MULTICHANNEL BLIND SOURCE SEPARATION BASED ON NON-NEGATIVE TENSOR FACTORIZATION IN WAVENUMBER DOMAIN

Yuki Mitsufuji[1, 2], Shoichi Koyama[1] and Hiroshi Saruwatari[1]

[1] The University of Tokyo, Graduate School of Information Science and Technology, Tokyo, Japan
[2] Sony Corporation, Audio Technology Development Department, Tokyo, Japan

ABSTRACT

Multichannel non-negative matrix factorization based on a spatial covariance model is one of the most promising techniques for blind source separation. However, this approach is not tractable for a large number of microphones, M, because the computational cost is of order $O(M^3)$ per time-frequency bin. To circumvent this drawback, we propose non-negative tensor factorization in the wavenumber domain, which reduces the cost to the order $O(M)$. It transforms microphone signals into the spatial frequency domain, a technique that is commonly used for soundfield reconstruction. The proposed method is compared to several blind source separation (BSS) methods in terms of separation quality and computational cost.

***Index Terms*—** Multichannel BSS, Non-negative Tensor Factorization, Wavenumber Domain, DoA, Spatial Covariance Model

1. INTRODUCTION

Non-negative matrix factorization (NMF) is one of the most prevalent techniques for blind source separation, and a number of studies have been carried out on single-channel [1–3] and multichannel scenarios [4–9]. In 2010, Ozerov and Févotte proposed a multichannel extension of NMF based on a spatial covariance model (SC-NMF), which incorporates spatial covariance matrices (SCMs) that encode the spatial positions of source signals [4]. Although the technique performs well under any type of mixing conditions, the convergence of the cost function is unstable and much slower than that of conventional NMF techniques. Sawada et al. mitigated this problem by introducing multiplicative update rules instead of using an expectation-maximization algorithm [7]. When the microphone positions are known, this technique was proved to improve separation quality because the direction of the source can be estimated from the interchannel coherence. Nikunen and Virtanen devised a direction-of-arrival (DoA)-based spatial covariance model, which provides a number of DoA kernels, namely, the outer products of steering vectors [8]. The weights of the DoA kernels are obtained by multiplicative update rules. It should be noted that the studies reported so far assume at most 3 or 4 microphones and are not suitable for a large number of microphones because the computational cost for SC-NMF is of order $O(M^3)$ per time-frequency (TF) bin.

Non-negative tensor factorization (NTF) is another promising technique for a multichannel scenario. The cost is only of order $O(M)$ [10, 11] per TF bin. The drawback is that, in contrast to SC-NMF, it cannot model interchannel phase differences; that is, only intensity level differences between microphones are taken into account as spatial properties of the observed mixture. However, thanks to its low computational cost, NTF is being rigorously investigated for many types of applications [12, 13], and a number of variants have been proposed [14–17].

In the field of soundfield reconstruction, in which a large number of microphones and loudspeakers are used, signal representation in the spatial frequency domain is regarded as an essential technique for reducing computational cost [18, 19]. This transformation allows essential information to be compressed and computational complexity to be reduced from order $O(M^3)$ to $O(M)$. The idea of using a spatial transform can also be applied to SC-NMF. To our knowledge, NMF-based source separation that explicitly takes advantage of spatial frequency representation has not yet been deeply investigated. Koyama et al. proposed a MAP estimation method to derive both spatial basis components and their weights, given the position of the primary source, so that spherical waves could be modeled with less spatial aliasing [20]. However, unlike NMF, the method does not take into account the source properties.

The focus of this work is twofold:

1. BSS in the wavenumber domain, in which plane waves can be represented as sparse spectrograms.
2. Efficient BSS for next-generation telecommunication systems, in which a large number of microphones in a uniform linear array are generally employed.

For example, Fig. 1 shows plane waves in the wavenumber domain, originating from three different directions. It is clear that the plane waves can be represented with little overlapping of the spectrograms, which is a great advantage for NMF-based BSS. Furthermore, we can derive an approximated modeling for SCM in the wavenumber domain, showing that SCM can be diagonalized and its inverse is efficiently calculated. This leads to faster implementation of NMF-based multichannel blind source separation with a large number of microphones, compared to the conventional SC-NMF. Also, we can confirm that the proposed method outperforms NTF in separation quality.

This paper is organized as follows: Section 2 briefly explains SCM and DoA-based SC-NMF. Section 3 describes our new method, which is based on an approximation of SCMs. Section 4 shows evaluation results on quality and computational cost. Finally, Section 5 presents some concluding remarks.

2. SPATIAL COVARIANCE MODEL

A spatial covariance model assumes that the TF bins of multichannel spectrograms for each source are represented by a complex Gaussian distribution:

$$\mathbf{s}^i_{fn} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{R}^i_{fn}\right), \tag{1}$$

where i, f, and n are the indices of sources, frequency bins, and frames, respectively; $\mathbf{s}^i_{fn} \in \mathbb{C}^M$ is an M-dimensional vector representing the ith source captured by M microphones; and $\mathbf{R}^i_{fn} \in \mathbb{C}^{M\times M}$ is an SCM. The superposition of multiple sources is represented by a sum of complex Gaussians:

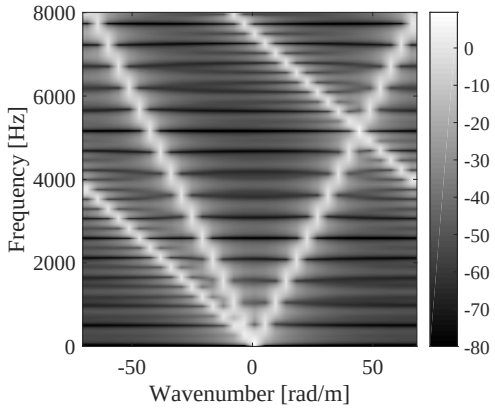


Fig. 1. Virtual plane waves represented in the wavenumber domain, originating from $\pi/20$, $8\pi/20$, and $14\pi/20$.

$$\mathbf{x}_{fn} = \sum_i \mathbf{s}^i_{fn} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \sum_i \mathbf{R}^i_{fn} \right), \tag{2}$$

with

$$\mathbf{R}_{fn} = \sum_i \mathbf{R}^i_{fn} = \mathbb{E}\left[\mathbf{x}_{fn} \mathbf{x}^H_{fn} \right], \tag{3}$$

where $\mathbf{x}_{fn} \in \mathbb{C}^M$ denotes a complex-valued short-time Fourier transform (STFT) of superposed sources captured by M microphones. The Hermitian transpose is denoted by H. To model the SCMs of a mixture, the element-wise divergence between estimated SCMs,

$$\begin{aligned} C(\theta) &= \sum_{fn} D_{\mathrm{IS}} \left(\mathbf{R}_{fn} | \hat{\mathbf{R}}_{fn} \right) \\ &= \sum_{fn} \mathrm{tr}\left(\mathbf{R}_{fn} \hat{\mathbf{R}}^{-1}_{fn} \right) - \log \det \mathbf{R}_{fn} \hat{\mathbf{R}}^{-1}_{fn} - M, \end{aligned} \tag{4}$$

is minimized, where θ is a set of hidden variables, $\hat{\mathbf{R}}_{fn}$ is an estimated SCM, and $\mathrm{tr}()$ is the trace function of linear algebra. The Itakura-Saito (IS) divergence, D_{IS}, is often preferred for minimizing SCMs owing to its scale-invariant nature [6, 7].

2.1. DoA-based SC-NMF

To take advantage of prior knowledge of microphone settings, Nikunen and Virtanen proposed in [8] a DoA-kernel-based approach to the estimation of SCMs:

$$\hat{\mathbf{R}}_{fn} = \sum_k \sum_o \mathbf{J}_{fo} z_{ko} w_{fk} h_{kn}, \tag{5}$$

where $\mathbf{J}_{fo}$ is a DoA kernel; z_{ko} is the weight of a kernel; w_{fk} is the frequency basis; h_{kn} is the activation; and o and k are indices for steering directions and NMF components, respectively.

A DoA kernel, $\mathbf{J}_{fo}$, is calculated from the outer product of steering vectors:

$$\mathbf{J}_{fo} = \mathbf{h}_{fo} \mathbf{h}^H_{fo}, \tag{6}$$

with

$$\mathbf{h}_{fo} = \begin{pmatrix} 1 \\ \vdots \\ e^{j\omega_f (M-1)\gamma_o} \end{pmatrix}, \tag{7}$$

where ω_f is the narrowband frequency and γ_o is the time difference of arrival (TDoA) between two adjacent microphones. This holds only for a uniform linear array with omnidirectional microphones.

3. PROPOSED METHOD

3.1. Minimization of IS divergence

Multiplicative update rules based on cost function (4) together with a model (5) can be derived by an auxiliary function method:

$$z_{ko} \leftarrow z_{ko} \sqrt{\frac{\sum_{fn} \mathrm{tr}\left(\hat{\mathbf{R}}^{-1}_{fn} \mathbf{R}_{fn} \hat{\mathbf{R}}^{-1}_{fn} \mathbf{J}_{fo} \right) w_{fk} h_{kn}}{\sum_{fn} \mathrm{tr}\left(\hat{\mathbf{R}}^{-1}_{fn} \mathbf{J}_{fo} \right) w_{fk} h_{kn}}}, \tag{8}$$

$$w_{fk} \leftarrow w_{fk} \sqrt{\frac{\sum_{on} \mathrm{tr}\left(\hat{\mathbf{R}}^{-1}_{fn} \mathbf{R}_{fn} \hat{\mathbf{R}}^{-1}_{fn} \mathbf{J}_{fo} \right) z_{ko} h_{kn}}{\sum_{on} \mathrm{tr}\left(\hat{\mathbf{R}}^{-1}_{fn} \mathbf{J}_{fo} \right) z_{ko} h_{kn}}}, \tag{9}$$

$$h_{kn} \leftarrow h_{kn} \sqrt{\frac{\sum_{fo} \mathrm{tr}\left(\hat{\mathbf{R}}^{-1}_{fn} \mathbf{R}_{fn} \hat{\mathbf{R}}^{-1}_{fn} \mathbf{J}_{fo} \right) z_{ko} w_{fk}}{\sum_{fo} \mathrm{tr}\left(\hat{\mathbf{R}}^{-1}_{fn} \mathbf{J}_{fo} \right) z_{ko} w_{fk}}}. \tag{10}$$

These formulations are simpler variants of an algorithm proposed by Higuchi and Kameoka [21]. As can be seen from the update rules, the matrix inversions of the estimated SCMs, which are of order $O(M^3)$, make the algorithm intractable for a large number of microphones (e.g., 32).

3.2. Approximation of SCM in spatial frequency domain

To tackle this problem, the SCM approach can be extended to a tensor-based approach by converting the signals on M channels into spatial frequency spectrograms. This transformation can be written and statistically modeled as

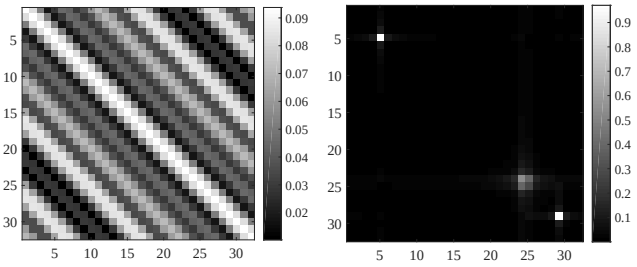
$$\mathbf{F}^H \mathbf{x}_{fn} = \mathbf{F}^H \sum_i \mathbf{s}^i_{fn} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \mathbf{F}^H \mathbf{R}_{fn} \mathbf{F} \right), \tag{11}$$

where $\mathbf{F} \in \mathbb{C}^{M \times M}$ denotes a unitary transform matrix, such as a DFT matrix. The cost function for signals in the spatial frequency domain can be modified to

$$C_{\mathrm{sp}}(\theta) = \sum_{fn} D_{\mathrm{IS}} \left(\mathbf{F}^H \mathbf{R}_{fn} \mathbf{F} | \mathbf{F}^H \hat{\mathbf{R}}_{fn} \mathbf{F} \right). \tag{12}$$

If we assume that $\mathbf{F}$ is a diagonalizing transform of $\mathbf{R}_{fn}$ and $\hat{\mathbf{R}}_{fn}$, that is, if we assume that $\mathbf{F}^H \mathbf{R}_{fn} \mathbf{F}$ and $\mathbf{F}^H \hat{\mathbf{R}}_{fn} \mathbf{F}$ are diagonal matrices or transformed covariance matrices (see Fig. 2), then $\mathbf{F}^H \mathbf{R}_{fn} \mathbf{F}$ and $\mathbf{F}^H \hat{\mathbf{R}}_{fn} \mathbf{F}$ can be well approximated by considering only their diagonal elements:

$$\mathbf{F}^H \mathbf{R}_{fn} \mathbf{F} \approx \begin{pmatrix} a_{1fn} & 0 & \dots & 0 \\ 0 & a_{2fn} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{mfn} \end{pmatrix}, \tag{13}$$



(a) $|\mathbf{R}_{fn}|$ (b) $|\mathbf{F}^H\mathbf{R}_{fn}\mathbf{F}|$

Fig. 2. SCM of mixed plane waves (1 kHz) originating from $\pi/20$, $8\pi/20$, and $14\pi/20$. The number of microphones, M, is 32. The function $|\ |$ denotes the element-wise absolute of the matrix.

$$\mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F} \approx \sum_k \sum_o \begin{pmatrix} b_{1fo} & 0 & \dots & 0 \\ 0 & b_{2fo} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_{mfo} \end{pmatrix} z_{ko} w_{fk} h_{kn}. \tag{14}$$

Thus, the cost function can also be approximated by considering only the diagonal elements:

$$\begin{aligned} C_{\mathrm{sp}}(\theta) &\approx \sum_{mfn} D_{\mathrm{IS}}\left(a_{mfn}|\hat{a}_{mfn}\right) \\ &= \sum_{mfn} D_{\mathrm{IS}}\left(a_{mfn}|\sum_k \sum_o b_{mfo} z_{ko} w_{fk} h_{kn}\right), \end{aligned} \tag{15}$$

under the assumption that $\mathbf{F}^H\hat{\mathbf{R}}_{fn}\mathbf{F} = \mathrm{diag}(\hat{a}_{1fn}, ..., \hat{a}_{Mfn})$. These formulations are analogous to the cost function for NTF.

3.3. Derivation of update rules

The auxiliary function can be constructed by applying Jensen's Inequality to the convex part of the cost function and the Taylor expansion to the concave part:

$$\begin{aligned} C^+_{\mathrm{sp}}(\theta, r_{mfnko}, u_{mfn}) &= \sum_{mfn}\left(\sum_{ko} r^2_{mfnko} \frac{a_{mfn}}{b_{mfo} z_{ko} w_{fk} h_{kn}}\right. \\ &\quad \left. + \log u_{mfn} + \frac{\hat{a}_{mfn} - u_{mfn}}{u_{mfn}}\right), \end{aligned} \tag{16}$$

where r_{mfnko} and u_{mfn} are hidden variables that satisfy $\sum_{ko} r_{mfnko} = 1, r_{mfnko} \geq 0$ and $u_{mfn} \geq 0$. The partial derivatives with respect to z_{ko}, w_{fk}, and h_{kn} are derived by minimizing the auxiliary function

$$\frac{\partial C^+_{\mathrm{sp}}}{\partial z_{ko}} = \sum_{mfn}\left(-r^2_{mfnko}\frac{a_{mfn}}{b_{mfo} z^2_{ko} w_{fk} h_{kn}} + \frac{b_{mfo} w_{fk} h_{kn}}{u_{mfn}}\right), \tag{17}$$

Table 1. Experimental setup

Number of sources	3
Number of channels	$M = 32$
Sampling rate	16 kHz
STFT frame size	1024
STFT frame shift	512
Number of iterations	100

$$\frac{\partial C^+_{\mathrm{sp}}}{\partial w_{fk}} = \sum_{mno}\left(-r^2_{mfnko}\frac{a_{mfn}}{b_{mfo} z_{ko} w^2_{fk} h_{kn}} + \frac{b_{mfo} z_{ko} h_{kn}}{u_{mfn}}\right), \tag{18}$$

$$\frac{\partial C^+_{\mathrm{sp}}}{\partial h_{kn}} = \sum_{mfo}\left(-r^2_{mfnko}\frac{a_{mfn}}{b_{mfo} z_{ko} w_{fk} h^2_{kn}} + \frac{b_{mfo} z_{ko} w_{fk}}{u_{mfn}}\right). \tag{19}$$

The equality of the auxiliary function and the cost function holds only when the hidden variables satisfy

$$r_{mfnko} = \frac{b_{mfo} z_{ko} w_{fk} h_{kn}}{\hat{a}_{mfn}}, \tag{20}$$

$$u_{mfn} = \hat{a}_{mfn}. \tag{21}$$

The update rules reflecting the approximation can be rewritten so that they no longer contain matrix inversions:

$$z_{ko} \leftarrow z_{ko}\sqrt{\frac{\sum_{fn}\sum_m \frac{a_{mfn}}{\hat{a}^2_{mfn}} b_{mfo} w_{fk} h_{kn}}{\sum_{fn}\sum_m \frac{1}{\hat{a}_{mfn}} b_{mfo} w_{fk} h_{kn}}}, \tag{22}$$

$$w_{fk} \leftarrow w_{fk}\sqrt{\frac{\sum_{on}\sum_m \frac{a_{mfn}}{\hat{a}^2_{mfn}} b_{mfo} z_{ko} h_{kn}}{\sum_{on}\sum_m \frac{1}{\hat{a}_{mfn}} b_{mfo} z_{ko} h_{kn}}}, \tag{23}$$

$$h_{kn} \leftarrow h_{kn}\sqrt{\frac{\sum_{fo}\sum_m \frac{a_{mfn}}{\hat{a}^2_{mfn}} b_{mfo} z_{ko} w_{fk}}{\sum_{fo}\sum_m \frac{1}{\hat{a}_{mfn}} b_{mfo} z_{ko} w_{fk}}}. \tag{24}$$

A comparison with the update rules in the previous section shows that the matrix inversions have been replaced with divisions, allowing the algorithm to run at a computational cost of order $O(M)$. We call the proposed method $wavenumberNTF$ (wnNTF), since our initial research target was signals in a uniform linear array, for which a wavenumber representation is practical.

4. EVALUATION

An evaluation was conducted by using the BSS Eval Toolbox, which calculates the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR), and the signal-to-artifact ratio (SAR) [22]. The proposed wnNTF method was compared with other BSS methods in two scenarios: anechoic and reverberant. The sound samples from SiSEC 2008 were used in combination with room impulse responses (RIRs) associated with source positions [23]. The test conditions (Table 1) were the same for both experiments.

Table 2. SDR, SIR, and SAR results

	DS with oracle DoAs			wnNTF		
	SDR	SIR	SAR	SDR	SIR	SAR
Hi-hat	-12.77	-9.88	24.30	7.43	-0.49	2.72
Snare	-3.36	-9.01	10.54	0.68	-7.04	0.70
Bass	3.28	15.93	19.38	16.56	15.13	19.51

4.1. Anechoic scenario

For the anechoic scenario, wnNTF was compared to independent vector analysis (IVA) [24] and NTF [11]. Multichannel observations for a uniform linear array were created simply by summing all the sources together with the addition of proper delays in the frequency domain. Separated signals for IVA were reconstructed by applying the projection back [25]. The distance between microphones was set to 0.384 m. The number of components for NTF and wnNTF was 18. The angle resolution for DoA kernels was limited to 10° in the range 0-180° due to computational cost. It should be noted that the approximation resulting from the extraction of the diagonal elements of SCMs is not correct for a uniform linear array because an SCM cannot be a circulant matrix. However, the approximation error can be mitigated by increasing the length of the array [26, 27].

The average improvement in SDR per file (Fig. 3) and the average improvement in SDR per angle between the two nearest sources (Fig. 4) show that the wnNTF method outperformed the other two methods for all three files at all source distances. The results for different distances show the same tendency as that in [8], namely, that the larger the source distance is, the better the DoA-based method performs with respect to non DoA-based methods. In addition, we assume that the proposed method has an advantage over NTF due to the sparse representation of propagated waves in the wavenumber domain, where the sources are less superposed. Unfortunately, due to the extreme computational complexity of SC-NMF, the comparison for 32 channels could not be completed in a reasonable amount of time. It took 1024 times longer than it took wnNTF ($32^3/32 = 1024$). Moreover, it is less likely that wnNTF will outperform SC-NMF because wnNTF approximates the SCMs and the DoA kernels in the wavenumber domain by using only diagonal elements.

4.2. Reverberant scenario

Evaluations for a reverberant scenario were conducted on the proposed method and on a delay-&-sum (DS) beamformer with oracle DoAs. The azimuths of the three sources were $\pi/20$, $8\pi/20$, and $14\pi/20$ radians. An image-source method was used to obtain reverberant RIRs [28]. The reverberation time, T_{60}, was 0.5 s for a room 7.68 m $\times$ 14 m $\times$ 6 m in size. A uniform linear array of microphones was assumed to be in the center of the room. The distance between microphones was set to 0.046 m. The other parameters were the same as those in Table 1.

The SDR and SIR results (Table 2) show that the proposed method outperformed the DS beamformer in the reverberant scenario, even with oracle DoAs for the beamformer. This is probably due to the fact that the proposed method is capable of modeling full-rank SCMs, even though a circulant matrix is approximated by a large Toeplitz matrix, whereas the DS beamformer can only steer in a single direction.

Table 3. Computation time for 1 iteration [s]

IVA	NTF	wnNTF	DoA-based SC-NMF
3.076e-01	1.2351e+00	1.6098e+01	1.0934e+03

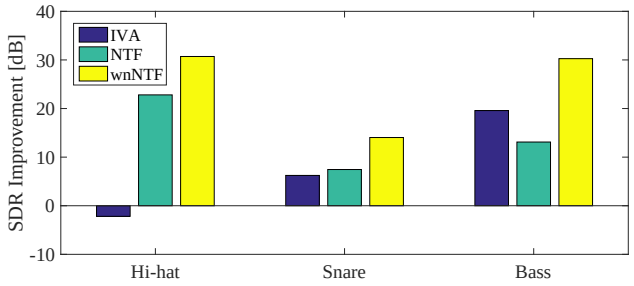


Fig. 3. Average improvement in SDR per source for hi-hat, snare drums, and bass guitar.

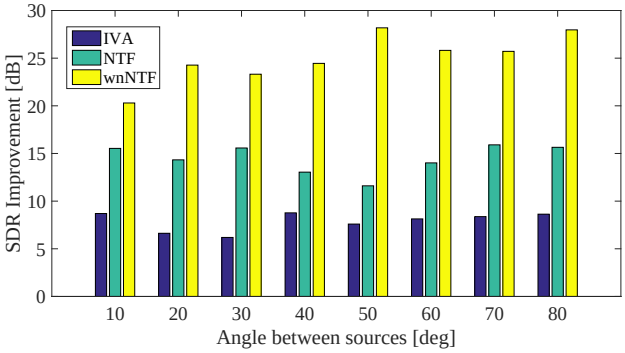


Fig. 4. Average improvement in SDR per direction, with the x-axis being the angle between the two nearest sources.

4.3. Computational cost

The computation time required for 1 iteration for each method is listed in Table 3. The parameter settings for the experiment are the same as those of section 4.1. The computation time was measured using MATLAB codes run on a PC with 18 Intel Xeon cores and 384 GB of memory. Although there is unexpected overhead in real implementation, wnNTF is still greatly faster than SC-NMF, confirming the proposed method's advantage in computational efficiency.

5. CONCLUSION AND FUTURE WORK

This paper describes the use of NTF in the wavenumber domain to reduce the computational cost of BSS for a large number of channels. The technique is based on the approximation of SCMs, which are transformed into the wavenumber domain in advance. With this approximation, the cost of running the algorithm is of order $O(M)$, whereas it is of order $O(M^3)$ for conventional SC-NMF. An evaluation conducted for anechoic and reverberant scenarios showed that the proposed method yielded good separation quality. Future plans call for a comparison with SC-NMF and an evaluation of the use other orthogonal transforms, for example, spherical harmonics.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS), USA*, 2000, pp. 556–562.

[2] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *the International Computer Music Conference (ICMC)*, 2003.

[3] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[4] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550 –563, march 2010.

[5] S. Arberet, A. Ozerov, N. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *10th International Conference on Information Sciences, Signal Processing and their Applications (ISSPA), Kuala Lumpur, Malaysia, 10-13 May*, 2010, pp. 1–4.

[6] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[7] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[8] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 3, pp. 727–739, 2014.

[9] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Queensland, Australia, April 19-24*, 2015, pp. 276–280.

[10] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations - Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 2009.

[11] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *Exploring Music Contents - 7th International Symposium (CMMR), Málaga, Spain, June 21-24. Revised Papers*, 2010, pp. 102–115.

[12] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML), Bonn, Germany, August 7-11*, 2005, pp. 792–799.

[13] F. Cong, A. H. Phan, Q. Zhao, A. K. Nandi, V. Alluri, P. Toiviainen, H. Poikonen, M. Huotilainen, A. Cichocki, and T. Ristaniemi, "Analysis of ongoing EEG elicited by natural music stimuli using nonnegative tensor factorization," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, August 27-31*, 2012, pp. 494–498.

[14] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008.

[15] Y. Mitsufuji and A. Roebel, "On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization," *EURASIP J. Adv. Sig. Proc.*, vol. 2014, pp. 40, 2014.

[16] N. D. Stein, "Nonnegative tensor factorization for directional blind audio source separation," *CoRR*, vol. abs/1411.5010, 2014.

[17] Y. Mitsufuji, M. Liuni, A. Baker, and A. Roebel, "Online nonnegative tensor deconvolution for source detection in 3DTV audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 4-9*, 2014, pp. 3082–3086.

[18] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of Audio Engineering Society*, pp. 1004–1025, 2005.

[19] S. Koyama, K. Furuya, Y. Hiwasaki, and Y. Haneda, "Analytical approach to wave field reconstruction filtering in spatio-temporal frequency domain," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 4, pp. 685–696, 2013.

[20] S. Koyama, K. Furuya, Y. Haneda, and H. Saruwatari, "Source-location-informed sound field recording and reproduction," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 881–894, 2015.

[21] T. Higuchi and H. Kameoka, "Unified approach for underdetermined BSS, VAD, dereverberation and DOA estimation with multichannel factorial HMM," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, USA, December 3-5*, 2014, pp. 562–566.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] "In signal separation evaluation campaign (SiSEC 2008):http://www.sisec.wiki.irisa.fr," .

[24] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, October 16-19*, 2011, pp. 189–192.

[25] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

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

[27] Robert M. Gray, "Toeplitz and circulant matrices: A review," Tech. Rep., 2001.

[28] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small room acoustics," vol. 65, no. 4, 1979.