APSIPA Transactions on Signal and Information Processing, 2025, 14, e18

Original Paper

Music Bleeding-sound Reduction Based on Time-channel Nonnegative Matrix Factorization

Yusaku Mizobuchi[1], Daichi Kitamura[1*], Tomohiko Nakamura[2], Norihiro Takamune[2], Hiroshi Saruwatari[2], Yu Takahashi[3] and Kazunobu Kondo[3]

[1] *National Institute of Technology, Kagawa College, Japan*
[2] *The University of Tokyo, Japan*
[3] *Yamaha Corporation, Japan*

ABSTRACT

When we place microphones close to a sound source near other sources in audio recording, obtained audio signals include undesired sound from the other sources, which is often called 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, 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. In the proposed method, a gamma prior distribution is introduced for the frequency-wise leakage gains of the bleeding sound component to estimate the mixing matrix. This estimation can be interpreted as maximum a posteriori probability estimation. From the experimental results, it is confirmed that the

[*] Corresponding author: kitamura-d@t.kagawa-nct.ac.jp. This work was partly supported by JSPS KAKENHI Grant Numbers 19K20306, 19H01116, and 22H03652.

Received 31 March 2025; revised 19 June 2025; accepted 22 June 2025
ISSN 2048-7703; DOI 10.1561/116.20250013

proposed method can reduce the music bleeding sound with higher accuracy than the other methods in both simulated and real situations. It is also confirmed that the proposed method achieves robust performance against parameter initializations, which is an important advantage in practical applications. The reason of this robustness is experimentally revealed.

Keywords: Blind source separation, bleeding sound, nonnegative matrix factorization, maximum a posteriori estimation

1 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 capture only specific source sounds. 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 Figure 1.

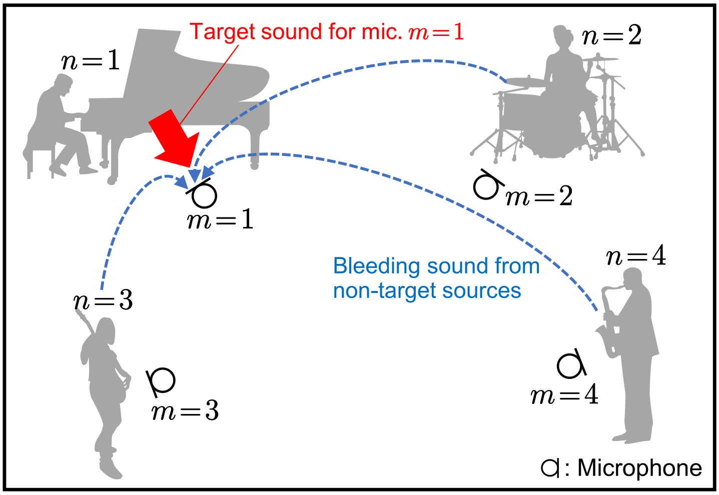

Target sound for mic. $m=1$
$n=1$
$n=2$
$m=1$
$m=2$
Bleeding sound from non-target sources
$n=4$
$n=3$
$m=3$
$m=4$
: Microphone

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

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.

Bleeding-sound reduction is similar to the well-investigated problem called multichannel audio source separation (MASS) [20, 34, 35, 18], 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 [20, 34] and independence-based blind source separation (BSS) [28, 17, 3, 31, 25, 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 [21, 15, 37, 33, 26] can be applied. Togami et al. [21] applied nonnegative matrix factorization (NMF) [7, 6] 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 [15, 37], although its effectiveness regarding music bleeding-sound reduction has not been investigated. A BSS-based method that ignores the phase information was proposed [33], 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. [26] 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 [1, 22, 16, 23, 32, 29] 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 [5], and we demonstrate that the proposed method is suitable for reducing music bleeding sound.

The rest of this paper is organized as follows. In Section 2, we describe the formulation of mixing and demixing systems. Also, phase-aware and unaware BSS algorithms are explained. In Section 3, we propose a new bleeding-sound reduction algorithm and derive its parameter update rules. Simulation-based and realistic experiments for comparing the performance of conventional and proposed methods are provided in Section 4. Also, an experimental analysis for initialization robustness of the proposed method is conducted in Section 5. Finally, conclusions are presented in Section 6. Note that this paper is partially based on an international conference paper [36] we wrote. The major new contributions of this paper are as follows:

- The influence of hyperparameters on the performance of both conventional and proposed methods is experimentally examined in Sections 4.3.2 and 4.1.2.

- New experimental results using impulse response signals recorded in an actual music studio with professionally used apparatuses are presented in Section 4.3, whereas the experiments provided in [36] was simply based on computer-generated artificial mixtures.

- The robustness against the parameter initialization of the proposed method is newly investigated and mathematically analyzed in Section 5.

2 Conventional Methods

In this section, we introduce the mathematical models used in conventional methods for MASS. Throughout the paper, scalars are denoted by regular lowercase letters, vectors by bold lowercase letters, and matrices by bold uppercase letters.

2.1 Formulation of Acoustic Signal

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

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

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

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

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, and $\cdot^{\mathrm{T}}$ denotes a transpose. Under the recording condition described in Section 1, 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.

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

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

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

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 Figure 2. Since the observed signal $\tilde{\boldsymbol{x}}(t)$ is "labeled," as explained in condition (b) in Section 1, 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 Figures 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 Section 1.

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

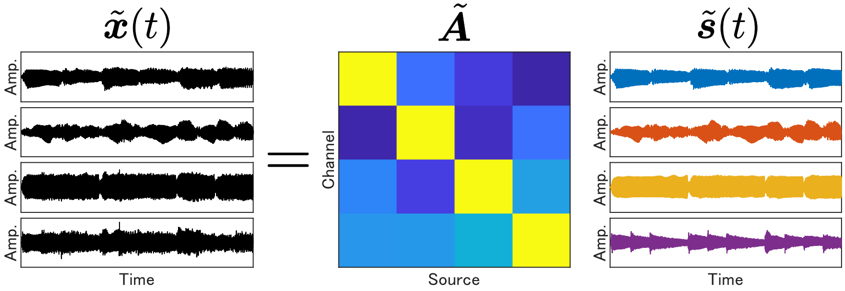


Figure 2: Instantaneous mixture model for bleeding-sound reduction, where $M = N = 4$. Color brightness in mixing matrix $\tilde{\boldsymbol{A}}$ shows amplitude level of each element (brighter is larger). Due to close miking setup, diagonal elements in $\tilde{\boldsymbol{A}}$ have larger amplitudes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

used in the short-time Fourier transform (STFT). This assumption enables us to respectively model the reverberant observed and estimated signals as

$$\boldsymbol{x}_{ij}^{(\mathrm{c})} = \boldsymbol{A}_i^{(\mathrm{c})} \boldsymbol{s}_{ij}^{(\mathrm{c})}, \tag{6}$$

$$\boldsymbol{y}_{ij}^{(\mathrm{c})} = \boldsymbol{W}_i^{(\mathrm{c})} \boldsymbol{x}_{ij}^{(\mathrm{c})}, \tag{7}$$

where

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

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

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

Here, $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$ are the indices of the frequency bin and time frame, respectively, and $\boldsymbol{A}_i^{(\mathrm{c})} \in \mathbb{C}^{M \times N}$ is the complex-valued frequency-wise mixing matrix. Also, $s_{ijn}^{(\mathrm{c})}$, $x_{ijm}^{(\mathrm{c})}$, and $y_{ijn}^{(\mathrm{c})}$ are the complex-valued time-frequency-wise elements of the source, observed, and estimated spectrograms $\boldsymbol{S}_n^{(\mathrm{c})} \in \mathbb{C}^{I \times J}$, $\boldsymbol{X}_m^{(\mathrm{c})} \in \mathbb{C}^{I \times J}$, and $\boldsymbol{Y}_n^{(\mathrm{c})} \in \mathbb{C}^{I \times J}$, respectively. Note that a superscript $\cdot^{(\mathrm{c})}$ denotes the complex-valued variable throughout this paper. In (6), the convolutive mixture is converted to the frequency-wise instantaneous mixture via STFT.

2.2 Phase-aware Method and Spatial Aliasing Problem

Typical beamformers [20, 34] and BSS methods [28, 17, 3, 31, 25, 10, 11] are used to estimate the complex-valued demixing matrix $\boldsymbol{W}_i^{(\mathrm{c})}$ 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. In particular, independent vector analysis (IVA) [3, 31, 25] and independent low-rank matrix analysis (ILRMA) [10, 11] become a common approach of BSS, and many variants of them have been proposed, e.g., [30, 19, 14]. These methods provide significant separation performance when we use a typical microphone array. However, when microphones are spatially apart from each other, phase-aware methods like IVA and ILRMA cannot precisely estimate $\boldsymbol{W}_i^{(\mathrm{c})}$ because of spatial aliasing. This problem is salient in bleeding-sound reduction, as we will confirm in the experimental section.

2.3 DMNMF

To cope with spatial aliasing, the power-based BSS method DMNMF was proposed [33]. DMNMF can be interpreted as a phase-insensitive version of ILRMA [18, 10, 11], and the observed signal is modeled as

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

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

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

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

where the dotted exponent $\cdot^{\cdot u}$ and absolute operation $\mathrm{abs}(\cdot)$ for vectors or matrices return the element-wise uth power and absolute, respectively; thus, $\boldsymbol{x}_{ij}^{\cdot 2}$ and $\boldsymbol{s}_{ij}^{\cdot 2}$ are the power spectrogram components of $\{\boldsymbol{X}_m^{(\mathrm{c})}\}_{m=1}^{M}$ and $\{\boldsymbol{S}_n^{(\mathrm{c})}\}_{n=1}^{N}$, respectively. DMNMF approximates (6) by the nonnegative frequency-wise mixing matrix $\boldsymbol{A}_i^{\cdot 2}$ 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 $\boldsymbol{A}_i^{\cdot 2}$ and $\boldsymbol{s}_{ij}^{\cdot 2}$ from $\boldsymbol{x}_{ij}^{\cdot 2}$, we can recover the estimated signal $\boldsymbol{y}_{ij}^{(\mathrm{c})}$ by Wiener filtering.

2.4 TCNMF

The amplitude-based BSS method TCNMF was proposed [21] and applied [15, 37] 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 Figure 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

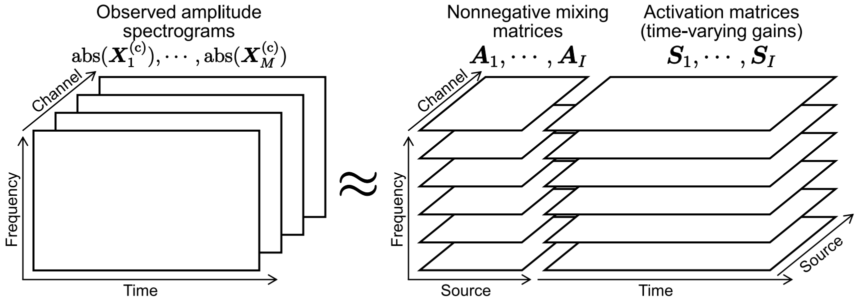

Observed amplitude spectrograms
$\mathrm{abs}(\boldsymbol{X}_1^{(\mathrm{c})}), \cdots, \mathrm{abs}(\boldsymbol{X}_M^{(\mathrm{c})})$
Channel
Frequency
Time
≈
Nonnegative mixing matrices
$\boldsymbol{A}_1, \cdots, \boldsymbol{A}_I$
Activation matrices (time-varying gains)
$\boldsymbol{S}_1, \cdots, \boldsymbol{S}_I$
Channel
Frequency
Source
Time
Source

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

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.

The variables $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ can be estimated by solving the following minimization problem [6]:

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

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. The generative model underlying (17) and (18) is explained in Section 3.2. 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 [21] as follows:

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

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

Update rules of a_{imn} and s_{inj} are derived as

$$a_{imn} \leftarrow a_{imn} \frac{\sum_j \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} s_{inj}}{\sum_j s_{inj}}, \tag{20}$$

$$s_{inj} \leftarrow s_{inj} \frac{\sum_m \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} a_{imn}}{\sum_m a_{imn} + \mu \frac{\sum_{n'} \sqrt{s_{in'j}}}{\sqrt{s_{inj}}}}. \tag{21}$$

The efficient matrix-form implementation of (20) and (21) is as follows:

$$\boldsymbol{A}_i \leftarrow \boldsymbol{A}_i \odot \frac{\frac{\boldsymbol{X}_i}{\boldsymbol{A}_i \boldsymbol{S}_i} \boldsymbol{S}_i^{\mathrm{T}}}{\mathbf{1}^{(N \times J)} \boldsymbol{S}_i^{\mathrm{T}}}, \tag{22}$$

$$\boldsymbol{S}_i \leftarrow \boldsymbol{S}_i \odot \frac{\boldsymbol{A}_i^{\mathrm{T}} \frac{\boldsymbol{X}_i}{\boldsymbol{A}_i \boldsymbol{S}_i}}{\boldsymbol{A}_i^{\mathrm{T}} \mathbf{1}^{(M \times J)} + \mu \frac{\mathbf{1}^{(N \times N)} \boldsymbol{S}_i^{\cdot 1/2}}{\boldsymbol{S}_i^{\cdot 1/2}}}, \tag{23}$$

where $\odot$ and the quotient symbol for matrices denote element-wise multiplication and division, respectively, and $\mathbf{1}^{(\cdot)}$ is a matrix of the size indicated by the superscript, with all elements being equal to one. It is guaranteed that the iterative calculation of (22) and (23) monotonically decreases the cost function in (19).

3 Proposed Method

3.1 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 $\boldsymbol{A}_i^{(\mathrm{c})}$) or a phase-aware BSS method is used for pre-estimation for DMNMF to stabilize and improve BSS performance [33]. TCNMF can estimate the source signals without phase information, even in asynchronous recording [15]. 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 [27], 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 address this issue, we propose the introduction of a novel regularization term specifically tailored for bleeding-sound reduction. Considering the characteristics (a) and (b) described in Section 1, both the diagonal and off-diagonal elements of the nonnegative mixing matrix $\boldsymbol{A}_i$ should be subject to regularization instead of $\boldsymbol{S}_i$. Such regularization also eliminates the risk of obtaining trivial solution of $\boldsymbol{A}_i$ in TCNMF. 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.

3.2 Generative Model of KL-divergence-based NMF

Cemgil [5] 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:

$$z_{imnj} \sim \mathcal{P}(z_{imnj}; a_{imn}s_{inj}), \tag{24}$$

$$\mathcal{P}(z;\lambda) = \frac{1}{\Gamma(z+1)} e^{-\lambda}\lambda^{z}, \tag{25}$$

where $z_{imnj} \in \mathbb{N}$ is a random variable that satisfies $x_{imj} = b + \sum_n z_{imnj}$, $\mathcal{P}(z;\lambda)$ 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 b 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

$$\begin{aligned}
&\log p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i)\\
&\quad = \log \prod_{m,j} \sum_{z_{imnj}} p(x_{inm}; z_{imnj}) p(z_{imnj}; a_{imn}s_{inj})\\
&\quad = \log \prod_{m,j} \mathcal{P}\left(x_{imj}; \sum_n a_{imn}s_{inj}\right)\\
&\quad = \sum_{m,j} \left[x_{imj} \log \sum_n a_{imn}s_{inj} - \sum_n a_{imn}s_{inj} - \log\Gamma(x_{imj}+1) \right].
\end{aligned} \tag{26}$$

The maximization of (26) w.r.t. a_{imn} and s_{inj} for all i (ML estimation) is equivalent to the minimization of (18).

3.3 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 $\boldsymbol{A}_i$:

$$a_{imn} \sim \begin{cases} \delta(a_{imn} - 1) & (m = n) \\ \mathcal{G}(a_{imn}; k, \theta) & (m \neq n) \end{cases}, \tag{27}$$

$$\mathcal{G}(a; k, \theta) = \frac{1}{\Gamma(k)\theta^k} a^{k-1} e^{-a/\theta}, \tag{28}$$

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 (24). 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

$$\begin{aligned} p(\boldsymbol{A}_i; k, \theta) &= \prod_{m,n=m} p(a_{imn}) \prod_{m,n\neq m} p(a_{imn}; k, \theta) \\ &= \prod_{m,n=m} \delta(a_{imn} - 1) \prod_{m,n\neq m} \mathcal{G}(a_{imn}; k, \theta). \end{aligned} \tag{29}$$

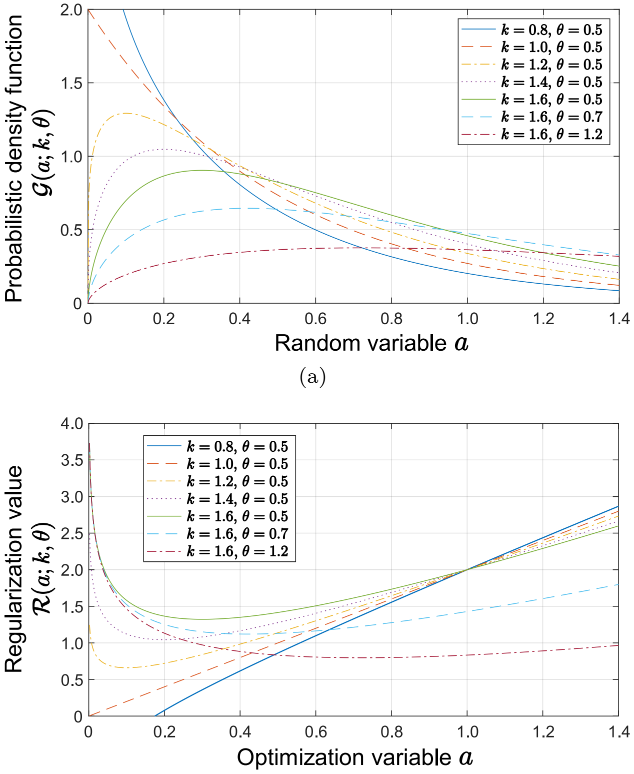
The prior (29) enables us to control the probability of off-diagonal elements of $\boldsymbol{A}_i$ (relative leakage levels of bleeding sound) by k and θ, while restricting all the diagonal elements to be unity. As shown in Figure 4, we can avoid $a_{imn} = 0$ for all $m \neq n$, which is the trivial 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:

$$\begin{aligned} s_{inj} &\sim \lim_{\beta \to \infty} \frac{1}{\beta} \mathcal{I}[0 \leq s_{inj} \leq \beta] \\ &\propto \mathcal{I}[0 \leq s_{inj}], \end{aligned} \tag{30}$$

where β is the normalized coefficient and $\mathcal{I}[\cdot]$ denotes a binary-valued function 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

$$\begin{aligned} p(\boldsymbol{S}_i) &= \prod_{n,j} p(s_{inj}) \\ &\propto \prod_{n,j} \mathcal{I}[0 \leq s_{inj}]. \end{aligned} \tag{31}$$



(a)

(b)

Figure 4: Shape of (a) probabilistic density function of gamma distribution and (b) its corresponding regularization function.

3.4 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

$$\prod_i p(\boldsymbol{A}_i, \boldsymbol{S}_i; \boldsymbol{X}_i) \propto \prod_i \underbrace{p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i)}_{\text{Likelihood}} \underbrace{p(\boldsymbol{A}_i; k, \theta) p(\boldsymbol{S}_i)}_{\text{Priors}}. \tag{32}$$

By taking a negative logarithm of (32), we can decompose the right side of (32) as

$$\mathcal{J} = -\sum_i \left[\log p(\boldsymbol{X}_i; \boldsymbol{A}_i, \boldsymbol{S}_i) + \log p(\boldsymbol{A}_i; k, \theta) + \log p(\boldsymbol{S}_i)\right]. \tag{33}$$

From (26), (29), and (31), the cost function $\mathcal{J}$ is obtained as

$$\begin{aligned}\mathcal{J} = &\sum_{i,m,j}\left[-x_{imj}\log\sum_{n} a_{imn}s_{inj} + \sum_{n} a_{imn}s_{inj} + \log\Gamma(x_{imj}+1)\right]\\ &+ \sum_{i,m,n=m}\mathbb{I}[a_{imn}=1] \; + \sum_{i,m,n\neq m}\left[-(k-1)\log a_{imn} + \frac{1}{\theta}a_{imn}\right]\\ &+ \sum_{i,n,j}\mathbb{I}[0\leq s_{inj}], \qquad (34)\end{aligned}$$

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 (34), and this minimization w.r.t. $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$ is equivalent to the following problem:

$$\begin{aligned}&\min_{\mathcal{A},\mathcal{S}}\sum_{i}\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i) + \sum_{i,m,n\neq m}\mathcal{R}(a_{imn};k,\theta)\\ &\text{s.t. } a_{imn}, s_{inj}\geq 0\ \forall i,m,n,j \text{ and } a_{imn}=1\ \forall m=n, \qquad (35)\end{aligned}$$

where

$$\mathcal{R}(a_{imn};k,\theta) = \left[-(k-1)\log a_{imn} + \frac{1}{\theta}a_{imn}\right] \qquad (36)$$

is the regularizer that corresponds to the gamma distribution prior (28) for the off-diagonal elements of $\boldsymbol{A}_i$.

3.5 Derivation of Optimization Algorithm

The minimization problem (35) can be solved using a majorization-minimization (MM) algorithm [6, 38], 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:

$$\begin{aligned}&\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i)\\ &\overset{\mathrm{c}}{=}\sum_{i,m,j}\left(-x_{imj}\log\sum_{n} a_{imn}s_{inj} + \sum_{n} a_{imn}s_{inj}\right)\\ &=\sum_{i,m,j}\left(-x_{imj}\log\sum_{n}\xi_{imnj}\frac{a_{imn}s_{inj}}{\xi_{imnj}} + \sum_{n} a_{imn}s_{inj}\right)\\ &\leq\sum_{i,m,j}\left(-x_{imj}\sum_{n}\xi_{imnj}\log\frac{a_{imn}s_{inj}}{\xi_{imnj}} + \sum_{n} a_{imn}s_{inj}\right)\\ &\equiv\mathcal{D}^{+}(\boldsymbol{A}_i,\boldsymbol{S}_i,\Xi), \qquad (37)\end{aligned}$$

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 (37) holds if and only if

$$\xi_{imnj} = \frac{a_{imn} s_{inj}}{\sum_{n'} a_{imn'} s_{in'j}} \ \forall i, m, j, n. \tag{38}$$

From (37), the MM problem is obtained as

$$\begin{aligned} &\min_{\mathcal{A},\mathcal{S},\Xi} \sum_i \mathcal{D}^+(\boldsymbol{A}_i, \boldsymbol{S}_i, \Xi) + \sum_{i,m,n\neq m} \mathcal{R}(a_{imn}; k, \theta) \\ &\text{s.t. } a_{imn}, s_{inj} \geq 0 \ \forall i, m, n, j, \quad \xi_{imnj} > 0 \ \forall i, m, n, j, \quad \sum_n \xi_{imnj} = 1 \ \forall i, m, j, \\ &\qquad \text{and } a_{imn} = 1 \ \forall m = n. \end{aligned} \tag{39}$$

By setting the derivative of the majorization function (39) w.r.t. a_{imn} and s_{inj} to zero and substituting (38) 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 [6] and expressed as

$$s_{inj} \leftarrow s_{inj} \frac{\sum_m \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} a_{imn}}{\sum_m a_{imn}}. \tag{40}$$

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

$$\sum_j \left(-x_{imj} \frac{\xi_{imnj}}{a_{imn}} + s_{inj} \right) - (k-1) \frac{1}{a_{imn}} + \frac{1}{\theta} = 0. \tag{41}$$

Therefore, we have

$$a_{imn} = \frac{(k-1) + \sum_j x_{imj} \xi_{imnj}}{\frac{1}{\theta} + \sum_j s_{inj}}. \tag{42}$$

The update rule of the off-diagonal elements a_{imn} is derived by substituting (38) as

$$a_{imn} \leftarrow \frac{(k-1) + a_{imn} \sum_j \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} s_{inj}}{\frac{1}{\theta} + \sum_j s_{inj}}. \tag{43}$$

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 (40) and (43) is as follows:

$$\boldsymbol{A}_i \leftarrow \frac{(k-1)+\boldsymbol{A}_i \odot \left(\frac{\boldsymbol{X}_i}{\boldsymbol{A}_i\boldsymbol{S}_i}\boldsymbol{S}_i^{\mathrm{T}}\right)}{\frac{1}{\theta}+\boldsymbol{1}^{(N\times J)}\boldsymbol{S}_i^{\mathrm{T}}} \quad \forall i, \tag{44}$$

$$\mathrm{diag}(\boldsymbol{A}_i) \leftarrow [1,1,\cdots,1]^{\mathrm{T}} \quad \forall i, \tag{45}$$

$$\boldsymbol{S}_i \leftarrow \boldsymbol{S}_i \odot \frac{\boldsymbol{A}_i^{\mathrm{T}}\frac{\boldsymbol{X}_i}{\boldsymbol{A}_i\boldsymbol{S}_i}}{\boldsymbol{A}_i^{\mathrm{T}}\boldsymbol{1}} \quad \forall i, \tag{46}$$

where $\mathrm{diag}(\cdot)$ returns a vector that consists of the diagonal elements of the input square matrix. Note that (44) will change the value of the diagonal elements of $\boldsymbol{A}_i$, but they are immediately replaced with unity by (45). It is guaranteed that the iterative calculation of (44)–(46) monotonically decreases the cost function (34).

3.6 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,

$$\mathcal{D}_{\mathrm{KL}}(\alpha\boldsymbol{X}_i|\alpha\boldsymbol{A}_i\boldsymbol{S}_i) = \alpha\mathcal{D}_{\mathrm{KL}}(\boldsymbol{X}_i|\boldsymbol{A}_i\boldsymbol{S}_i)\,, \tag{47}$$

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 (35).

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:

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

$$v = \max\bigl(\{\mathrm{abs}(\tilde{\boldsymbol{x}}(t))\}_{t=1}^{T}\bigr)\,, \tag{49}$$

where $\max(\cdot)$ returns the maximum scalar value of the input set. After the normalization (48), 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.

3.7 Reconstruction of Estimated Signals

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

$$y_{ijn}^{(\mathrm{c})} = \frac{(a_{imm}s_{imj})^2}{\sum_n (a_{imn}s_{inj})^2} x_{ijm}^{(\mathrm{c})}. \tag{50}$$

Since $a_{imm} = 1$, (50) can be implemented as

$$y_{ijn}^{(\mathrm{c})} = \left[\frac{\boldsymbol{S}_i^{.2}}{\boldsymbol{A}_i^{.2}\boldsymbol{S}_i^{.2}} \right]_{m,j} x_{ijm}^{(\mathrm{c})}, \tag{51}$$

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

$$\tilde{\boldsymbol{y}}(t) \leftarrow \frac{v}{\alpha}\tilde{\boldsymbol{y}}(t) \quad \forall t. \tag{52}$$

4 Experimental Results and Discussion

To assess the efficacy of the proposed method, we conducted three experiments to evaluate blind bleeding-sound reduction: (i) a simulation-based experiment using randomly produced various mixing matrices $\boldsymbol{A}_i$, (ii) simulation-based experiment using various sound sources produced by a musical instrument digital interface (MIDI), and (iii) realistic experiment employing impulse responses obtained from an actual music studio with professionally used apparatuses and a dataset comprising professionally produced music recordings. In both experiments, we compared five methods, i.e., IVA [25], ILRMA [11], DMNMF [33], the conventional TCNMF [21], and the proposed method. Hereafter, we respectively denote the conventional and proposed TCNMF methods by $L_{0.5}$TCNMF and GammaTCNMF throughout the paper. IVA and ILRMA estimate the complex-valued demixing matrix $\boldsymbol{W}_i^{(\mathrm{c})}$, thus are phase-aware BSS methods. The other methods are the phase-insensitive methods that only use amplitude or power spectrograms, so it is likely to work even when spatial aliasing problems occur.

4.1 Simulation-based Experiment Using Various Random Mixing Matrices

4.1.1 Conditions

The observed music mixture signal was simulated using *songKitamura* [9, 8], which is an MIDI-based artificial music dataset. We chose four musical instruments, clarinet (Cl.), oboe (Ob.), piano (Pf.), and trombone (Tb.), as dry sources $\boldsymbol{S}_n^{(\mathrm{c})}$ and prepared a four-channel observed signal $\boldsymbol{x}_{ij}^{(\mathrm{c})}$ so that $M = N = 4$. To simulate bleeding sound, we mixed these instrumental sounds $\boldsymbol{s}_{ij}^{(\mathrm{c})}$ using the frequency-wise nonnegative random mixing matrix $\overline{\boldsymbol{A}}_i \in \mathbb{R}_{\geq 0}^{M \times N}$ as follows:

$$\boldsymbol{x}_{ij}^{(\mathrm{c})} = \overline{\boldsymbol{A}}_i \boldsymbol{s}_{ij}^{(\mathrm{c})}, \tag{53}$$

where the diagonal and off-diagonal elements of $\overline{\boldsymbol{A}}_i$ were set to unity and uniformly distributed random values in the range $(0, 0.2)$ for all i, respectively. In this experiment, 10 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.

For DMNMF, $L_{0.5}$-TCNMF, and Gamma-TCNMF, 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 $(0, 1)$. The initial value of $\boldsymbol{W}_i^{(\mathrm{c})}$ for IVA and ILRMA was set to an inverse matrix of the initial mixing matrix used in DMNMF, $L_{0.5}$-TCNMF, and Gamma-TCNMF. 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 [24] to the estimated signal to recover the frequency-wise signal scales. For ILRMA and DMNMF, the number of basis vectors in the NMF source model, L, was set to 10, 30, and 80.

As an evaluation criterion, we used the source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifact ratio (SAR), which can be calculated using the `bss_ eval_sources` function implemented in `BSS_EVAL_Toolbox` [13]. SIR and SAR are used to quantify the amount of interference rejection and the absence of artificial distortion of the estimated signal, respectively. SDR is used to quantify the overall separation performance, as SDR is in good agreement with both SIR and SAR for BSS. As described in condition (a) in Section 1, the SDR of the observed signals (input SDR) is relatively high. We calculated the improvements from the input SDR for each source to evaluate the performance of each method.

4.1.2 Influence of Hyperparameters

$L_{0.5}$-TCNMF has a hyperparameter μ which controls the intensity of the regularization in (19). Figure 5 shows SDR improvements obtained by $L_{0.5}$-TCNMF with various weight coefficients μ, where each score is the average

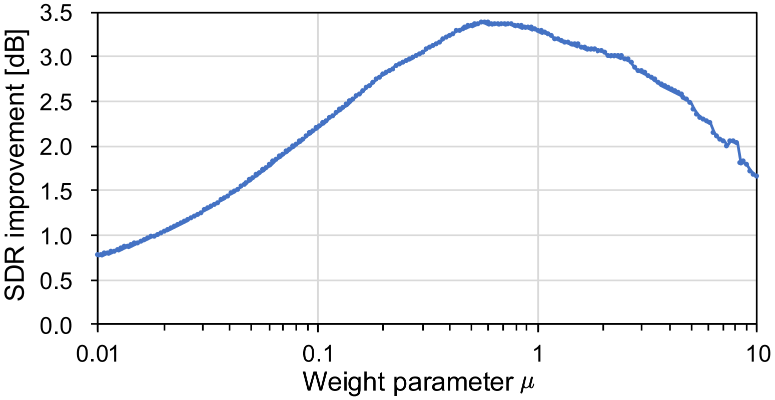


Figure 5: SDR improvements for simulated data using various random mixing matrices obtained by $L_{0.5}$-TCNMF with various weight coefficients μ, where each plot is average over 10 different observed mixtures and four instrumental sources.

over ten random mixing matrices and four instrumental sources. We changed μ 200 times on logarithmic scale in the range [0.01, 10]; thus, we can confirm that the conventional TCNMF can achieve more than 3 dB SDR improvement when we select the appropriate hyperparameters. $L_{0.5}$-TCNMF performs best for $\mu = 0.56$ under these conditions.

Gamma-TCNMF has three hyperparameters, k, θ, and α. For the shape and scale parameters, k and θ, the shape of the probabilistic density function $\mathcal{G}(a_{imn}; k, \theta)$ or the regularizer $\mathcal{R}(a_{imn}; k, \theta)$, which are respectively illustrated in Figure 4 (a) or (b), may be useful for the hyperparameter tuning. Since the off-diagonal elements of $\boldsymbol{A}_i$ represent the relative leakage levels of bleeding sound, they should be in the appropriate range, e.g., $[0.05, 0.6]$. On the basis of this range, we can tune k and θ. However, α directly affects the performance of Gamma-TCNMF because this parameter controls the intensity of the regularization as well as μ in (19) of $L_{0.5}$-TCNMF.

Figure 6 shows average SDR improvements of Gamma-TCNMF with various hyperparameter settings. We chose three values for each shape and scale parameter, namely, $k = 1.10$, 1.25, and 1.60 and $\theta = 0.3$, 0.6, and 1.3, resulting in nine patterns, as shown in Figure 6. We can confirm that Gamma-TCNMF can achieve more than 5 dB SDR improvement when we select the appropriate hyperparameters. Gamma-TCNMF performs best for $k = 1.25$, $\theta = 0.6$, and $\alpha = 0.006$ under these conditions. Although the optimal k and θ depend on the weight parameter α, the achievable best performance is almost the same for different parameter settings.

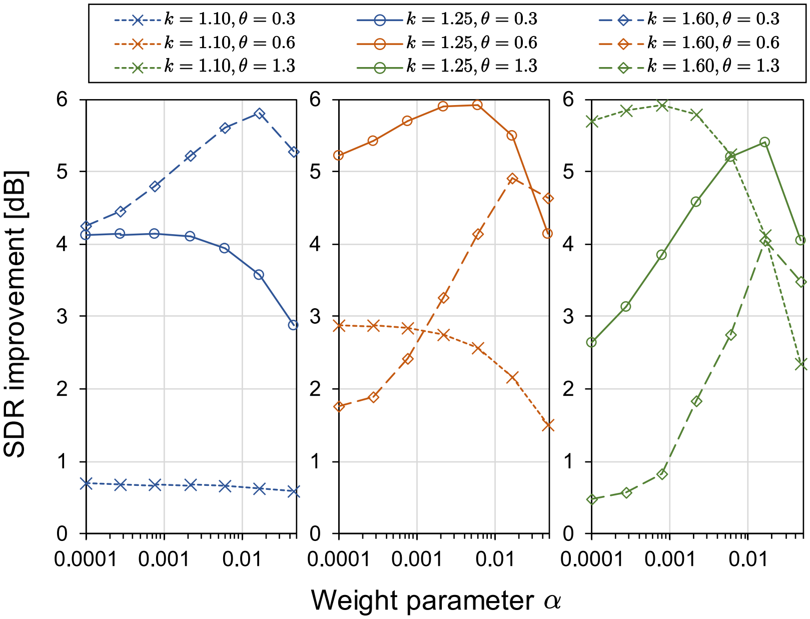


Figure 6: SDR improvements for simulated data using various random mixing matrices obtained by Gamma-TCNMF with various α, k, and θ, where each plot is average over 10 different observed mixtures and four instrumental sources.

4.1.3 Performance Comparison

Figure 7 shows the comparison of SDR improvements among the five methods, where the hyperparameters of $L_{0.5}$- and Gamma-TCNMFs were set to their optimal values. The violin plots in Figure 7 shows the distributions of the results for 10 different random mixing matrices. The white circle indicates a median value, the gray vertical line shows the range of 25–75 percentiles, and the violin curve is an estimated distribution. In addition, Table 1 summarizes the average evaluation scores for the input, output, and improvements. The input SDR and SIR refer to the scores of the observed signals recorded by close microphones for each source, while the output SDR, SIR, and SAR represent the scores of the estimated signals. The improvement is calculated as the difference between the output and input scores. Since SAR measures the absence of artifacts introduced by BSS, the input SAR is infinity. From these results, we can confirm that the phase-aware BSS methods, IVA and ILRMA, cannot reduce the 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 (53), and the phase information is useless for estimating the demixing matrix. As a result, the demixing matrix $\boldsymbol{W}_i^{(\mathrm{c})}$ estimated by

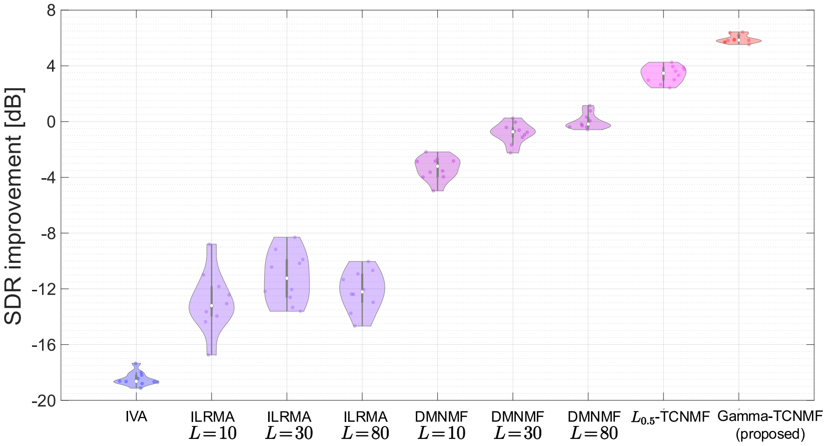


Figure 7: Violin plots of SDR improvements for simulated data using various random mixing matrices. In each method, white circle indicates median value, gray vertical line shows range of 25–75 percentiles, and violin curve is estimated distribution.

IVA or ILRMA contains many errors, which simultaneously degrade SIR and introduce harmful distortions into the estimated signal $\boldsymbol{y}_{ij}^{(\mathrm{c})}$. The SNR of the observed signal is relatively high in the context of bleeding-sound reduction, resulting in higher input SDR/SIR as shown in Table 1. In such situations, even slight errors in the estimated signals can lead to significant SIR and SAR degradations, resulting in the poor SDR improvements observed for IVA and ILRMA in Figure 7. DMNMF has the potential to reduce bleeding sound, but its SDR improvements did not substantially exceed 0 dB. This result indicates the difficulty of parameter optimization in DMNMF. For both $L_{0.5}$- and Gamma-TCNMFs, we can confirm that the average SDR improvements exceed 0 dB. In particular, Gamma-TCNMF outperformed $L_{0.5}$-TCNMF by more than 2 dB. This improvement is significant to achieve high-quality post-processing or sound reinforcement of a musical performance.

4.2 Simulation-based Experiment Using Various Source Signals

In this subsection, we evaluated the performance using various MIDI-based source signals, where the mixing matrix $\overline{\boldsymbol{A}}_i$ was fixed to a single matrix generated in the same manner as described in Section 4.1.1. Furthermore, to ensure equitable comparison of the performance, the optimal hyperparameters for $L_{0.5}$- and Gamma-TCNMFs were experimentally determined using a development dataset. Subsequently, the performance of each method was assessed using a separate test dataset.

Table 1: Average SDR, SIR, and SAR values for simulated data using various random mixing matrices.

Method	Input	Output	SDR	Input	Output	SIR	Output
	SDR	SDR	imp.	SIR	SIR	imp.	SAR
	[dB]	[dB]	[dB]	[dB]	[dB]	[dB]	[dB]
IVA	14.27	-4.18	-18.45	14.44	-2.80	-17.24	8.90
ILRMA (L=10)		1.35	-12.92		4.45	-10.00	14.45
ILRMA (L=30)		3.09	-11.18		6.41	-8.03	14.42
ILRMA (L=80)		2.15	-12.12		5.52	-8.92	13.93
DMNMF (L=10)		10.92	-3.35		22.08	7.63	11.46
DMNMF (L=30)		13.44	-0.83		24.46	10.01	13.90
DMNMF (L=80)		14.30	0.03		**25.18**	**10.74**	14.72
$L_{0.5}$-TCNMF		17.65	3.38		19.86	5.42	22.30
Gamma-TCNMF (proposed)		**20.20**	**5.93**		23.27	8.82	**23.34**

4.2.1 Conditions

To confirm the efficacy for various source signals, we used the dataset called *Slakh2100-redux* [12], which contains 1710 songs featuring at least four sources: Pf., bass (Ba.), guitar (Gt.), and drums (Dr.). We extracted the segments lasting 30 to 60 seconds from all songs for these four sources. From these segments, we randomly selected 30 songs in which each source is active for at least 80% of the duration. Among them, 10 songs were used as development data, and the remaining 20 were used as test data. Other experimental conditions were the same as those described in Section 4.1.1. The numbers of frequency bins and time frames were $I = 2049$ and $J = 647$, respectively.

4.2.2 Experimental Analysis of Optimal Hyperparameters Using Development Dataset

To determine the optimal hyperparameter settings for $L_{0.5}$- and Gamma-TCNMFs, we used a development dataset comprising 10 songs. Figures 8 and 9 represent SDR improvements for various hyperparameter settings in $L_{0.5}$- and Gamma-TCNMFs, under the same conditions as those described in Section 4.1.1. Similar to the trends observed in Figures 5 and 6, both methods exhibit significant SDR improvements when their hyperparameters are optimally tuned. The optimal settings identified in Figures 8 and 9 ($\mu = 0.78$ for $L_{0.5}$-TCNMF and $k = 1.1$, $\theta = 1.3$, and $\alpha = 0.0002$ for Gamma-TCNMF) were used in subsequent experiments using test data.

4.2.3 Performance Comparison Using Test Dataset

Figure 10 shows the comparison of SDR improvements among the five methods, where each violin plot includes the results of 20 songs. Also, Table 2 summarizes the average evaluation scores. In this experiment, both $L_{0.5}$ and GammaTCNMFs achieve comparable SDR improvements. However, the re-

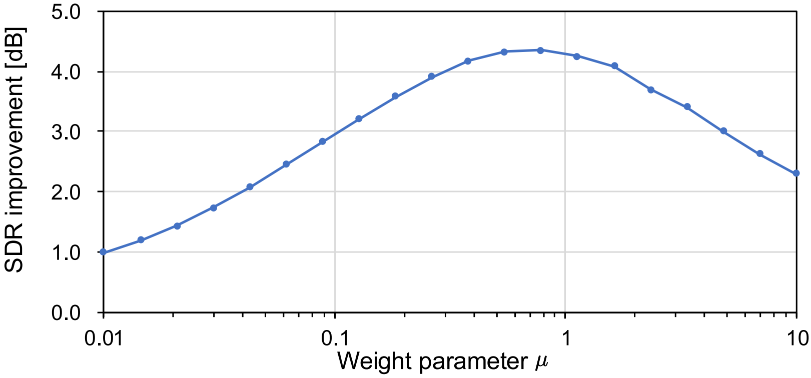


Figure 8: SDR improvements for simulated data (development data) using various source signals obtained by $L_{0.5}$-TCNMF with various weight coefficients μ, where each plot is average over 10 different source signals and four instrumental sources.

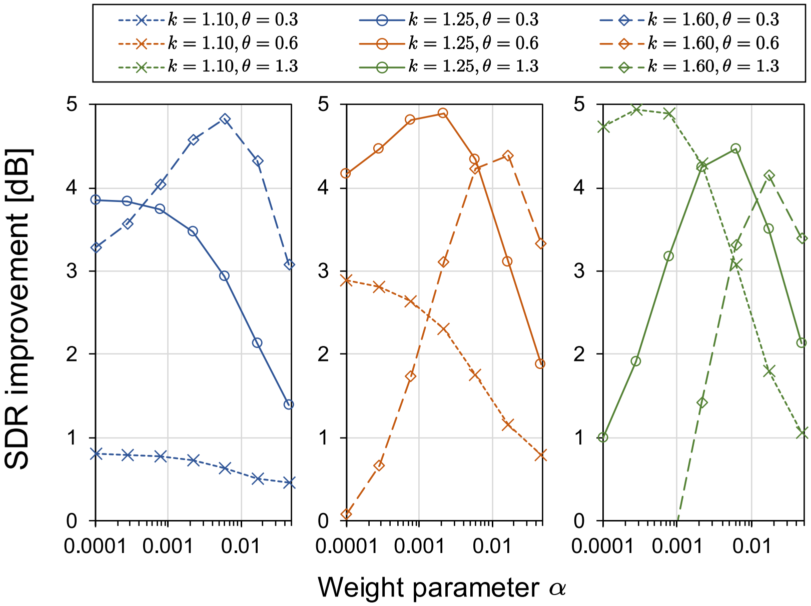


Figure 9: SDR improvements for simulated data (development data) using various source signals obtained by Gamma-TCNMF with various α, k, and θ, where each plot is average over 10 different source signals and four instrumental sources.

sults in Table 2 show that GammaTCNMF outperforms $L_{0.5}$TCNMF by approximately 0.58 dB on average, demonstrating its superiority for the various source signals.

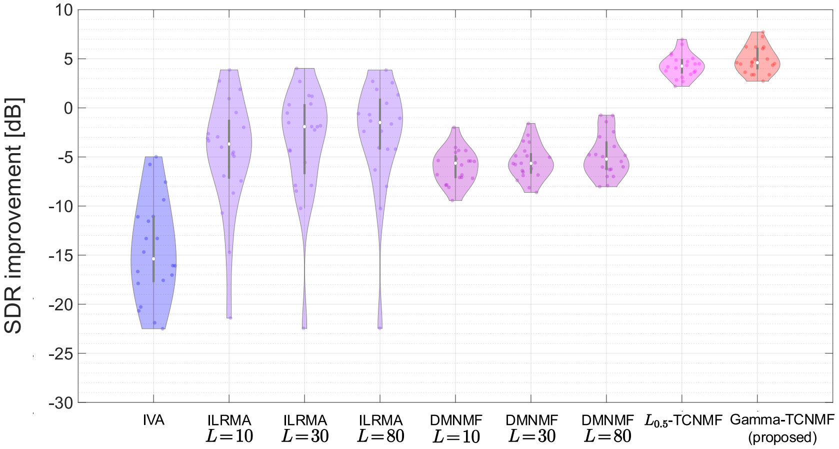


Figure 10: Violin plots of SDR improvements for simulated data using various source signals. In each method, white circle indicates median value, gray vertical line shows range of 25–75 percentiles, and violin curve is estimated distribution.

Table 2: Average SDR, SIR, and SAR values for simulated data using various source signals (test data).

Method	Input SDR [dB]	Output SDR [dB]	SDR imp. [dB]	Input SIR [dB]	Output SIR [dB]	SIR imp. [dB]	Output SAR [dB]
IVA	14.72	5.20	-14.46	15.16	9.10	-6.06	4.16
ILRMA ($L=10$)		10.07	-4.65		23.01	7.85	11.65
ILRMA ($L=30$)		11.16	-3.56		24.64	9.48	12.51
ILRMA ($L=80$)		11.89	-2.84		25.17	10.01	13.17
DMNMF ($L=10$)		8.72	-6.00		24.93	9.77	8.89
DMNMF ($L=30$)		9.22	-5.50		25.14	9.97	9.42
DMNMF ($L=80$)		9.84	-4.88		**25.56**	**10.40**	10.03
$L_{0.5}$-TCNMF		19.02	4.30		22.39	7.23	**21.97**
Gamma-TCNMF (proposed)		**19.60**	**4.88**		24.30	9.14	21.58

4.3 Realistic Experiment Using Impulse Responses

In this subsection, to imitate actual bleeding sounds within the observed signals, impulse responses were measured at an authentic music studio environment. In addition, a dataset comprising professionally produced music recordings was used as dry source signals.

4.3.1 Conditions

As the dry source signals, we used 20 songs randomly selected from *DSD100* [4], divided equally into 10 development songs and 10 test songs. This dataset consists of full lengths music tracks along with their isolated Dr., Ba., vocals and others signals. The dry sources of each source were extracted from segments lasting 60 to 90 seconds from each track.

To replicate realistic bleeding sounds observed in professional musical performances or music recordings, impulse responses were measured at an actual music studio with professionally used apparatuses. The setup for this measurement is depicted in Figure 11. Within this recording environment, the audio captured by the first microphone (Microphone 1 in Figure 11) was input to a mixing console and subsequently emitted through a monitor loudspeaker positioned at the corner of the room. Such conditions simulate the common setup found in professional music studios. As a result, the first source (Source 1 in Figure 11) is recorded with a larger volume as bleeding sound into the other microphones compared to the other sources. The reverberation time of these impulse responses was around $T_{60} = 330$ ms.

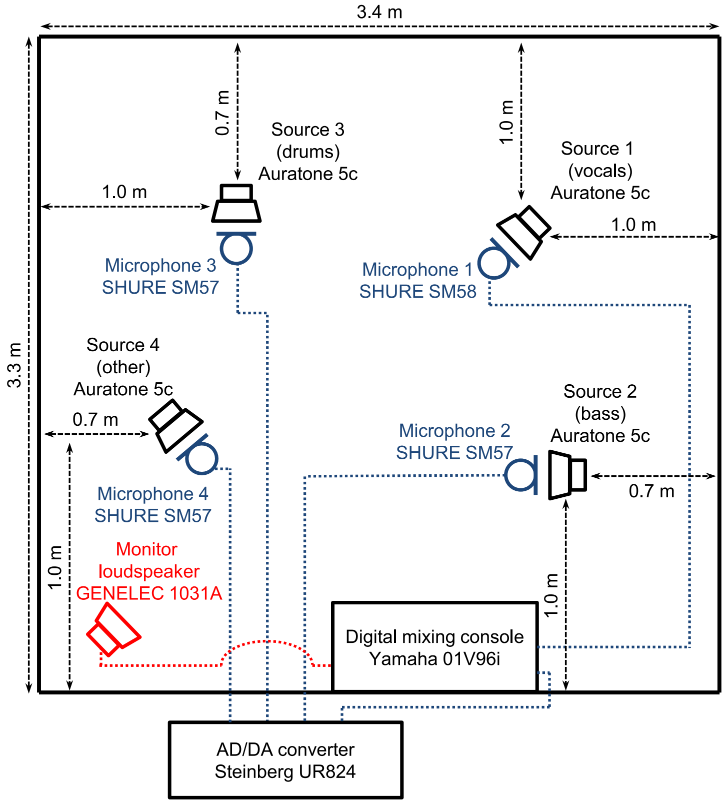


Figure 11: Recording environment of impulse response signals, which consists of an actual music studio with professionally used apparatuses. Only first source (Source 1) is amplified and emitted by monitor loudspeaker.

To set the initial values of each method, we used the same manner as those in Section 4.1.1. We used 10 pseudo-random seeds for these initializations, resulting in 10 results for each of 10 songs. The other experimental conditions were the same as those in Section 4.1.1. The numbers of frequency bins and time frames were $I = 2049$ and $J = 647$, respectively.

4.3.2 Experimental Analysis of Optimal Hyperparameters Using Development Dataset

To determine the optimal hyperparameter settings for $L_{0.5}$- and Gamma-TCNMFs, we used a development dataset comprising 10 songs. Figures 12 and 13 show the average SDR improvements achieved by $L_{0.5}$- and Gamma-TCNMFs, respectively, across different hyperparameter settings. Under these experimental conditions, $L_{0.5}$-TCNMF yielded unsatisfactory results, whereas Gamma-TCNMF exhibited a notable improvement of over 1.5 dB with appropriate hyperparameter selections. The best hyperparameters obtained from this experiment were $\mu = 0.0749$ for $L_{0.5}$-TCNMF, and $k = 1.02$, $\theta = 2.15$, and $\alpha = 0.0008$ for Gamma-TCNMF.

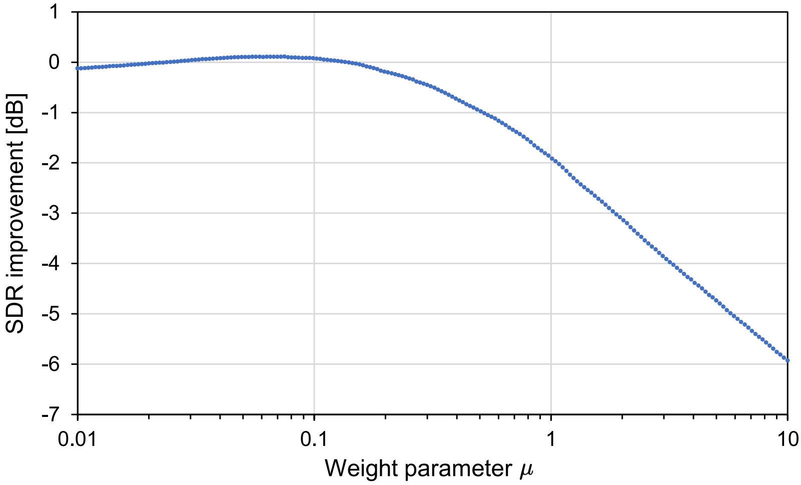


Figure 12: SDR improvements for realistic data (development data) obtained by $L_{0.5}$-TCNMF with various weight coefficients μ, where each plot is average over 10 songs, 10 random seeds, and four instrumental sources.

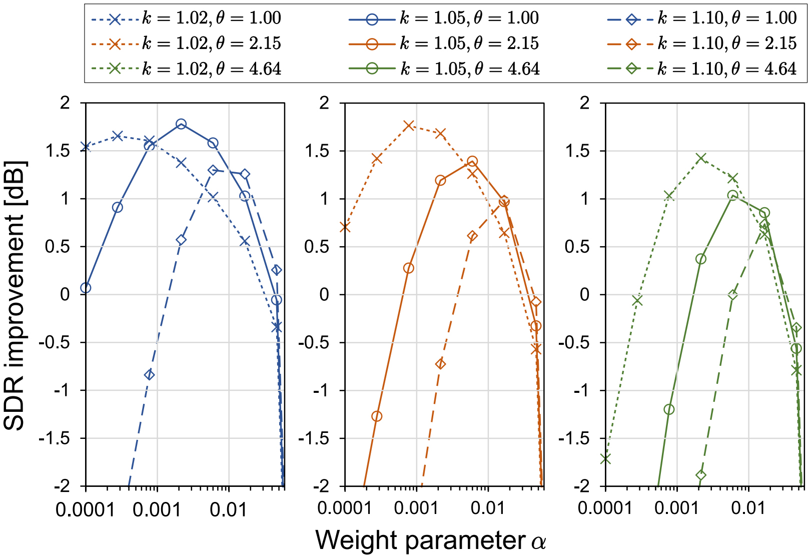


Figure 13: SDR improvements for realistic data (development data) obtained by Gamma-TCNMF with various α, k, and θ, where each plot is average over 10 songs, 10 random seeds, and four instrumental sources.

4.3.3 Performance Comparison Using Test Dataset

Figure 14 shows the results of SDR improvements comparing the performance of five methods on the test 10 songs. Also, Figure 15 shows only the results of $L_{0.5}$- and Gamma-TCNMFs in Figure 14. The colored dots represent the average of SDR improvement of four sources, and there are 100 plots in each method, including 10 songs with 10 random initialization patterns. Furthermore, Table 3 summarizes the average evaluation scores. We can confirm that the phase-aware BSS methods, IVA and ILRMA, cannot reduce the bleeding sound in the observed signal. This is because spatial aliasing problems occurred due to the distance between the microphones, as depicted in Figure 11. DMNMF also fails to achieve bleeding-sound reduction. Compared to the results in Section 4.1.3, the performance of both $L_{0.5}$- and Gamma-TCNMFs has deteriorated. However, Gamma-TCNMF still achieves SDR improvement scores exceeding 0 dB, demonstrating the effectiveness even in realistic environments.

In addition, to rigorously compare $L_{0.5}$- and Gamma-TCNMFs, 100 random initialization patterns were applied to each of the 10 songs in the test dataset. The average and standard deviation (SD) values obtained by this experiment are summarized in Table 4. This result also shows the efficacy of Gamma-TCNMF. Furthermore, we can confirm that the SD of Gamma-

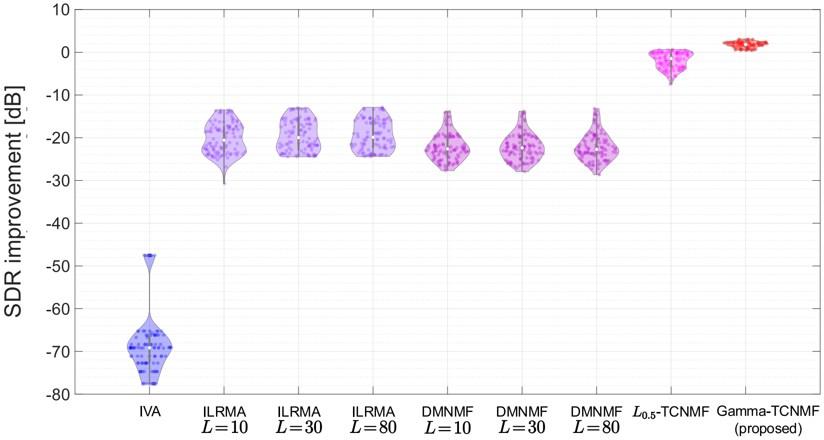


Figure 14: Violin plots of SDR improvements for realistic data (test data). In each method, white circle indicates median value, gray vertical line shows range of 25–75 percentiles, and violin curve is estimated distribution.

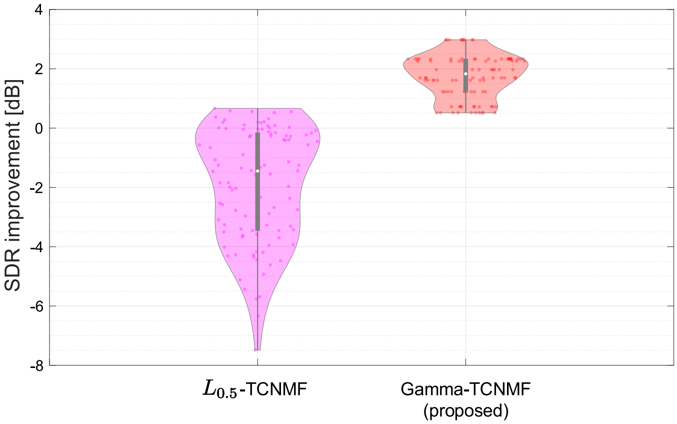


Figure 15: Violin plots of SDR improvements for realistic data (test data) obtained by $L_{0.5}$- and Gamma-TCNMFs.

TCNMF is significantly small in all the songs. While typical NMF-based algorithms often face challenges due to the significant influence of initial parameter values, Gamma-TCNMF exhibits robustness against parameter initialization, yielding superior results, as demonstrated in Table 4. This robustness is a desirable property, particularly in practical applications.

Table 3: Average SDR, SIR, and SAR values for realistic data (test data).

Method	Input SDR [dB]	Output SDR [dB]	SDR imp. [dB]	Input SIR [dB]	Output SIR [dB]	SIR imp. [dB]	Output SAR [dB]
IVA	26.82	-41.38	-68.20	30.34	-7.02	-37.36	-28.99
ILRMA (L=10)		6.67	-20.15		22.41	-7.92	7.97
ILRMA (L=30)		7.37	-19.46		23.64	-6.70	8.43
ILRMA (L=80)		7.47	-19.35		23.83	-6.51	8.44
DMNMF (L=10)		7.91	-18.92		30.70	0.36	7.98
DMNMF (L=30)		8.67	-18.15		31.06	0.72	8.74
DMNMF (L=80)		9.18	-17.64		31.21	0.87	9.23
$L_{0.5}$-TCNMF		24.93	-1.90		33.11	2.77	26.56
Gamma-TCNMF (proposed)		**28.59**	**1.76**		**35.75**	**5.41**	**30.16**

Table 4: Average and SD values [dB] of SDR improvements for realistic data (test data) over 100 parameter initializations for each song.

Method / Music no.	Conventional TCNMF		Proposed TCNMF	
	Average	SD	Average	SD
4	-0.15	0.23	**1.23**	$\mathbf{5.33\times10^{-7}}$
5	-0.95	0.51	**2.33**	$\mathbf{1.83\times10^{-6}}$
19	-1.93	1.16	**1.61**	$\mathbf{1.42\times10^{-4}}$
20	-0.18	0.61	**2.25**	$\mathbf{7.66\times10^{-6}}$
34	-3.19	0.78	**0.52**	$\mathbf{6.82\times10^{-6}}$
70	-0.32	0.49	**2.34**	$\mathbf{3.57\times10^{-5}}$
71	-0.19	0.81	**2.98**	$\mathbf{8.71\times10^{-7}}$
77	-5.02	1.36	**1.97**	$\mathbf{1.09\times10^{-3}}$
79	-2.52	0.84	**1.69**	$\mathbf{1.22\times10^{-6}}$
99	-4.15	0.56	**0.72**	$\mathbf{2.94\times10^{-3}}$

5 Initialization Robustness Analysis

In Section 4, we demonstrated that Gamma-TCNMF has robustness against initialization of the optimization parameters. In this section, we further investigate the factors behind this robustness through theoretical and experimental analysis.

5.1 Theoretical Analysis of Convexity of Cost Function

The cost function of TCNMF is based on KL divergence (18). It is well known that KLNMF becomes a convex optimization problem when either two variables, $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$, is fixed [2]. TCNMF typically does not become convex optimization due to the simultaneous optimization of both variables, $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$. Moreover, $L_{0.5}$-TCNMF (19) includes a non-convex regularization term represented by the $L_{0.5}$ norm, making it non-convex even if one variable is fixed. On the other hand, the gamma-distribution-based regularization term (36) employed in Gamma-TCNMF is convex with respect to parameter $a > 0$:

$$\begin{aligned}\frac{d^2\mathcal{R}(a)}{da^2} &= \frac{d^2}{da^2}\left[-(k-1)\log a + \frac{a}{\theta}\right] \\ &= (k-1)a^{-2} \\ &\geq 0 \quad (\forall k > 1).\end{aligned} \tag{54}$$

Since the sum of convex functions retains convexity, Gamma-TCNMF becomes convex optimization when one of the two variables is fixed. Although Gamma-TCNMF simultaneously estimates $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$, the diagonal elements of $\boldsymbol{A}_i$ are constrained to unity by the Dirac's delta distribution prior. This constraint may transforms the problem (35) into convex optimization, namely, if KLNMF with fixed diagonal elements of $\boldsymbol{A}_i$ is a convex optimization problem, Gamma-TCNMF that consists of KL divergence and the convex regularization term also becomes convex. On the basis of this hypothesis, we conduct a theoretical analysis to ascertain the convexity of the cost function employed in Gamma-TCNMF.

To check the convexity of KL divergence with fixed diagonal elements, we calculate the Hessian matrix. We consider the simplest case of KLNMF ($I = J = 1$ and $M = 2$) as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \approx \begin{bmatrix} 1 & a_1 \\ a_2 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}, \tag{55}$$

where x_1 and x_2 are nonnegative data and a_1, a_2, s_1, and s_2 are nonnegative optimization parameters. KL divergence between left- and right-hand sides becomes

$$\mathcal{J} = -x_1 \log(s_1 + a_1 s_2) - x_2 \log(a_2 s_1 + s_2) + (1 + a_2)s_1 + (1 + a_1)s_2. \tag{56}$$

The second derivatives of $\mathcal{J}$ are as follows:

$$\frac{\partial^2 \mathcal{J}}{\partial a_1^2} = \frac{x_1 s_2^2}{(s_1 + a_1 s_2)^2}, \tag{57}$$

$$\frac{\partial^2 \mathcal{J}}{\partial a_1 \partial a_2} = 0, \tag{58}$$

$$\frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_1} = \frac{x_1 s_2}{(s_1 + a_1 s_2)^2}, \tag{59}$$

$$\frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_2} = \frac{-x_1(s_1 + 2a_1 s_2)}{(s_1 + a_1 s_2)^2} + 1, \tag{60}$$

$$\frac{\partial^2 \mathcal{J}}{\partial a_2^2} = \frac{x_2 s_1^2}{(a_2 s_1 + s_2)^2}, \tag{61}$$

$$\frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_1} = \frac{-x_2(s_2 + 2a_2 s_1)}{(a_2 s_1 + s_2)^2} + 1, \tag{62}$$

$$\frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_2} = \frac{x_2 a_2}{(a_2 s_1 + s_2)^2}, \tag{63}$$

$$\frac{\partial^2 \mathcal{J}}{\partial s_1^2} = \frac{x_1}{(s_1 + a_1 s_2)^2} + \frac{x_2 a_2^2}{(a_2 s_1 + s_2)^2}, \tag{64}$$

$$\frac{\partial^2 \mathcal{J}}{\partial s_1 \partial s_2} = \frac{x_1 a_1}{(s_1 + a_1 s_2)^2} + \frac{x_2 a_2}{(a_2 s_1 + s_2)^2}, \tag{65}$$

$$\frac{\partial^2 \mathcal{J}}{\partial s_2^2} = \frac{x_1 a_1^2}{(s_1 + a_1 s_2)^2} + \frac{x_2}{(a_2 s_1 + s_2)^2}. \tag{66}$$

The Hessian matrix $\boldsymbol{H}$ can be obtained as

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 \mathcal{J}}{\partial a_1^2} & \frac{\partial^2 \mathcal{J}}{\partial a_1 \partial a_2} & \frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_1} & \frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_2} \\ \frac{\partial^2 \mathcal{J}}{\partial a_1 \partial a_2} & \frac{\partial^2 \mathcal{J}}{\partial a_2^2} & \frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_1} & \frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_2} \\ \frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_1} & \frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_1} & \frac{\partial^2 \mathcal{J}}{\partial s_1^2} & \frac{\partial^2 \mathcal{J}}{\partial s_1 \partial s_2} \\ \frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_2} & \frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_2} & \frac{\partial^2 \mathcal{J}}{\partial s_1 \partial s_2} & \frac{\partial^2 \mathcal{J}}{\partial s_2^2} \end{bmatrix}. \tag{67}$$

Let $\boldsymbol{z}$ be an arbitrary vector as $\boldsymbol{z} = [z_1,\ z_2,\ z_3,\ z_4]^{\mathrm{T}} \in \mathbb{R}^4$. To check the positive-semidefiniteness of $\boldsymbol{H}$, we obtain $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{z}$ as follows:

$$\begin{aligned} \boldsymbol{z}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{z} &= \frac{\partial^2 \mathcal{J}}{\partial a_1^2} z_1^2 + \frac{\partial^2 \mathcal{J}}{\partial a_2^2} z_2^2 + \frac{\partial^2 \mathcal{J}}{\partial s_1^2} z_3^2 + \frac{\partial^2 \mathcal{J}}{\partial s_2^2} z_4^2 + 2\left(\frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_1} z_1 z_3 \right. \\ &\quad \left. + \frac{\partial^2 \mathcal{J}}{\partial a_1 \partial s_2} z_1 z_4 + \frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_1} z_2 z_3 + \frac{\partial^2 \mathcal{J}}{\partial a_2 \partial s_2} z_2 z_4 + \frac{\partial^2 \mathcal{J}}{\partial s_1 \partial s_2} z_3 z_4 \right) \\ &= p_1(z_1 + q_1)^2 + p_2(z_2 + q_2)^2 + (p_3 - p_4)(z_3 + q_3)^2 \\ &\quad + (p_5 - p_6)(z_4 + q_4)^2 + (p_7 - p_8), \end{aligned} \tag{68}$$

where, $p_1, p_2, \cdots, p_8$ and $q_1, q_2, \cdots, q_8$ are positive and real constants, respectively. From (68), $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{z}$ can take either positive and negative values depending on the third, fourth, and fifth terms in (68). This fact shows that the Hessian matrix $\boldsymbol{H}$ is not always a positive semidefinite matrix. Thus, KL-NMF and Gamma-TCNMF are not always a convex optimization problem even if the diagonal elements of $\boldsymbol{A}_i$ are fixed to unity for all i.

5.2 Experimental Analysis of Initialization Robustness and Value Range of Diagonal Elements

As described in Section 5.1, Gamma-TCNMF is not a convex optimization problem similar to simple KLNMF with simultaneous optimization of $\boldsymbol{A}_i$ and $\boldsymbol{S}_i$. Nevertheless, the experimental results in Table 4 show obvious robustness against the parameter initialization. A key disparity between simple KLNMF and the proposed approach lies in the imposition of constraints ensuring the unity of diagonal elements within $\boldsymbol{A}_i$. On the basis of these facts, in this subsection, we conduct and experimental analysis to elucidate the relationship between initialization robustness and the constraint imposed on the diagonal elements in $\boldsymbol{A}_i$.

In this experiment, the range of the diagonal elements in $\boldsymbol{A}_i$ is parametrically adjusted by introducing the gamma distribution prior with a shape parameter $k_d > 1$ and scale parameter $\theta_d = 1/(k_d - 1)$. The probability density function of this prior distribution is depicted in Figure 16, where the mode, denoted as $(k_d - 1)\theta_d$, is fixed at unity. The variance of the gamma distribution can be controlled by the shape parameter k_d. When $k_d \to \infty$, this method coincides with Gamma-TCNMF, i.e., the diagonal elements of $\boldsymbol{A}_i$ are constrained to unity. Conversely, as the shape parameter approaches unity, the gamma distribution converges towards a uniform distribution, thereby manifesting a noninformative prior for the diagonal elements of $\boldsymbol{A}_i$. The update rule of $\boldsymbol{A}_i$ with the above-mentioned prior can be derived as

$$a_{imn} \leftarrow \begin{cases} \frac{(k_d - 1) + a_{imn} \sum_j \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} s_{inj}}{(k_d - 1) + \sum_j s_{inj}} & (\forall m = n) \\ \frac{(k - 1) + a_{imn} \sum_j \frac{x_{imj}}{\sum_{n'} a_{imn'} s_{in'j}} s_{inj}}{\frac{1}{\theta} + \sum_j s_{inj}} & (\forall m \neq n) \end{cases}. \tag{69}$$

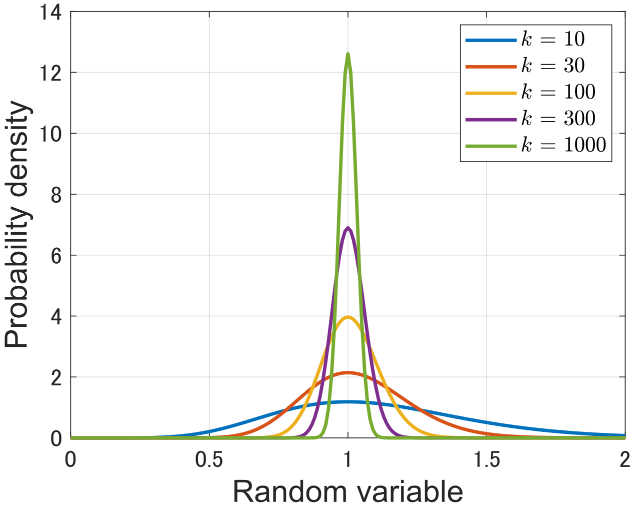


Figure 16: Gamma distribution prior for diagonal elements of gain matrix $\boldsymbol{A}_i$ whose mode is fixed at unity.

The shape parameter k_d was varied across 11 discrete values spanning from 1.004 to 10^6. For the observed signals, we used song no. 70 selected from the test dataset in Section 4.3.3. The other experimental conditions were the same as those in Section 4.3.1. Also, the other hyperparameters k, θ, and α were determined as the same values in Section 4.3.3, which are the optimal values obtained by the development dataset.

Figure 17 presents the violin plot of the SDR improvement corresponding to each value of k_d, while Table 5 provides the average and SD values. From these results, it is clearly confirmed that the distribution of results converges towards optimal performance with increasing values of k_d. Thus, we deduce that the robustness against parameter initializations observed in Gamma-TCNMFS stems from the constraint imposed on diagonal elements, facilitated by the Dirac's delta function (28), despite the non-convex property of the optimization problem. This observation may imply the uniqueness of the solution, warranting further investigation in future studies.

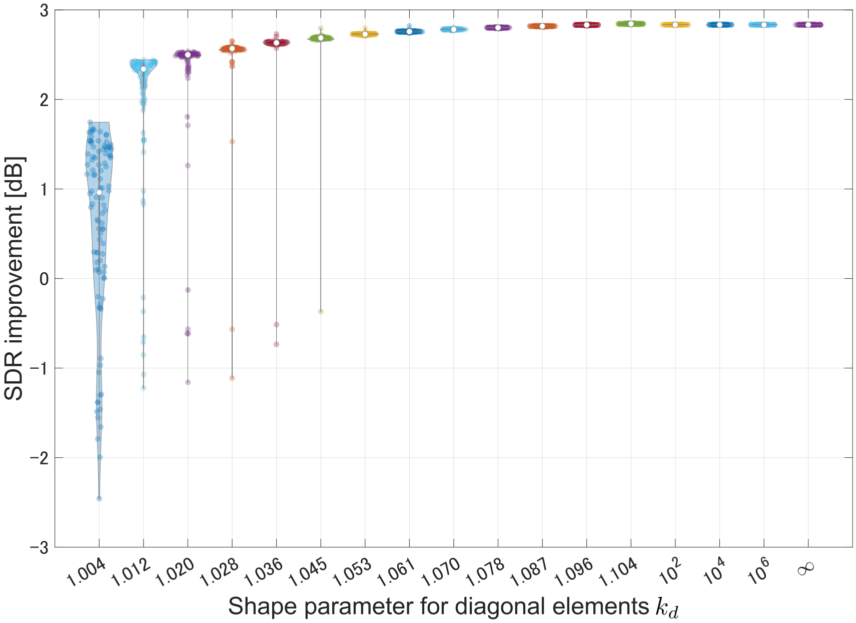


Figure 17: Violin plots of SDR improvements for song no. 70 with various k_d and 100 initialization patterns.

Table 5: Average and SD values [dB] of SDR improvements for song no. 70 with various k_d over 100 initialization patterns.

k_d	Average [dB]	SD [dB]
1.004	0.61	1.025
1.012	2.01	8.241×10^{-1}
1.020	2.30	6.964×10^{-1}
1.028	2.48	4.912×10^{-1}
1.036	2.57	4.590×10^{-1}
1.045	2.66	3.061×10^{-1}
1.053	2.73	9.995×10^{-3}
1.061	2.76	8.157×10^{-3}
1.070	2.78	4.741×10^{-3}
1.078	2.80	3.478×10^{-3}
1.087	2.82	3.012×10^{-3}
1.096	2.83	3.160×10^{-3}
1.104	2.84	3.760×10^{-3}
10^2	2.83	1.547×10^{-5}
10^4	2.83	1.517×10^{-5}
10^6	2.83	1.516×10^{-5}
∞	2.83	1.516×10^{-5}

6 Conclusion

We aimed to reduce the bleeding sound in the observed signal obtained with close microphones. We proposed a new 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 and realistic mixture signals demonstrated that the proposed method could achieve the highest bleeding-sound-reduction performance. In addition, we confirmed that the proposed method is robust to the parameter initializations, which is a desirable property in practical applications. We also revealed that this robustness stems from the constraint imposed on diagonal elements of the mixing matrix.

Since the proposed method has three hyperparameters, an efficient parameter-tuning method is necessary and is for future work. Further theoretical investigation is also required regarding the uniqueness of the solution in the proposed method.

References

[1] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks", *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 24(9), 2016, 1652–64.

[2] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multiway Data Analysis and Blind Source Separation*, UK: John Wiley and Sons, 2009.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ICA using multivariate probability density functions", in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation*, 2006, 601–8.

[4] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign", in *Proc. LVA/ICA*, 2017, 323–32.

[5] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models", *Comput. Intell. Neurosci.*, 2009(785152), 2009.

[6] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", in *Proc. Neural Info. Process. Syst.* 2000, 556–62.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, 401(6755), 1999, 788–91.

[8] D. Kitamura, "Open dataset: songKitamura", http://d-kitamura.net/dataset_en.html, Accessed 30 March 2025.

[9] 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.*, 23(4), 2015, 654–69.

[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.*, 24(9), 2016, 1626–41.

[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*, ed. S. Makino, Cham: Springer, 2018, 125–55.

[12] E. Manilow, G. Wichern, P. Seetharaman, and J. L. Roux, "Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity", in *Proc. IEEE Worksh. Appl. Signal Process. Audio Acoust.* IEEE, 2019.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation", *IEEE Trans. Audio, Speech, and Lang. Process.*, 14(4), 2006, 1462–9.

[14] G.-Y. Chen and C.-N. Wang, “Determined blind source separation combining independent low-rank matrix analysis with optimized parameters and Q-learning”, *Circuits Syst. Signal Process.*, 42, 2023, 6854–70.

[15] 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”, in *Proc. Int. Workshop Acoustic Signal Enhancement*, 2014, 203–7.

[16] H. Kameoka, L. Li, S. Inoue, and S. Makino, “Supervised determined source separation with multichannel variational autoencoder”, *Neural Comput.*, 31(9), 2019, 1891–914.

[17] 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.*, 14(2), 2006, 666–78.

[18] 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.*, 8(e12), 2019, 1–14.

[19] K. Yatabe and D. Kitamura, “Determined BSS based on time-frequency masking and its application to harmonic vector analysis”, *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, 29, 2021, 1609–25.

[20] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag Berlin Heidelberg, 2001.

[21] M. Togami, Y. Kawaguch, H. Kokubo, and Y. Obuchi, “Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization”, in *Proc. Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf.* 2010, 522–5.

[22] 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.*, 27(10), 2019, 1601–15.

[23] 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.*, 178, 2021, 107753.

[24] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals”, *Neurocomputing*, 41(1–4), 2001, 1–24.

[25] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique”, in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.* 2011, 189–92.

[26] O. Das, J. O. Smith, and J. S. Abel, "Microphone cross-talk cancellation in ensemble recordings with maximum likelihood estimation", in *Proc. Audio Eng. Soc. Convention*, 2021.

[27] O. Ylmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", *IEEE Trans. Signal Process.*, 52(7), 2004, 1830–47.

[28] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain", *Neurocomputing*, 22, 1998, 21–34.

[29] R. Rajesh and P. Rajan, "Neural networks for interference reduction in multi-track recordings", in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.* 2023, 1–5.

[30] S. Robin and N. Ono, "Fast and stable blind source separation with rank-1 updates", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* 2020, 236–40.

[31] 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.*, 15(1), 2007, 70–9.

[32] 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.*, 29, 2021, 1687–701.

[33] T. Taniguchi and T. Masuda, "Linear demixed domain multichannel nonnegative matrix factorization for speech enhancement", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* 2017, 476–80.

[34] H. L. V. Trees, *Optimum Array Processing*, New York: John Wiley and Sons, 2002.

[35] X. Yu, D. Hu, and J. Xu, *Blind Source Separation: Theory and Applications*, New York: John Wiley and Sons, 2014.

[36] Y. Mizobuchi, D. Kitamura, T. Nakamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Prior distribution design for music bleeding-sound reduction based on nonnegative matrix factorization", in *Proc. Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf.* 2021, 651–8.

[37] 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", in *Proc. Asia-Pacific Signal Info. Process. Assoc. Annu. Summit Conf.* 2014.

[38] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning", *IEEE Trans. Signal Process.*, 65(3), 2017.