Prior Distribution Design for Music Bleeding-Sound Reduction Based on Nonnegative Matrix Factorization

Yusaku Mizobuchi*, Daichi Kitamura*, Tomohiko Nakamura†, Hiroshi Saruwatari†,
Yu Takahashi‡, and Kazunobu Kondo‡
* National Institute of Technology, Kagawa College, Kagawa, Japan
† The University of Tokyo, Tokyo, Japan
‡ Yamaha Corporation, Shizuoka, Japan

***Abstract*—When we place microphones close to a sound source near other sources in audio recording, the obtained audio signal includes undesired sound from the other sources, which is often called cross-talk or bleeding sound. For many audio applications including onstage sound reinforcement and sound editing after a live performance, it is important to reduce the bleeding sound in each recorded signal. However, since microphones are spatially apart from each other in this situation, typical phase-aware blind source separation (BSS) methods cannot be used. We propose a phase-insensitive method for blind bleeding-sound reduction. This method is based on time-channel nonnegative matrix factorization, which is a BSS method using only amplitude spectrograms. With the proposed method, we introduce the gamma-distribution-based prior for leakage levels of bleeding sounds. Its optimization can be interpreted as maximum a posteriori estimation. The experimental results of music bleeding-sound reduction indicate that the proposed method is more effective for bleeding-sound reduction of music signals compared with other BSS methods.**

I. Introduction

When we record a live musical performance, many microphones are usually arranged among the players. Some are located very close to each of the audio sources, such as musical instruments, vocals, and amplifiers. These close microphones are placed to pick up the sound from the source other than that which is intended. However, undesirable audio leakage from the non-target audio sources is also captured, which is often called "cross-talk" or "bleeding sound," as shown in Fig. 1.

In onstage mixing, sound engineers control the balance of sound levels of individual sources, and the processed sounds are provided to the audience through loudspeakers and performers through monitor speakers. Bleeding sound makes such sound reinforcement difficult, degrading musical performance quality. It is also necessary to avoid sound bleeding for high-quality audio editing (remixing) of the recorded signals after a live performance. For these reasons, sound engineers carefully place close microphones so that the as much bleeding sound is reduced as possible. Putting acoustic barriers between the sound sources and reducing sound reflection in the recording room are also effective. However, completely avoiding bleeding sound is almost impossible. In other words, sound bleeding essentially occurs in a live-recording situation.

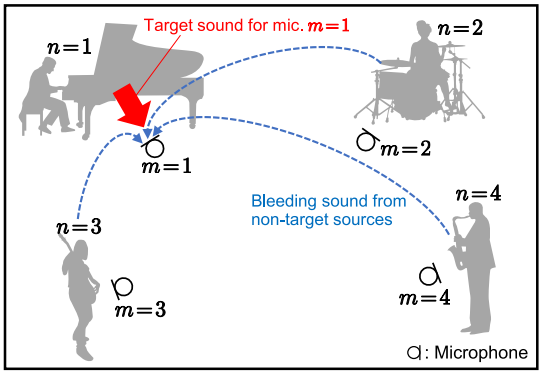

Fig. 1. Spatial arrangement of sources and close microphones, where $M = N = 4$. Target sound is contaminated with bleeding sound from other non-target sources.

Bleeding-sound reduction is similar to the well-investigated problem called multichannel audio source separation (MASS) [1], [2], [3], [4], but some conditions are different from those in MASS, which are listed as follows.

(a) The signal-to-noise ratio (SNR) of the observed signal is relatively high because of a close miking setup, where the "signal" is a target source for the close microphone and the "noise" is the leakage from the other sources.
(b) The observed multichannel signals are already "labeled," namely, the target source for each microphone is known because each microphone is located close to each sound source.
(c) The microphones are spatially apart from each other (e.g., more than 2 m), resulting in serious spatial aliasing.
(d) The requirement of separation quality is relatively high so as not to degrade the artistic value of the music signal.

Conditions (a) and (b) are advantages of bleeding-sound reduction, which make resolving bleeding sound easier than MASS. However, conditions (c) and (d) are difficult. In particular, condition (c) is critical because typical high-quality MASS, including beamformers [1], [2] and independence-based blind source separation (BSS) [5], [6], [7], [8], [9], [10], [11], uses

phase differences between microphones, which are unreliable in bleeding-sound reduction due to spatial aliasing. To tackle this problem, phase-insensitive (amplitude- or power-based) MASS [12], [13], [14], [15] can be applied. Togami et al. [12] applied nonnegative matrix factorization (NMF) [17], [18] to the time-channel domain in each frequency (hereafter, time-channel NMF: TCNMF), where both the nonnegative mixing matrix and amplitude activation of each source are estimated in each frequency bin. TCNMF performs well even under condition (c) or an asynchronous recording condition [13], [14], although its effectiveness regarding music bleeding-sound reduction has not been investigated. A BSS-based method that ignores the phase information was proposed [15], which is called linear demixed domain multichannel NMF (DMNMF). Similar to TCNMF, this method also estimates the frequency-wise nonnegative mixing matrix. Das et al. [16] introduced supervised information to accurately reduce the bleeding sound, where the frequency-wise nonnegative mixing matrix (i.e., leakage levels of the non-target sources for each close microphone) is measured before the musical performance or calculated using the solo-played time segments of each source. However, to reduce the onsite recording cost for sound reinforcement, such supervision should not be used. Also, a mismatch between the obtained mixing matrix and actual condition may markedly degrade reduction performance.

We aimed to reduce bleeding sound in a fully blind manner, namely, the spatial locations of sources and microphones are unknown. We also did not use supervision of sources, such as solo-played music datasets, to avoid the mismatch between training and test data; thus, supervised deep-neural-network-based approaches [19], [20], [21], [22], [23] are out of the scope of this paper. We propose a phase-insensitive method for blind bleeding-sound reduction, which is a modification of TCNMF: we introduce an a priori generative model for diagonal and off-diagonal elements of the frequency-wise mixing matrix to model relative leakage levels of bleeding sounds. This method is based on NMF with maximum a posteriori (MAP) estimation, which was originally proposed by Cemgil [24], and we demonstrate that the proposed method is suitable for reducing music bleeding sound.

II. Conventional Methods

A. Mixture Model

Let M and N be the numbers of microphones (channels) and sources, respectively. The source, observed, and estimated signals are respectively denoted as

$$\tilde{\boldsymbol{s}}(t) = [\tilde{s}_1(t), \tilde{s}_2(t), \cdots, \tilde{s}_n(t), \cdots, \tilde{s}_N(t)]^{\mathrm{T}} \in \mathbb{R}^N, \quad (1)$$

$$\tilde{\boldsymbol{x}}(t) = [\tilde{x}_1(t), \tilde{x}_2(t), \cdots, \tilde{x}_m(t), \cdots, \tilde{x}_M(t)]^{\mathrm{T}} \in \mathbb{R}^M, \quad (2)$$

$$\tilde{\boldsymbol{y}}(t) = [\tilde{y}_1(t), \tilde{y}_2(t), \cdots, \tilde{y}_n(t), \cdots, \tilde{y}_N(t)]^{\mathrm{T}} \in \mathbb{R}^N, \quad (3)$$

where $t = 1, 2, \cdots, T$, $n = 1, 2, \cdots, N$, and $m = 1, 2, \cdots, M$ are the indices of discrete time, source, and microphone, respectively. Under the recording condition described in Sect. I, the mixing system becomes determined ($M = N$) or overdetermined ($M > N$). In this study, we focused only on the determined case, which is the most difficult situation in bleeding-sound reduction.

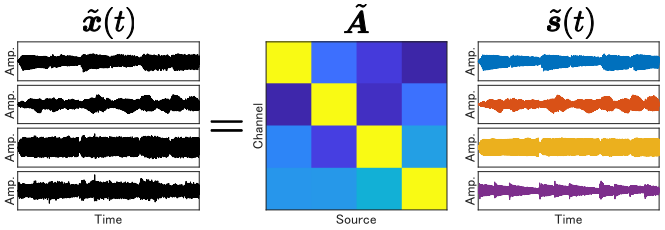

Fig. 2. Instantaneous mixture model for bleeding-sound reduction, where $M = N = 4$. Color brightness in mixing matrix $\tilde{\boldsymbol{A}}$ shows amplitude level of each element (brighter is larger). Due to close miking setup, diagonal elements in $\tilde{\boldsymbol{A}}$ have larger amplitudes compared with off-diagonal elements.

In an instantaneous mixture, the observed and estimated signals can respectively be modeled as

$$\tilde{\boldsymbol{x}}(t) = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{s}}(t), \quad (4)$$

$$\tilde{\boldsymbol{y}}(t) = \tilde{\boldsymbol{W}}\tilde{\boldsymbol{x}}(t), \quad (5)$$

where $\tilde{\boldsymbol{A}} \in \mathbb{R}^{M\times N}$ and $\tilde{\boldsymbol{W}} \in \mathbb{R}^{N\times M}$ are the time-invariant mixing and demixing matrices, respectively. The mixture model (4) is illustrated in Fig. 2. Since the observed signal $\tilde{\boldsymbol{x}}(t)$ is "labeled," as explained in condition (b) in Sect. I, we define that $\tilde{x}_m(t)$ is the close-microphone signal for the mth source $\tilde{s}_m(t)$ ($n = m$), as shown in Figs. 1 and 2. Thus, $\tilde{x}_m(t)$ mainly contains the sound from the target source $\tilde{s}_m(t)$, although the bleeding sound from the non-target sources $\tilde{s}_{m'}(t)$ is also included, where $m' \neq m$. For this reason, the absolute values of diagonal elements in $\tilde{\boldsymbol{A}}$ should be large enough, and those of off-diagonal elements become small, which results in high-SNR condition (a) in Sect. I. Blind bleeding-sound reduction is formulated as an estimation problem of the mixing matrix $\tilde{\boldsymbol{A}}$ or demixing matrix that satisfies $\tilde{\boldsymbol{W}} = \tilde{\boldsymbol{A}}^{-1}$ from only $\tilde{\boldsymbol{x}}(t)$.

In actual recording, the mixing system (4) becomes a convolutive mixture due to time difference of arrival and room reverberation. To simply model the convolutive mixture, we assume that the impulse responses (reverberation time) between microphones and sources are shorter than the window length used in the short-time Fourier transform (STFT). This assumption enables us to respectively model the reverberant observed and estimated signals as[1]

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij}, \quad (6)$$

$$\mathbf{y}_{ij} = \mathbf{W}_i \mathbf{x}_{ij}, \quad (7)$$

where

$$\mathbf{s}_{ij} = [\mathrm{s}_{ij1}, \mathrm{s}_{ij2}, \cdots, \mathrm{s}_{ijn}, \cdots, \mathrm{s}_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N, \quad (8)$$

$$\mathbf{x}_{ij} = [\mathrm{x}_{ij1}, \mathrm{x}_{ij2}, \cdots, \mathrm{x}_{ijm}, \cdots, \mathrm{x}_{ijM}]^{\mathrm{T}} \in \mathbb{C}^M, \quad (9)$$

$$\mathbf{y}_{ij} = [\mathrm{y}_{ij1}, \mathrm{y}_{ij2}, \cdots, \mathrm{y}_{ijn}, \cdots, \mathrm{y}_{ijN}]^{\mathrm{T}} \in \mathbb{C}^N. \quad (10)$$

Here, $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$ are the indices of the frequency bin and time frame, respectively, and $\mathbf{A}_i \in \mathbb{C}^{M\times N}$

[1]Note that roman font signal denotes complex values and italic font signal denotes real (or nonnegative) values.

is the complex-valued frequency-wise mixing matrix. Also, s_{ijn}, x_{ijm}, and y_{ijn} are the complex-valued time-frequency-wise elements of the source, observed, and estimated spectrograms $\mathbf{S}_n \in \mathbb{C}^{I\times J}$, $\mathbf{X}_m \in \mathbb{C}^{I\times J}$, and $\mathbf{Y}_n \in \mathbb{C}^{I\times J}$, respectively. In (6), the convolutive mixture is converted to the frequency-wise instantaneous mixture via STFT.

Typical beamformers [1], [2] and BSS methods [5], [6], [7], [8], [9], [10], [11] are used to estimate the complex-valued demixing matrix $\mathbf{W}_i$ on the basis of a principle of microphone arrays, e.g., time difference of arrival, and these methods rely on the phase differences between microphones. When microphones are spatially apart from each other, these methods cannot precisely estimate $\mathbf{W}_i$ because of spatial aliasing. This problem is salient in bleeding-sound reduction.

B. DMNMF

To cope with spatial aliasing, the power-based BSS method DMNMF was proposed [15]. DMNMF can be interpreted as a phase-insensitive version of independent low-rank matrix analysis (ILRMA) [4], [10], [11], and the observed signal is modeled as

$$\boldsymbol{x}_{ij}^{.2} \approx \boldsymbol{A}_i \boldsymbol{s}_{ij}^{.2} \quad \forall i,j, \tag{11}$$

$$\boldsymbol{A}_i = \mathrm{abs}(\mathbf{A}_i) \in \mathbb{R}_{\geq 0}^{M\times N}, \tag{12}$$

$$\boldsymbol{x}_{ij} = \mathrm{abs}(\mathbf{x}_{ij}) \in \mathbb{R}_{\geq 0}^{M}, \tag{13}$$

$$\boldsymbol{s}_{ij} = \mathrm{abs}(\mathbf{s}_{ij}) \in \mathbb{R}_{\geq 0}^{N}, \tag{14}$$

where the dotted exponent $\cdot^{.q}$ and absolute operation $\mathrm{abs}(\cdot)$ for vectors or matrices return the element-wise qth power and absolute, respectively; thus, $\boldsymbol{x}_{ij}^{.2}$ and $\boldsymbol{s}_{ij}^{.2}$ are the power spectrogram components of $\{\mathbf{X}_m\}_{m=1}^M$ and $\{\mathbf{S}_n\}_{n=1}^N$, respectively. DMNMF approximates (6) by the nonnegative frequency-wise mixing matrix $\boldsymbol{A}_i$ in the power-spectrogram domain to ignore the phase information. In addition, the power spectrogram of each source is modeled by a low-rank matrix using NMF. After estimating $\boldsymbol{A}_i$ and $\boldsymbol{s}_{ij}^{.2}$ from $\boldsymbol{x}_{ij}^{.2}$, we can recover the estimated signal $\mathbf{y}_{ij}$ by Wiener filtering.

C. TCNMF

The amplitude-based BSS method TCNMF was proposed [12] and applied [13], [14] to speech enhancement. Whereas typical NMF is a low-rank decomposition of time-frequency matrices, TCNMF decomposes frequency-wise time-channel matrices in the amplitude domain as

$$\boldsymbol{X}_i \approx \boldsymbol{A}_i \boldsymbol{S}_i \quad \forall i, \tag{15}$$

$$\boldsymbol{X}_i = [\boldsymbol{x}_{i1}\ \boldsymbol{x}_{i2}\ \cdots\ \boldsymbol{x}_{iJ}] \in \mathbb{R}_{\geq 0}^{M\times J}, \tag{16}$$

which is illustrated in Fig. 3, where $\boldsymbol{X}_i$ is the frequency-wise time-channel observed signal in the amplitude domain and $\boldsymbol{S}_i \in \mathbb{R}_{\geq 0}^{N\times J}$ is a time-source activation matrix: $\boldsymbol{S}_i$ involves time-varying gains of each source as the row vectors. By estimating $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ in the same manner as typical NMF, we can reconstruct the estimated sources using Wiener filtering.

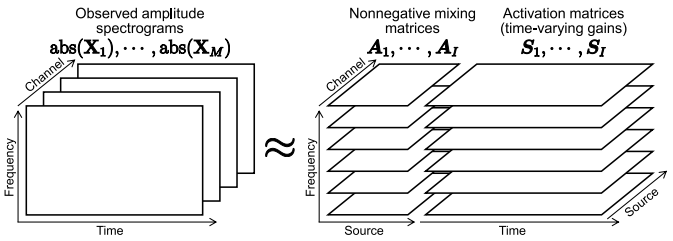

Fig. 3. Decomposition model of TCNMF, where $M = N = 4$ and $I = 6$. Note that $\mathrm{abs}(\mathbf{X}_m)$ is channel-wise time-frequency matrix, but $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ are frequency-wise source-channel and time-source matrices, respectively.

The variables $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ can be estimated by solving the following minimization problem [18]:

$$\min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i) \ \text{ s.t. } a_{imn}, s_{inj} \geq 0 \ \ \forall i,m,n,j, \tag{17}$$

where

$$\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i) = \sum_{m,j}\left(x_{imj}\log\frac{x_{imj}}{\sum_n a_{imn}s_{inj}} - x_{imj} + \sum_n a_{imn}s_{inj}\right) \tag{18}$$

is the generalized Kullback-Leibler (KL) divergence that measures the similarity between $\boldsymbol{X}_i$ and $\boldsymbol{A}_i\boldsymbol{S}_i$, $\mathcal{A}$ and $\mathcal{S}$ are the sets $\{\boldsymbol{A}_i\}_{i=1}^I$ and $\{\boldsymbol{S}_i\}_{i=1}^I$, respectively, and x_{imj}, a_{imn}, and s_{inj} are the elements of $\boldsymbol{X}_i$, $\boldsymbol{A}_i$, and $\boldsymbol{S}_i$, respectively. However, since $\boldsymbol{A}_i$ is an $M \times N$ square matrix in the determined case, the minimization problem (17) has a trivial solution, namely, $\boldsymbol{A}_i = \boldsymbol{I}$ for all i, where $\boldsymbol{I}$ is an identity matrix. To avoid this trivial solution, an $L_{0.5}$-norm-based sparse regularizer was introduced for each time frame [12] as follows:

$$\min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i) + \mu\sum_{i,j}\|\boldsymbol{s}_{ij}\|_{0.5} \quad \text{s.t. } a_{imn}, s_{inj} \geq 0 \ \ \forall i,m,n,j, \tag{19}$$

where μ is a weight coefficient for regularization. Note that $\boldsymbol{s}_{ij}$ is a time-frame-wise vector in $\boldsymbol{S}_i$, namely, $\boldsymbol{S}_i = [\boldsymbol{s}_{i1}\ \boldsymbol{s}_{i2}\ \cdots\ \boldsymbol{s}_{iJ}]$.

III. Proposed Method

A. Motivation

In bleeding-sound reduction, phase information cannot be used because of the close miking setup and serious spatial aliasing. As a phase-insensitive method, DMNMF is a reasonable approach. However, full-blind parameter optimization of DMNMF is difficult and unstable. In fact, a priori information of steering vectors (column vectors of $\mathbf{A}_i$) or a phase-aware BSS method is used for pre-estimation [15] to stabilize and improve BSS performance. TCNMF can estimate the source signals without phase information, even in asynchronous recording [13]. However, its performance for music BSS or bleeding-sound reduction has not been investigated. In particular, the sparse regularizer $\sum_{i,j}\|\boldsymbol{s}_{ij}\|_{0.5}$

in (19) may degrade the sound quality of estimated signals in music mixture. This is because the regularizer is based on a W-disjoint-orthogonality assumption in the time-frequency domain [25], which is suitable only for speech mixtures. Since music mixtures frequently include both spectral and temporal overlaps of sources, the sparse regularizer for $\boldsymbol{S}_i$ may be inappropriate.

To avoid the trivial solution of $\boldsymbol{A}_i$ in TCNMF, we use our proposed method to regularize both the diagonal and off-diagonal elements of the nonnegative mixing matrix $\boldsymbol{A}_i$ instead of regularizing $\boldsymbol{S}_i$. The proposed method can be interpreted as a MAP estimation, where the bleeding-sound levels are assumed to be generated by the gamma distribution prior.

B. Generative Model of KL-Divergence-Based NMF

Cemgil [24] revealed the generative model of NMF with KL divergence (KLNMF): the minimization problem in KLNMF is equivalent to the maximum likelihood (ML) estimation with the Poisson generative model. For (17), the following generative model is assumed:

$$z_{imnj} \sim \mathcal{P}(z_{imnj}; a_{imn}s_{inj}), \tag{20}$$

$$\mathcal{P}(z;\lambda) = \frac{1}{\Gamma(z+1)} e^{-\lambda}\lambda^z, \tag{21}$$

where $z_{imnj} \in \mathbb{N}$ is a random variable that satisfies $x_{imj} = e + \sum_n z_{imnj}$, $\mathcal{P}(z;\lambda)$ is the Poisson distribution with the random variable $z \in \mathbb{N}$ and parameter $\lambda > 0$, $\Gamma(z+1) = z!$ is the gamma function, and e is a random variable that obeys uniform distribution in the range $[0, 1)$. Also, z_{imnj} is assumed to be mutually independent w.r.t. i, m, n, and j. The Poisson random variables have the superposition property, namely, when $z_n \sim \mathcal{P}(z_n;\lambda_n)$ and $x = \sum_n z_n$, the marginal probability is given by $p(x) = \mathcal{P}(x; \sum_n \lambda_n)$. Therefore, the marginal log-likelihood of $\boldsymbol{X}_i$ is given by

$$\begin{aligned}
&\log p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i) \\
&\quad = \log \prod_{m,j} \sum_{z_{imnj}} p(x_{inm}; z_{imnj}) p(z_{imnj}; a_{imn}s_{inj}) \\
&\quad = \log \prod_{m,j} \mathcal{P}(x_{imj}; \textstyle\sum_n a_{imn}s_{inj}) \\
&\quad = \sum_{m,j} \Bigg[x_{imj} \log \sum_n a_{imn}s_{inj} \\
&\qquad - \sum_n a_{imn}s_{inj} - \log\Gamma(x_{imj}+1) \Bigg].
\end{aligned} \tag{22}$$

The maximization of (22) w.r.t. a_{imn} and s_{inj} for all i (ML estimation) is equivalent to the minimization of (18).

C. A Priori Generative Model for Bleeding-Sound Levels

With the proposed method, to avoid the trivial solution of $\boldsymbol{A}_i$, we introduce the following a priori generative model into the diagonal and off-diagonal elements of $\boldsymbol{A}_i$:

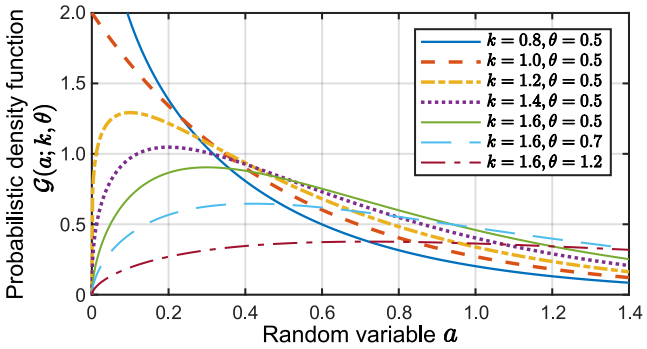

Fig. 4. Probabilistic density function of gamma distribution.

$$a_{imn} \sim \begin{cases} \delta(a_{imn}-1) & (m=n) \\ \mathcal{G}(a_{imn};k,\theta) & (m\neq n) \end{cases}, \tag{23}$$

$$\mathcal{G}(a;k,\theta) = \frac{1}{\Gamma(k)\theta^k} a^{k-1} e^{-a/\theta}, \tag{24}$$

where $\delta(a)$ is the Dirac's delta distribution and $\mathcal{G}(a;k,\theta)$ is the gamma distribution with the random variable $a \geq 0$ and shape and scale parameters $k > 0$ and $\theta > 0$. Note that the gamma distribution is a conjugate prior of the Poisson generative model (20). In addition, a_{imn} is assumed to be mutually independent w.r.t. i, m, and n; thus, the prior distribution of $\boldsymbol{A}_i$ becomes

$$\begin{aligned}
p(\boldsymbol{A}_i;k,\theta) &= \prod_{m,n=m} p(a_{imn}) \prod_{m,n\neq m} p(a_{imn};k,\theta) \\
&= \prod_{m,n=m} \delta(a_{imn}-1) \prod_{m,n\neq m} \mathcal{G}(a_{imn};k,\theta).
\end{aligned} \tag{25}$$

The prior (25) enables us to control the probability of off-diagonal elements of $\boldsymbol{A}_i$ (relative leakage levels of bleeding sound) by k and θ, while restricting all the diagonal elements to be unity. As shown in Fig. 4, we can avoid $a_{imn} = 0$ for all $m \neq n$, which is the trivial solution of $\boldsymbol{A}_i$, by setting the shape parameter to $k > 1$. Hereafter, we consider $k > 1$ only.

For the activation matrix $\boldsymbol{S}_i$, we do not assume explicit structure, but only the nonnegativity prior is used as follows:

$$\begin{aligned}
s_{inj} &\sim \lim_{\beta\to\infty} \frac{1}{\beta} \mathcal{I}[0 \leq s_{inj} \leq \beta] \\
&\propto \mathcal{I}[0 \leq s_{inj}],
\end{aligned} \tag{26}$$

where β is the normalized coefficient and $\mathcal{I}[\cdot]$ denotes a binary distribution that has value one when its argument is true and zero otherwise. Similar to $\boldsymbol{A}_i$, s_{inj} is mutually independent w.r.t. i, n, and j, and the prior distribution of $\boldsymbol{S}_i$ becomes

$$\begin{aligned}
p(\boldsymbol{S}_i) &= \prod_{n,j} p(s_{inj}) \\
&\propto \prod_{n,j} \mathcal{I}[0 \leq s_{inj}].
\end{aligned} \tag{27}$$

D. Cost Function for MAP Estimation

On the basis of the above-mentioned prior distributions, we estimate variables $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ in the MAP sense. The posterior distribution can be obtained as

$$\prod_i p(\boldsymbol{A}_i, \boldsymbol{S}_i; \boldsymbol{X}_i) \propto \prod_i \underbrace{p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i)}_{\text{Likelihood}} \underbrace{p(\boldsymbol{A}_i; k, \theta) p(\boldsymbol{S}_i)}_{\text{Priors}}. \tag{28}$$

By taking a negative logarithm of (28), we can decompose the right side of (28) as

$$\mathcal{J} = -\sum_i \left[\log p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i) + \log p(\boldsymbol{A}_i; k, \theta) + \log p(\boldsymbol{S}_i)\right]. \tag{29}$$

From (22), (25), and (27), the cost function $\mathcal{J}$ is obtained as

$$\begin{aligned}\mathcal{J} = \sum_{i,m,j} \Bigg[& -x_{imj} \log \sum_n a_{imn} s_{inj} \\ & + \sum_n a_{imn} s_{inj} + \log \Gamma(x_{imj} + 1) \Bigg] \\ & + \sum_{i,m,n=m} \mathbb{I}[a_{imn} = 1] \\ & + \sum_{i,m,n\neq m} \left[-(k-1) \log a_{imn} + \frac{1}{\theta} a_{imn} \right] \\ & + \sum_{i,n,j} \mathbb{I}[0 \leq s_{imn}], \end{aligned} \tag{30}$$

where $\mathbb{I}[\cdot] = -\log \mathcal{I}[\cdot]$ denotes an indicator function that has value zero when its argument is true and ∞ otherwise. The MAP estimation of $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ is a minimization problem of (30), and this minimization w.r.t. $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ is equivalent to the following problem:

$$\begin{aligned} & \min_{\mathcal{A},\mathcal{S}} \sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) + \sum_{i,m,n\neq m} \mathcal{R}(a_{imn}; k, \theta) \\ & \text{s.t. } a_{imn}, s_{inj} \geq 0 \ \forall i, m, n, j \text{ and } a_{imn} = 1 \ \forall m = n, \end{aligned} \tag{31}$$

where

$$\mathcal{R}(a_{imn}; k, \theta) = \left[-(k-1) \log a_{imn} + \frac{1}{\theta} a_{imn} \right] \tag{32}$$

is the regularizer that corresponds to the gamma distribution prior (24) for the off-diagonal elements of $\boldsymbol{A}_i$.

E. Derivation of Optimization Algorithm

The minimization problem (31) can be solved using a majorization-minimization (MM) algorithm [18], [26], which is often used in the context of NMF optimization. The majorization function of the fidelity term $\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i)$ can be designed using Jensen's inequality as follows:

$$\begin{aligned} & \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i | \boldsymbol{A}_i \boldsymbol{S}_i) \\ & \overset{\mathrm{c}}{=} \sum_{i,m,j} \left(-x_{imj} \log \sum_n a_{imn} s_{inj} + \sum_n a_{imn} s_{inj} \right) \\ & = \sum_{i,m,j} \left(-x_{imj} \log \sum_n \xi_{imnj} \frac{a_{imn} s_{inj}}{\xi_{imnj}} + \sum_n a_{imn} s_{inj} \right) \\ & \leq \sum_{i,m,j} \left(-x_{imj} \sum_n \xi_{imnj} \log \frac{a_{imn} s_{inj}}{\xi_{imnj}} + \sum_n a_{imn} s_{inj} \right) \\ & \equiv \mathcal{D}^+(\boldsymbol{A}_i, \boldsymbol{S}_i, \Xi), \end{aligned} \tag{33}$$

where $\overset{\mathrm{c}}{=}$ denotes equality up to a constant, $\xi_{imnj} > 0$ is an auxiliary variable that satisfies $\sum_n \xi_{imnj} = 1$, and Ξ is a set of ξ_{imnj} for all i, m, j, and n. The equality in (33) holds if and only if

$$\xi_{imnj} = \frac{a_{imn} s_{inj}}{\sum_{n'} a_{imn'} s_{in'j}} \ \forall i, m, j, n. \tag{34}$$

From (33), the MM problem is obtained as

$$\begin{aligned} & \min_{\mathcal{A},\mathcal{S},\Xi} \sum_i \mathcal{D}^+(\boldsymbol{A}_i, \boldsymbol{S}_i, \Xi) + \sum_{i,m,n\neq m} \mathcal{R}(a_{imn}; k, \theta) \\ & \text{s.t. } a_{imn}, s_{inj} \geq 0 \ \forall i, m, n, j, \quad \xi_{imnj} > 0 \ \forall i, m, n, j, \\ & \qquad \text{and } a_{imn} = 1 \ \forall m = n. \end{aligned} \tag{35}$$

By setting the derivative of the majorization function (35) w.r.t. a_{imn} and s_{inj} to zero and substituting (34) for ξ_{imnj}, we can derive the update rules. Since the regularizer does not affect s_{inj}, the update rule of s_{inj} is the same as that of simple KLNMF [18] and expressed as

$$s_{inj} \leftarrow s_{inj} \frac{\sum_m \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{inj'}} a_{imn}}{\sum_m a_{imn}}. \tag{36}$$

For the off-diagonal elements a_{imn} ($m \neq n$), we have the following equations from the derivative of the majorization function:

$$\sum_j \left(-x_{imj} \frac{\xi_{imnj}}{a_{imn}} + s_{inj} \right) - (k-1) \frac{1}{a_{imn}} + \frac{1}{\theta} = 0. \tag{37}$$

Therefore, we have

$$a_{imn} = \frac{(k-1) + \sum_j x_{imj} \xi_{imnj}}{\frac{1}{\theta} + \sum_j s_{inj}}. \tag{38}$$

The update rule of the off-diagonal elements a_{imn} is derived by substituting (34) as

$$a_{imn} \leftarrow \frac{(k-1) + a_{imn} \sum_j \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} s_{inj}}{\frac{1}{\theta} + \sum_j s_{inj}}. \tag{39}$$

The nonnegativity of a_{imn} and s_{inj} can hold by setting their initial values to nonnegative values. Since the value of the diagonal elements of $\boldsymbol{A}_i$ is restricted, we initialize the diagonal elements a_{imn} ($m = n$) with unity and fix them during the iterative optimization of the other variables.

The efficient matrix-form implementation of (36) and (39) is as follows:

$$\boldsymbol{A}_i \leftarrow \frac{(k-1)+\boldsymbol{A}_i \odot \left(\frac{\boldsymbol{X}_i}{\boldsymbol{A}_i\boldsymbol{S}_i}\boldsymbol{S}_i^{\mathrm{T}}\right)}{\frac{1}{\theta}+\mathbf{1}\boldsymbol{S}_i^{\mathrm{T}}} \quad \forall i, \tag{40}$$

$$\mathrm{diag}(\boldsymbol{A}_i) \leftarrow [1,1,\cdots,1]^{\mathrm{T}} \quad \forall i, \tag{41}$$

$$\boldsymbol{S}_i \leftarrow \boldsymbol{S}_i \odot \frac{\boldsymbol{A}_i^{\mathrm{T}}\frac{\boldsymbol{X}_i}{\boldsymbol{A}_i\boldsymbol{S}_i}}{\boldsymbol{A}_i^{\mathrm{T}}\mathbf{1}} \quad \forall i, \tag{42}$$

where $\odot$ and the quotient symbol for matrices denote element-wise multiplication and division, respectively, $\mathbf{1}$ is an $M \times J$ matrix containing only ones, and $\mathrm{diag}(\cdot)$ returns a vector that consists of the diagonal elements of the input square matrix. Note that (40) will change the value of the diagonal elements of $\boldsymbol{A}_i$, but they are immediately replaced with unity by (41). It is guaranteed that the iterative calculation of (40)–(42) monotonically decreases the cost function (31).

F. Balancing Between Fidelity Term and Regularizer

With the proposed method, the diagonal elements of $\boldsymbol{A}_i$ are restricted to be unity so that the off-diagonal elements correspond to the relative leakage levels of bleeding sound. The KL divergence (18) also has a scale-dependent property, namely,

$$\mathcal{D}_{\mathrm{KL}}(\alpha\boldsymbol{X}_i|\alpha\boldsymbol{A}_i\boldsymbol{S}_i) = \alpha\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i)\,, \tag{43}$$

where $\alpha \geq 0$ is an arbitrary coefficient. These facts mean that an observed gain of $\boldsymbol{X}_i$, i.e., the signal amplitude in each microphone, affects the balance of the fidelity term $\sum_i \mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i)$ and regularizer $\sum_{i,m,n\neq m}\mathcal{R}(a_{imn};k,\theta)$ in (31).

To solve this problem, we also parameterize the observed gain. The following normalization is carried out for the observed signal $\tilde{\boldsymbol{x}}(t)$ before we apply the proposed method:

$$\tilde{\boldsymbol{x}}(t) \leftarrow \frac{\alpha}{v}\tilde{\boldsymbol{x}}(t) \quad \forall t, \tag{44}$$

$$v = \max\left(\{\mathrm{abs}(\tilde{\boldsymbol{x}}(t))\}_{t=1}^{T}\right), \tag{45}$$

where $\max(\cdot)$ returns the maximum scalar value of the input set. After the normalization (44), a dynamic range of $\{\tilde{\boldsymbol{x}}(t)\}_{t=1}^{T}$ becomes $\pm\alpha$. Similar to μ in (19), we can control the balance between the fidelity term and regularizer by α. If we set α to a small value, the regularizer strongly affects the optimization.

G. Reconstruction of Estimated Signals

Similar to conventional TCNMF, the complex-valued estimated signal $\mathbf{Y}_n$ can be recovered by applying Wiener filtering to the complex-valued observed signal x_{ijm} as follows:

$$\mathrm{y}_{ijn} = \frac{(a_{imm}s_{imj})^2}{\sum_n (a_{imn}s_{inj})^2}\mathrm{x}_{ijm}. \tag{46}$$

Since $a_{imm} = 1$, (46) can be implemented as

$$\mathrm{y}_{ijn} = \left[\frac{\boldsymbol{S}_i^{\cdot 2}}{\boldsymbol{A}_i^{\cdot 2}\boldsymbol{S}_i^{\cdot 2}}\right]_{m,j}\mathrm{x}_{ijm}, \tag{47}$$

where $[\cdot]_{m,j}$ denotes an (m,j) element of the input matrix. After Wiener filtering, the estimated signal $\mathbf{Y}_n$ is converted to the time-domain signal $\tilde{y}_n(t)$ via the inverse STFT. Then, the signal gain is recovered by

$$\tilde{\boldsymbol{y}}(t) \leftarrow \frac{v}{\alpha}\tilde{\boldsymbol{y}}(t) \quad \forall t. \tag{48}$$

IV. Experiments

A. Conditions

To evaluate the performance of the proposed method (proposed TCNMF), we conducted an experiment of blind bleeding-sound reduction. The observed music mixture signal was simulated using *songKitamura* [27], [28], which is an artificial music dataset. We chose four musical instruments, clarinet (Cl.), oboe (Ob.), piano (Pf.), and trombone (Tb), as dry sources $\mathbf{S}_n$ and prepared a four-channel observed signal $\mathbf{x}_{ij}$ so that $M = N = 4$. To simulate bleeding sound, we mixed these instrumental sounds $\mathbf{s}_{ij}$ using the frequency-wise nonnegative random mixing matrix $\overline{\boldsymbol{A}}_i \in \mathbb{R}_{\geq 0}^{M\times N}$ as follows:

$$\mathbf{x}_{ij} = \overline{\boldsymbol{A}}_i\mathbf{s}_{ij}, \tag{49}$$

where the diagonal and off-diagonal elements of $\overline{\boldsymbol{A}}_i$ are set to unity and uniformly distributed random values in the range $(0, 0.2)$ for all i, respectively. This mixing system is an approximation of (6). In this experiment, ten observed mixtures were prepared using different pseudo-random seeds, i.e., ten different mixing matrices $\overline{\boldsymbol{A}}_i$.

For all signals, we performed STFT using a 4096-point-long hamming window with half-overlap shifting, where a sampling frequency of the signals was 44.1 kHz. The numbers of frequency bins and time frames were $I = 2049$ and $J = 109$, respectively. The update rules in the optimization algorithm were iterated 200 times, and we confirmed the convergence of the cost function value.

For DMNMF and the conventional and proposed TCNMFs, the initial value of $\boldsymbol{A}_i$ was set as follows: the diagonal and off-diagonal elements were set to unity and the uniformly distributed random value in the range $(0, 0.1)$, respectively. The other parameters were initialized by the uniformly distributed random value in the range $(0, 1)$.

As an evaluation criterion, we used the source-to-distortion ratio (SDR) [29], which indicates total separation quality including both degree of separation (source-to-interference ratio: SIR) and absence of artificial distortion (sources-to-artificial ratio: SAR). As described in condition (a) in Sect. I, the SNR and SDR of the observed signals for the bleeding-sound reduction are high. In our experiment, the average SDRs over the ten observed mixture signals of Cl., Ob., Pf., and Tb. were 18.8, 15.0, 14.7, and 8.6 dB, respectively. We calculated the improvements from these input SDRs for each source to evaluate the performance of each method.

We compared five methods, i.e., independent vector analysis (IVA) [9], ILRMA [11], DMNMF [15], the conventional TCNMF [12], and the proposed TCNMF. IVA and ILRMA estimate the complex-valued demixing matrix $\mathbf{W}_i$, thus are

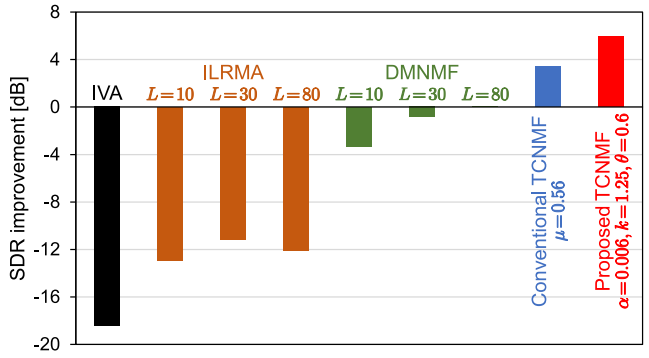

Fig. 5. Comparison of SDR improvements, where each bar is average over 10 different observed mixtures and 4 instrumental sources.

phase-aware BSS methods. The other methods are the phase-insensitive methods that only use amplitude or power spectrograms. The initial value of $\mathbf{W}_i$ for IVA and ILRMA was set to an inverse matrix of the initial mixing matrix used in DMNMF and the conventional and proposed TCNMFs. We also used the numerically stable update rule of the demixing matrix in both IVA and ILRMA, which is called iterative source steering [30], and the estimated source was recovered using (7). We then applied the projection-back technique [31] to the estimated signal to recover the frequency-wise signal scales. For DMNMF and the conventional and proposed TCNMFs, we used Wiener filtering (46) to obtain the estimated source. For ILRMA and DMNMF, the number of basis vectors in the NMF source model, L, was set to 10, 30, and 80.

B. Results

Figure 5 shows the average performance comparison among the five methods, where the hyperparameters of the conventional and proposed TCNMF were experimentally determined and set to $\mu = 0.56$, $k = 1.25$, $\theta = 0.6$, and $\alpha = 0.006$, which provided the best performance in this experiment. We can confirm that the phase-aware BSS methods, IVA and ILRMA, cannot reduce bleeding sound. This is because the observed mixture signal in this experiment was produced using the nonnegative random mixing matrix $\overline{\boldsymbol{A}}_i$ as (49), and the phase information is useless for estimating the demixing matrix. As a result, many artificial distortions are produced in the estimated signals of IVA and ILRMA, degrading their SDR performance. DMNMF has the potential to reduce bleeding sound, but its performance did not exceed 0 dB. This result indicates the difficulty of parameter optimization in DMNMF. For both the conventional and proposed TCNMFs, we can confirm that the average SDR improvements exceed 0 dB. In particular, the proposed TCNMF outperformed the conventional TCNMF by more than 2.5 dB. This improvement is significant to achieve high-quality post-processing or sound reinforcement of a musical performance.

V. Conclusion

We aimed to reduce the bleeding sound in the observed signal obtained with close microphones. We proposed a TCNMF method that regularizes the relative leakage levels of bleeding sounds and is based on MAP estimation with the gamma distribution prior. Experiments using simulated mixture signals showed that the proposed method could achieve the highest bleeding-sound-reduction performance. Since the proposed method has three hyperparameters, an efficient parameter-tuning method is necessary and is for future work.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Numbers 19K20306 and 19H01116.

References

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag Berlin Heidelberg, 2001.
[2] H. L. Van Trees, *Optimum Array Processing*, John Wiley and Sons, New York, 2002.
[3] X. Yu, D. Hu, J. Xu, *Blind Source Separation: Theory and Applications*, John Wiley and Sons, New York, 2014.
[4] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: Two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. Signal and Info. Process.*, vol. 8, no. e12, pp. 1–14, 2019.
[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.
[6] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 2, pp. 666–678, 2006.
[7] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions," *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, pp. 601–608, 2006.
[8] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007.
[9] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 189–192, 2011.
[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.
[11] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, S. Makino, Ed., pp. 125–155. Springer, Cham, 2018.
[12] M. Togami, Y. Kawaguch, H. Kokubo, and Y. Obuchi, "Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization," *Proc. Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf.*, pp. 522–525, 2010.
[13] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," *Proc. Int. Workshop Acoustic Signal Enhancement*, pp. 203–207, 2014.
[14] Y. Murase, H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "On microphone arrangement for multichannel speech enhancement based on nonnegative matrix factorization in time-channel domain," *Proc. Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf.*, 2014.
[15] T. Taniguchi and T. Masuda, "Linear demixed domain multichannel nonnegative matrix factorization for speech enhancement," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 476–480, 2017.
[16] O. Das, J. O. Smith, and J. S. Abel, "Microphone cross-talk cancellation in ensemble recordings with maximum likelihood estimation," *Proc. Audio Eng. Soc. Convention*, 2021.
[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization" *Proc. Neural Info. Process. Syst.*, pp. 556–562, 2000.

[19] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[20] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, 2019.

[21] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Comput.*, vol. 31, no. 9, pp. 1891–1914, 2019.

[22] N. Makishima, Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Independent deeply learned matrix analysis with automatic selection of stable microphone-wise update and fast sourcewise update of demixing matrix," *Signal Process.*, vol. 178, 107753, 2021.

[23] T. Nakamura, S. Kozuka, and H. Saruwatari, "Time-domain audio source separation with neural networks based on multiresolution analysis," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 29, pp. 1687–1701, 2021.

[24] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, no. 785152, 2009.

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

[26] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, 2017.

[27] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 4, pp. 654–669, 2015.

[28] D. Kitamura, "Open dataset: songKitamura," http://d-kitamura.net/dataset_en.html. Accessed 10 July 2021.

[29] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[30] S. Robin and N. Ono, "Fast and stable blind source separation with rank-1 updates," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp.236–240, 2020.

[31] 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.