556–562.

[16] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2009, pp. 3437–3440.

[17] M. Nakano, H. Kameoka, J. L. Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. IEEE Int. Workshop Machine Learn. Signal Process. (MLSP)*, Kittilä, Aug. 2010, pp. 283–288.

[18] N. Murata, S. Koyama, H. Kameoka, N. Takamune, and H. Saruwatari, "Sparse sound field decomposition with multichannel extension of complex NMF," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Shanghai, Mar. 2016, pp. 395–399.

[19] 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.

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

[21] 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.

[22] RWCP Speech Resources Consortium, "Japanese speech database (RWCP-SP99)," `http://research.nii.ac.jp/src/RWCP-SP99.html` [accessed 1 Sep. 2016].

[23] W. H. Press, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge University Press, New York, 2007.

[24] C. Walck, *Handbook on Statistical Distributions for Experimentalists*, University of Stockholm Internal Report SUF-PFY/96-01, available from www. physto. se/˜ walck, 2007.