SPARSE SOUND FIELD DECOMPOSITION WITH MULTICHANNEL EXTENSION OF COMPLEX NMF

[1]*Naoki Murata,* [1]*Shoichi Koyama,* [1,2]*Hirokazu Kameoka,* [1]*Norihiro Takamune, and* [1]*Hiroshi Saruwatari*

[1]Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[2]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation,
3-1 Morinosato Wakamiya, Atsugi-shi, Kanagawa 243-0198, Japan

ABSTRACT

A sparse sound field decomposition method using prior information on source signals in the time-frequency domain is proposed. Sparse sound field decomposition has been proved to be effective for various acoustic signal processing applications. Current methods for sparse decomposition are based only on the spatial sparsity of the source distribution. However, it can be assumed that possible source signals to be decomposed are approximately known in advance. To exploit this prior information, we incorporated the complex nonnegative factorization model into sparse sound field decomposition. Since the magnitude spectrum of the possible source signals can be trained in advance, accuracy of the sparse decomposition can be improved even when the source signals are highly correlated and the sources are in a highly noisy environment. In addition, the proposed decomposition algorithm is derived using the auxiliary function method. Numerical experiments indicated that the sparse decomposition performance was significantly improved using the proposed method.

Index Terms— Sound field decomposition, source localization, sparse representation, complex nonnegative matrix factorization, auxiliary function method

1. INTRODUCTION

Sound field decomposition forms the foundation of various acoustic signal processing applications, such as sound field analysis, reconstruction, and visualization. Its aim is to represent a sound field as a linear combination of fundamental solutions of the wave equation (or Helmholtz equation) from the results of the pressure measurements. Using this representation, the entire sound field can be estimated from the signals received by multiple microphones. It can be considered that this problem involves source localization or direction of arrival (DOA) estimation because the power distribution of the pressure field indicates the locations or directions of the sound sources.

A conventional method for sound field decomposition is based on spatial Fourier analysis of the sound field, which corresponds to plane wave decomposition [1]. In acoustic holography, the method based on spatial Fourier analysis is known as near-field acoustic holography (NAH). NAH is used to measure the pressure or velocity distribution on a surface close to acoustic sources. In the context of sound field recording and reproduction, which is targeted at high-fidelity audio systems, a method of converting signals received by microphones into driving signals of loudspeakers for reproduction is necessary. The wave field reconstruction filtering method is an efficient and stable signal conversion method based on spatial Fourier analysis [2, 3]. Higher-order Ambisonics [4, 5] also enables signal conversion by encoding and decoding processes in the spherical harmonic domain.

Source localization and DOA estimation have been developed in a different context from sound field decomposition [6, 7]. Delay and sum (DAS) is one of the most classical approaches [6]. Subspace methods, such as the multiple signal classification (MUSIC) method [8], are also well-established methods based on the assumption of orthogonality of signal and noise subspaces. Since signal reconstruction is not necessary in source localization, a wide range of methods have been proposed compared with for sound field decomposition.

In recent years, the sparse representation of a sound field has attracted interest in various research fields involving acoustic signal processing. Sparse sound field decomposition has been applied to acoustic holography [9] and sound field recording and reproduction [10, 11] to improve their spatial resolution limits. Several methods for source localization based on sparse representation have been proposed [12–14], which also make higher spatial resolution possible. Generally, these methods are formulated in a common framework of an overcomplete linear equation of the observations and are based on an assumption of a spatially sparse distribution of sound sources. Therefore, improving the performance of sparse sound field representation has important implications for various applications.

In current methods for sparse sound field decomposition, no assumptions are imposed on the structure of the source signals in the time-frequency domain. However, in some applications, it can be assumed that the source signals of interest are known in advance. For instance, speech signals are the main target of sparse sound field decomposition in telecommunication systems. This assumption is widely used in the blind source separation (BSS) problem. In nonnegative matrix factorization (NMF) [15], which is one of the well-known methods for BSS, a time-frequency spectrogram of the source signal is approximated as a product of two nonnegative low-rank matrices. This means that acoustic source signals are represented by a limited number of spectrogram components.

We propose a sparse sound field decomposition method that takes the time-frequency-spectrum structures of the source signals into consideration. The proposed method is derived by incorporating complex NMF (CNMF) [16] into sparse sound field decomposition. The source signal is modeled as a sum of products of a rank 1 magnitude spectrogram and its phase spectrogram. Therefore, a complex spectrogram of the source signals can be incorporated into sparse decomposition without transformation into nonnegative values. The decomposition algorithm is obtained by an auxiliary function approach. By exploiting time-frequency-spectrum structures trained in advance, it is possible to improve the decomposition accuracy even when the source signals are highly correlated and the sources are in a highly noisy environment. Numerical simulations are conducted to evaluate the proposed method in terms of DOA estimation and signal reconstruction.

2. SPARSE SOUND FIELD DECOMPOSITION

First, we formulate a signal model for sparse sound field decomposition and its optimization criterion for current methods in a generalized form [9–14]. Figure 1 shows this signal model in the special case of a linear microphone array and plane wave decomposition. The number of sources is assumed to be small in the region of interest. We consider M microphones receiving signals from I sources. The locations (or directions) of the sources are denoted by position vectors in an arbitrary number of dimensions $\mathbf{v}_i$ ($i \in \{1, \cdots, I\}$). A model of observed signals in the time-frequency domain at a single time frame can be represented as

$$\mathbf{y}(\omega) = \mathbf{A}(\omega)\mathbf{s}(\omega) + \mathbf{z}(\omega), \tag{1}$$

where ω denotes the frequency, $\mathbf{y}(\omega) \in \mathbb{C}^M$ and $\mathbf{s}(\omega) \in \mathbb{C}^I$ respectively denote the signals received by the microphones and the complex amplitudes of the source signals, and $\mathbf{z}(\omega) \in \mathbb{C}^M$ is an unknown noise vector. Hereafter, we omit ω for notational simplicity. The quantity

$$\mathbf{A} = [\mathbf{d}(\mathbf{v}_1)\, \mathbf{d}(\mathbf{v}_2)\, \cdots\, \mathbf{d}(\mathbf{v}_I)] \tag{2}$$

is referred to as a steering matrix, where each column $\mathbf{d}(\mathbf{v}_i) \in \mathbb{C}^M$ is a vector whose elements consist of a fundamental solution of the Helmholtz equation defined by the locations of the sources and microphones, such as the Green's function and plane wave function. In the source localization problem, arbitrary transfer functions can be used for $\mathbf{A}$. Our objective is to estimate both the parameters $\mathbf{v}_i$ and $\mathbf{s}$. Note that only $\mathbf{v}_i$ is necessary in the source localization.

To solve (1) as a sparse representation problem, we generalize the steering matrix $\mathbf{A}$ as an overcomplete dictionary matrix $\mathbf{D}$ comprising all possible source locations $\hat{\mathbf{v}}_n$ ($n \in \{1, \cdots, N\}$), which are referred to as *grids*, such that

$$\mathbf{D} = [\mathbf{d}(\hat{\mathbf{v}}_1)\, \mathbf{d}(\hat{\mathbf{v}}_2)\, \cdots\, \mathbf{d}(\hat{\mathbf{v}}_N)], \tag{3}$$

where N denotes the number of grids and both $N \gg I$ and $N \gg M$ are assumed. The source vector $\mathbf{s}$ can be accordingly modified to $\mathbf{x} \in \mathbb{C}^N$, where the nth element x_n has a nonzero value and equals s_i when the ith source originates from $\hat{\mathbf{v}}_n$, and is zero otherwise. Using $\mathbf{D}$ and $\mathbf{x}$, (1) can be reformulated as

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{z}. \tag{4}$$

When the source distribution is spatially sparse, $\mathbf{x}$ becomes a sparse vector, i.e., $\mathbf{x}$ has a few nonzero elements. Therefore, sparse decomposition algorithms can be applied to estimate $\mathbf{x}$ from $\mathbf{y}$ [17].

The optimization criterion for the sparse sound field decomposition can be formulated as

$$\underset{\mathbf{x}}{\text{minimize}} \left\{ \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \eta J(\mathbf{x}) \right\}, \tag{5}$$

where $J(\mathbf{x})$ is a penalty term for inducing the sparsity of $\mathbf{x}$ and η is a parameter that balances the approximation error and the penalty $J(\mathbf{x})$ [17]. In general, the ℓ_p-norm of $\mathbf{x}$ is used as $J(\mathbf{x})$,

$$J^p(\mathbf{x}) = \|\mathbf{x}\|_p = \left(\sum_{n=1}^{N} |x_n|^p \right)^{1/p}, \tag{6}$$

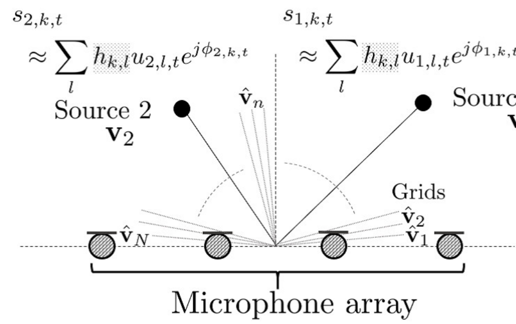

Fig. 1. Signal model for plane wave decomposition using linear array of microphones.

where $0 < p \leq 1$. When observations of multiple time frames are available and the source locations can be assumed to be static during the observations, a matrix in which each column is an observation is constructed, and a row-sparse penalty term of the matrix is usually used for decomposition, such as the $\ell_{p/q}$-norm ($q \geq 2$) [11, 18, 19]. This problem is generally referred to as the multiple measurement vector (MMV) problem.

3. PROPOSED SIGNAL MODEL AND OBJECTIVE FUNCTION

In (4) and (5), no assumptions are imposed on $\mathbf{x}$ for the temporal structure; therefore, the signal decomposition is based only on the spatial sparsity of the sources. Although it is assumed that $\mathbf{x}$ for multiple time frames has the same sparsity pattern in the MMV problem, a time-frequency-spectrum structure that is unique in wide-band acoustic signals is not exploited.

Kameoka *et al.* proposed CNMF [16], in which a complex spectrogram in the time-frequency domain is represented as a sum of products of a rank 1 magnitude spectrogram and its phase spectrogram. We consider now to incorporate the signal model of CNMF into sparse sound field decomposition since the complex spectrogram of the observed signals is necessary to model signals received by multiple microphones. In CNMF, a complex spectrogram in the time-frequency domain is modeled as

$$f_{k,t} \approx \sum_l h_{k,l} u_{l,t} e^{j\phi_{k,l,t}}, \tag{7}$$

where k, t, and l respectively denote the frequency, time frame, and basis index, $f_{k,t}$ is the complex spectrogram of a source signal, $h_{k,l}$ is a static magnitude spectrum, $u_{l,t}$ is a time-varying activation coefficient, and $\phi_{n,k,t}$ is a time-varying phase spectrum. This model allows any observed complex spectrum to be accurately represented only using a few active magnitude spectrum bases, each of which is paired with an arbitrary phase spectrum.

The signal model (7) can be introduced into the source component $\mathbf{x}$ in (4). Again, n denotes the index of the grids. When $x_{n,k,t}$ represents the complex spectrogram in the time-frequency domain at each grid, $x_{n,k,t}$ can be described as

$$x_{n,k,t} \approx \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}. \tag{8}$$

We assume that the magnitude spectrum $h_{k,l}$ is static and trained in advance using possible monaural source signals. Although $\phi_{n,k,t}$ should have the basis index l when the CNMF model is straightforwardly applied, we omit this index and assume that each grid has its own phase spectrogram. Denoting the microphone index as m, the observed signal of the mth microphone, $y_{m,k,t}$, can be represented as

$$y_{m,k,t} \approx \sum_n d_{m,n,k} \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}, \tag{9}$$

where $d_{m,n,k}$ is the element of the dictionary matrix between the nth grid and the mth microphone in the kth frequency bin.

Finally, we formulate the objective function under the assumption that $d_{m,n,k}$ is known and the basis magnitude spectrum $h_{k,l}$ is trained in advance. We define the modeling error of (9) as

$$R(\mathbf{U}, \mathbf{\Phi}) = \sum_{m,k,t} |y_{m,k,t} - \sum_n d_{m,n,k} \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2, \tag{10}$$

where $\mathbf{U}$ and $\mathbf{\Phi}$ represent third-rank tensors whose elements are $u_{n,l,t}$ and $\phi_{n,k,t}$, respectively. To obtain a spatially sparse solution of $\mathbf{U}$, the following penalty term is introduced:

$$J^{p,2}(\mathbf{U}) = \sum_n \Big(\sum_{l,t} u_{n,l,t}^2 \Big)^{p/2}. \tag{11}$$

Therefore, the optimization criterion is formulated as

$$\underset{\mathbf{U},\mathbf{\Phi}}{\text{minimize}} \left\{ R(\mathbf{U}, \mathbf{\Phi}) + 2\lambda J^{p,2}(\mathbf{U}) \right\}, \tag{12}$$

where λ is a parameter used to balance the modeling error $R(\mathbf{U}, \mathbf{\Phi})$ and the penalty $J^{p,2}(\mathbf{U})$.

4. OPTIMIZATION ALGORITHM BASED ON AUXILIARY FUNCTION METHOD

We derive an optimization algorithm using the auxiliary function method [16], which can give stable and fast update rules. First, we develop an auxiliary function of the objective function using the following inequalities:

$$|y_{m,k,t} - \sum_n d_{m,n,k} \sum_l h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2 \le \sum_{n,l} \frac{|\overline{y}_{m,n,k,l,t} - d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2}{\beta_{m,n,k,l,t}}, \tag{13}$$

and

$$2(\sum_{l,t} u_{n,l,t}^2)^{p/2} \le p(\sum_{l,t} \overline{u}_{n,l,t}^2)^{p/2-1}(\sum_{l,t} u_{n,l,t}^2 - \overline{u}_{n,l,t}^2) + 2(\sum_{l,t} \overline{u}_{n,l,t}^2)^{p/2}. \tag{14}$$

Here, $\overline{y}_{m,n,k,l,t}$ and $\overline{u}_{n,l,t}$ are auxiliary variables and $\beta_{m,n,k,l,t}$ is an arbitrary constant. The inequality (13) is satisfied when $\sum_{n,l} \overline{y}_{m,n,k,l,t} = y_{m,k,t}$, and the inequality (14) is satisfied when $0 < p \le 2$, $0 \le \beta_{m,n,k,l,t} \le 1$, and $\sum_{n,l} \beta_{m,n,k,l,t} = 1$. Then, an upper bound function of $f(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}})$, $f^+(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}})$, can be defined as

$$\begin{aligned} f^+(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}}) &= \sum_{m,n,k,l,t} \frac{|\overline{y}_{m,n,k,l,t} - d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}}|^2}{\beta_{m,n,k,l,t}} \\ &+\lambda \sum_n \Big\{ p\Big(\sum_{l,t} \overline{u}_{n,l,t}^2\Big)^{p/2-1} \Big(\sum_{l,t} u_{n,l,t}^2 - \overline{u}_{n,l,t}^2\Big) \\ &+2\Big(\sum_{l,t} \overline{u}_{n,l,t}^2\Big)^{p/2} \Big\}. \end{aligned} \tag{15}$$

This function satisfies $f(\mathbf{U}, \mathbf{\Phi}) \le f^+(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}})$, and the equality is satisfied when

$$\begin{aligned} \overline{y}_{m,n,k,l,t} &= d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}} \\ &+\beta_{m,n,k,l,t} \Big(y_{m,k,t} - \sum_{n,l} d_{m,n,k} h_{k,l} u_{n,l,t} e^{j\phi_{n,k,t}} \Big) \end{aligned} \tag{16}$$

and

$$\overline{u}_{n,l,t} = u_{n,l,t}. \tag{17}$$

Therefore, the conditions required for $f^+(\mathbf{U}, \mathbf{\Phi}, \overline{\mathbf{Y}}, \overline{\mathbf{U}})$ to be an auxiliary function of $f(\mathbf{U}, \mathbf{\Phi})$ are satisfied [16]. Here, $\overline{\mathbf{Y}}$ and $\overline{\mathbf{U}}$ are tensors whose elements consist of $\overline{y}_{m,n,k,l,t}$ and $\overline{u}_{n,l,t}$, respectively.

Using the auxiliary function (15), we can derive iterative update rules for $\mathbf{U}$ and $\mathbf{\Phi}$. Although $u_{n,l,t}$ should be updated under the nonnegative constraint, we here update $u_{n,l,t}$ without this constraint. First, by solving $\partial f^+ / \partial u_{n,l,t} = 0$, we obtain the following update rule:

$$u_{n,l,t} \leftarrow \frac{\sum_{m,k} \Re[\overline{y}_{m,n,k,l,t}^* d_{m,n,k} h_{k,l} e^{j\phi_{n,k,t}}] / \beta_{m,n,k,l,t}}{\sum_{m,k} |d_{m,n,k}|^2 h_{k,l}^2 / \beta_{m,n,k,l,t} + \lambda p (\sum_{l,t} \overline{u}_{n,l,t}^2)^{p/2-1}}. \tag{18}$$

Here, $\Re[\cdot]$ denotes the operator that returns the real part of a complex number and $*$ denotes the operator that returns the complex conjugate value. To obtain the update rule for $\mathbf{\Phi}$, we use the following properties. When θ' satisfies $e^{j\theta'} = \sum_i a_i b_i^* / |\sum_i a_i b_i^*|$, θ' minimizes $\sum_i |a_i - b_i e^{j\theta}|^2$, where $a_i,\ b_i \in \mathbb{C}$ and $\theta \in \mathbb{R}$. Then, the update rule is obtained as

$$e^{j\phi_{n,k,t}} \leftarrow \frac{\sum_{m,l} \overline{y}_{m,n,k,l,t} d_{m,n,k}^* h_{k,l} u_{n,l,t} / \beta_{m,n,k,l,t}}{|\sum_{m,l} \overline{y}_{m,n,k,l,t} d_{m,n,k}^* h_{k,l} u_{n,l,t} / \beta_{m,n,k,l,t}|}. \tag{19}$$

We here set the arbitrary variable $\beta_{m,n,k,l,t}$ to

$$\beta_{m,n,k,l,t} = \frac{|d_{m,n,k}| h_{k,l} u_{n,l,t}}{\sum_{n,l} |d_{m,n,k}| h_{k,l} u_{n,l,t}}. \tag{20}$$

We summarize the proposed algorithm as follows. In the training stage, the magnitude spectrum bases $h_{k,l}$ are trained in advance using possible monaural source signals and $d_{m,n,k}$ is obtained by setting the locations of microphones and grids. In the decomposition stage, $\mathbf{U}$ and $\mathbf{\Phi}$ are first initialized. Then, $\overline{\mathbf{Y}}$, $\overline{\mathbf{U}}$, $\mathbf{U}$, and $\mathbf{\Phi}$ are iteratively updated using (16), (17), (18), and (19), respectively.

5. EXPERIMENTS

Numerical experiments on DOA estimation and signal reconstruction in terms of plane wave decomposition were conducted to evaluate the proposed method. We compared the proposed method (Proposed), M-FOCUSS [18], which is a sparse decomposition method for the MMV problem based only on the simultaneous spatial sparsity of the sources, and conventional DOA estimation methods, i.e., DAS, the Capon beamformer (CB) [7], and MUSIC.

Omnidirectional microphones were linearly aligned along the x-axis with their center at the origin. The number of microphones was four and they were arranged at intervals of 0.24 m; therefore, the array length was 0.72 m. Two plane wave sources, whose angles of arrival were randomly determined within the range of $(-90^\circ, 90^\circ)$, impinged on the microphones. Identical source signals were generated by artificial MIDI sounds, i.e., the two source signals were the same. The musical instrument used to generate the source signals as above was the oboe. The sampling frequency was 16 kHz. The short-time Fourier transform (STFT) was performed to obtain time-frequency spectrograms of the observed signals. A square-root Hanning window of 32 ms length with a 16 ms overlap was used in the STFT. The training data were two octaves of notes that covered all the notes of the source signal. They were also generated by artificial MIDI sounds. The number of bases was 25. A white Gaussian noise was added to the signals received by the microphones so that the signal-to-noise ratio became 20 dB. The dictionary matrix, i.e., the steering vector for the conventional DOA estimation methods, was generated so that the intervals between the grids were 1.0° within the range of $(-90^\circ, 90^\circ)$ as shown in Fig. 1. The angles of each grid is denoted as θ_n.

In M-FOCUSS, the parameters for the row-sparse penalty (p and q in [18]) were set to $p = 1$ and $q = 2$. The maximum number of iterations was 400. The parameter corresponding to η in (5) was set to 1.0. In Proposed, λ in (12) was set to 10^{-5} in the first five iterations, and 2.0×10^2 in the remaining 95 iterations.

To evaluate the source localization performance, we used the DOA spectrum [7], which indicates the power distribution of plane waves from each angle θ_n. For example, in Proposed, the DOA spectrum $P(\theta_n)$ was obtained as

$$P(\theta_n) = 10 \log_{10} \left(\sum_{l,t} u_{n,,l,t}^2 \right). \tag{21}$$

For a quantitative evaluation, we define the F-measure as

$$F_{\text{msr}} = 2 \frac{|\text{supp}\{P_{\text{est}}(\theta_n)\} \cap \text{supp}\{P_{\text{true}}(\theta_n)\}|}{|\text{supp}\{P_{\text{est}}(\theta_n)\}| + |\text{supp}\{P_{\text{true}}(\theta_n)\}|}, \tag{22}$$

where $P_{\text{est}}(\theta)$ and $P_{\text{true}}(\theta)$ are the estimated and true DOA spectra, respectively. The operator $\text{supp}(\cdot)$ extracts a set of indexes such that the value of each element of the solution $P(\theta_n)$ is larger than a threshold value μ,

$$\text{supp}\{P(\theta_n)\} = \{n \in \{1, ..., N\} \mid P(\theta_n) > \mu\}. \tag{23}$$

Since there exists no optimal threshold value for all the methods, we here changed μ from –15 dB to –1 dB in this experiment. The true indexes of θ_n were set to those for the two grids nearest the true angle. F_{msr} was averaged over 100 trials, i.e., 100 pairs of DOAs randomly selected.

We also evaluated the signal reconstruction performance, using the signal-to-distortion ratio (SDR) defined as

$$\text{SDR} = 10 \log_{10} \frac{\sum_{m,k,t} |y_{\text{org},m,k,t}|^2}{\sum_{m,k,t} |y_{\text{org},k,m,t} - y_{\text{est},m,k,t}|^2}, \tag{24}$$

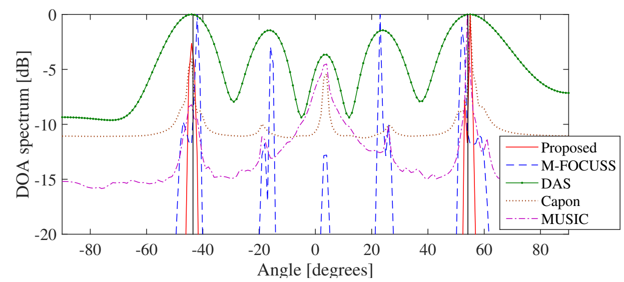

Fig. 2. Examples of DOA spectra for each method when the true angles were set to –43.5° and 54.2°.

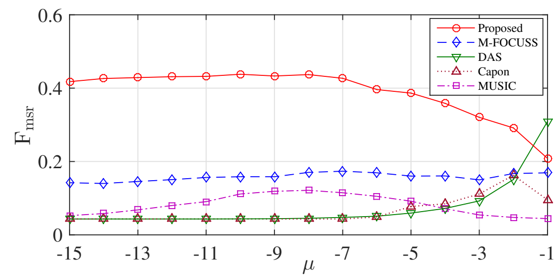

Fig. 3. Relation between F_{msr} and threshold value μ.

where $y_{\text{org},m,k,t}$ is the original signal in the observations and $y_{\text{est},m,k,t}$ is the signal in the observations reconstructed using the decomposed signals.

Figure 2 shows examples of the DOA spectra of Proposed, M-FOCUSS, DAS, CB, and MUSIC when the true angles were –43.5° and 54.2°. The DOA spectra of the conventional methods (DAS, CB, and MUSIC) had a false peak around 0°. This is because these methods were derived under the assumption that the source signals are uncorrelated. Since the source signals were identical in this experimental setup, it is very difficult to accurately estimate the true source angles. In the DOA spectra of M-FOCUSS and Proposed, the spatial resolution was significantly improved by assuming the spatial sparsity of the sources. However, the DOA spectrum of M-FOCUSS still contained multiple false peaks. In contrast, that of Proposed had two peaks at accurate angles. The average values of F_{msr} were shown in Figure 3. When the true angles were –43.5° and 54.2°, the SDRs of Proposed and M-FOCUSS were 5.27 dB and 2.13 dB, respectively. This indicates the efficacy of the proposed method, particularly for the decomposition of multiple coherent sources.

6. CONCLUSION

We proposed a sparse sound field decomposition method using prior information on source signals in the time-frequency domain. The proposed method was derived by incorporating the CNMF model of source signals into sparse sound field decomposition. Whereas current methods are based only on the spatial sparsity of the source distribution, the proposed method exploits magnitude spectrum bases pretrained using possible monaural source signals. The decomposition algorithm was derived using the auxiliary function method. Numerical simulation results indicated that the DOA estimation and signal reconstruction performance of the proposed method was significantly improved from that of current methods, particularly when the multiple source signals were highly correlated.

7. REFERENCES

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

[2] 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.

[3] 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.

[4] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonics format," in *Proc. 23rd AES Int. Conf.*, Copenhagen, May 2003.

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

[6] M. Brandstein and W. Darren, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business, 2001.

[7] V. Trees and L. Harry, *Optimum array processing: part IV of detection, estimation, and modulation*, John Wiley & Sons, 2002.

[8] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

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

[10] 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.

[11] 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.

[12] D. Malioutov, M. Çetin, 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.

[13] 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.

[14] T. Noohi, N. Epain, and C. T. Jin, "Super-resolution acoustic imaging using sparse recovery with spatial priming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2015, pp. 2414–2418.

[15] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Int. Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2003, pp. 177–180.

[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)*, Taipei, Apr. 2009, pp. 3437–3440.

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

[18] S. F. Cotter, B. 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.

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