Demixing Filter Estimation for Bleeding-Sound Reduction of a Vocal Microphone

Soushi Taninomiya*, Daichi Kitamura*§, Norihiro Takamune†, Kouei Yamaoka†, Hiroshi Saruwatari†,
Yu Takahashi‡, Kazunobu Kondo‡, and Hayato Yamakawa‡
* National Institute of Technology, Kagawa College, Japan
† The University of Tokyo, Japan
‡ Yamaha Corporation, Japan

***Abstract*—In live music performances, multiple microphones are used to capture sound from individual sources for a sound reinforcement (SR) process. However, the observed signals often contain unwanted leakage from non-target sources, known as bleeding sounds. Bleeding sounds in the vocal (Vo.) microphone signal can particularly degrade SR quality. In this work, to reduce bleeding sounds specifically in the Vo. microphone signal, we propose a semi-blind extension of independent low-rank matrix analysis, in which the observed signals from microphones other than the Vo. microphone are treated as reference signals. Experiments using impulse responses measured in an actual live music venue demonstrate that the proposed method can robustly estimate a demixing filter for bleeding sounds, even under conditions with spatial aliasing.**

I. Introduction

In live music performances, microphones are typically placed close to each sound source to capture and amplify the sound for the audience, which is known as sound reinforcement (SR). This close-miking technique aims to isolate the target source with high clarity. In particular, achieving high isolation for the vocal (Vo.) source is crucial for both mixing and SR processes. However, as illustrated in Fig. 1, many other sound sources—such as instrument amplifiers, monitor loudspeakers, and front-of-house (FoH) loudspeakers—are also present on the stage. As a result, other sounds from nearby sources often leak into the Vo. microphone, leading to unintended signal mixing. This phenomenon, known as "bleeding sound," can adversely affect the SR process.

Multichannel audio source separation (MASS) techniques have the potential to reduce the bleeding sound. In particular, methods based on nonnegative matrix factorization (NMF) [1], which are called time-channel NMF (TCNMF) [2]–[5], have been investigated. TCNMF estimates frequency-wise mixing matrices in the amplitude domain and applies a time-frequency (TF) mask to the observed signal to suppress bleeding sounds. However, such TF masks often introduce artificial distortion and degrade the sound quality of the output signal.

Blind source separation (BSS) based on the statistical independence between sources [6] is also a reliable approach for solving the MASS problem. In this paper, we focus on a method called independent low-rank matrix analysis (ILRMA) [7], [8], as it is well suited to the music-related MASS problem due to its low-rank time-frequency (TF) modeling. ILRMA estimates frequency-wise demixing matrices in the complex domain (including both amplitude and phase), resulting in a linear time-invariant demixing process. This property helps preserve sound quality and makes ILRMA suitable for reducing bleeding sounds in music signals. However, directly applying ILRMA to the MASS problem in live music performances often fails due to severe spatial aliasing. When microphones are spaced far apart (e.g., more than 2 m), as shown in Fig. 1, phase differences across multiple microphones cannot be accurately captured, leading to poor estimation of the demixing matrices [4].

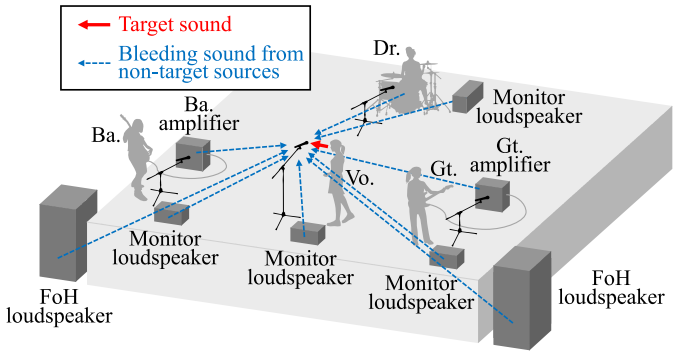


Fig. 1: Spatial arrangement of close microphones and sound sources. Solid and dashed arrows indicate the target and bleeding sounds for the Vo. microphone, respectively.

In actual live music performance setups, bleeding sounds captured by the Vo. microphone pose a serious problem. In contrast, bleeding sounds entering other microphones are typically of relatively low energy and therefore have limited impact, which will be verified in our recording experiment. On the basis of this condition, we focus on bleeding-sound reduction specifically for the Vo. microphone. In the proposed method, we assume that the signals obtained from the other microphones can serve as reference signals for each source, and we introduce a semi-blind demixing model into ILRMA. This approach is interpreted as an estimation of a linear time-invariant demixing filter, which enables robust reduction of bleeding sounds in the Vo. microphone signal. We conducted an experiment using impulse responses measured in an actual

§Corresponding author: kitamura-d@t.kagawa-nct.ac.jp

live music venue to validate the effectiveness of the proposed method.

II. Conventional BSS

A. Formulation

Let N and M be the number of sources and microphones, respectively. The source, observed, and estimated signals at each time-frequency slot calculated via short-time Fourier transform (STFT) are respectively defined as

$$\boldsymbol{s}_{ft} = [s_{ft1}, \cdots, s_{ftn}, \cdots, s_{ftN}]^{\mathrm{T}} \in \mathbb{C}^N, \tag{1}$$

$$\boldsymbol{x}_{ft} = [x_{ft1}, \cdots, x_{ftm}, \cdots, x_{ftM}]^{\mathrm{T}} \in \mathbb{C}^M, \tag{2}$$

$$\boldsymbol{y}_{ft} = [y_{ft1}, \cdots, y_{ftn}, \cdots, y_{ftN}]^{\mathrm{T}} \in \mathbb{C}^N, \tag{3}$$

where $f \in \{1, 2, \cdots, F\}$, $t \in \{1, 2, \cdots, T\}$, $n \in \{1, 2, \cdots, N\}$, and $m \in \{1, 2, \cdots, M\}$ are the indices for frequency bins, time frames, sources, and microphones, respectively, and $\cdot^{\mathrm{T}}$ denotes the transpose.

In BSS, the observed signal is assumed to obey the following mixing model:

$$\boldsymbol{x}_{ft} = \boldsymbol{A}_f \boldsymbol{s}_{ft}, \tag{4}$$

where $\boldsymbol{A}_f \in \mathbb{C}^{M \times N}$ is a frequency-wise time-invariant mixing matrix. In the determined observation case, i.e., $M = N$, BSS can be performed by estimating the demixing matrix

$$\boldsymbol{W}_f = \begin{bmatrix} w_{f11}^* & \cdots & w_{f1M}^* \\ \vdots & \ddots & \vdots \\ w_{fN1}^* & \cdots & w_{fNM}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{f1}^{\mathrm{H}} \\ \vdots \\ \boldsymbol{w}_{fN}^{\mathrm{H}} \end{bmatrix} \in \mathbb{C}^{N \times M}, \tag{5}$$

where $\cdot^*$ and $\cdot^{\mathrm{H}}$ denote the complex conjugate and Hermitian transpose, respectively. Hereafter, we only consider the determined situation $M = N$. In the context of BSS, for overdetermined situation $M > N$, principal component analysis is often applied to $\boldsymbol{x}_{ft}$ for dimensionality reduction so that $M = N$.

If the demixing matrix satisfies $\boldsymbol{W}_f = \boldsymbol{A}_f^{-1}$, the estimated (separated) signal can be obtained as

$$\boldsymbol{y}_{ft} = \boldsymbol{W}_f \boldsymbol{x}_{ft}. \tag{6}$$

Note that $\boldsymbol{w}_{fn}$ can be interpreted as a linear time-invariant demixing filter for the nth source, and the estimated signal can be computed as the inner product between $\boldsymbol{w}_{fn}$ and $\boldsymbol{x}_{ft}$, given by

$$y_{ftn} = \boldsymbol{w}_{fn}^{\mathrm{H}} \boldsymbol{x}_{ft}. \tag{7}$$

B. ILRMA

ILRMA [7], [8] is a powerful approach for accurately estimating the demixing matrix $\boldsymbol{W}_f$. This method simultaneously optimizes $\boldsymbol{W}_f$ and sourcewise NMF variables by maximizing the statistical independence between sources and modeling the power spectrogram of each estimated signal. Since the NMF-based low-rank TF modeling is effective for music sources, ILRMA can achieve high-quality BSS, particularly for music mixtures.

The cost function in ILRMA is defined as

$$\mathcal{J} = -T \sum_f \log |\det \boldsymbol{W}_f|^2 + \sum_{f,t,n} \left[\frac{|\boldsymbol{w}_{fn}^{\mathrm{H}} \boldsymbol{x}_{ft}|^2}{\sum_k b_{fkn} v_{ktn}} + \log \sum_k b_{fkn} v_{ktn} \right], \tag{8}$$

where b_{fkn} and v_{ktn} are the nonnegative elements of basis and activation matrices $\boldsymbol{B}_n \in \mathbb{R}_+^{F \times K}$ and $\boldsymbol{V}_n \in \mathbb{R}_+^{K \times T}$ in NMF, respectively, and $k \in \{1, 2, \cdots, K\}$ is the index of basis vectors (i.e., the columns of $\boldsymbol{B}_n$). The rank-K matrix $\boldsymbol{B}_n \boldsymbol{V}_n$ represents the power spectrogram model for the nth estimated source and plays a key role in promoting the accurate estimation of $\boldsymbol{W}_f$.

The variables $\boldsymbol{W}_f$, $\boldsymbol{B}_n$, and $\boldsymbol{V}_n$ are optimized by minimizing the cost function (8). A fast and stable update rule for $\boldsymbol{W}_f$, called iterative projection (IP) [9], and the well-known multiplicative update rules for $\boldsymbol{B}_n$ and $\boldsymbol{V}_n$ [10] are integrated in ILRMA [7], resulting in a convergence-guaranteed optimization algorithm:

$$\boldsymbol{U}_{fn} = \frac{1}{T} \sum_t \frac{1}{\sum_k b_{fkn} v_{ktn}} \boldsymbol{x}_{ft} \boldsymbol{x}_{ft}^{\mathrm{H}}, \tag{9}$$

$$\boldsymbol{w}_{fn} \leftarrow (\boldsymbol{W}_f \boldsymbol{U}_{fn})^{-1} \boldsymbol{e}_n, \tag{10}$$

$$\boldsymbol{w}_{fn} \leftarrow \frac{\boldsymbol{w}_{fn}}{\sqrt{\boldsymbol{w}_{fn}^{\mathrm{H}} \boldsymbol{U}_{fn} \boldsymbol{w}_{fn}}}, \tag{11}$$

$$b_{fkn} \leftarrow b_{fkn} \sqrt{\frac{\sum_t |\boldsymbol{w}_{fn}^{\mathrm{H}} \boldsymbol{x}_{ft}|^2 v_{ktn} \left(\sum_{k'} b_{fk'n} v_{k'tn}\right)^{-2}}{\sum_t v_{ktn} \left(\sum_{k'} b_{fk'n} v_{k'tn}\right)^{-1}}}, \tag{12}$$

$$v_{ktn} \leftarrow v_{ktn} \sqrt{\frac{\sum_f |\boldsymbol{w}_{fn}^{\mathrm{H}} \boldsymbol{x}_{ft}|^2 b_{fkn} \left(\sum_{k'} b_{fk'n} v_{k'tn}\right)^{-2}}{\sum_f b_{fkn} \left(\sum_{k'} b_{fk'n} v_{k'tn}\right)^{-1}}}, \tag{13}$$

where $\boldsymbol{e}_n \in \{0, 1\}^N$ is a one-hot vector whose nth element is unity. All the variables must randomly be initialized before the iteration. After convergence, the estimated signals are obtained by (6).

III. Proposed Method

A. Motivations

In actual live music performance setups, bleeding sounds captured by the Vo. microphone pose a more serious problem than those captured by other microphones. In particular, drums (Dr.) bleeding into the Vo. microphone is a major issue. As illustrated in Fig. 2, bleeding sounds from the Dr. source tend to be clearly audible due to three main factors: the inherently loud sound of the drums, the typical positions of the Vo. microphone and the Dr. source, and the high head amplifier (HA) gain settings for the Vo. microphone. This contamination degrades the SR quality, even though the Vo. plays the most central role in musical performances. In contrast, bleeding sounds entering other microphones, such as those for guitar (Gt.) amplifiers, bass (Ba.) amplifiers, and Dr., are typically of

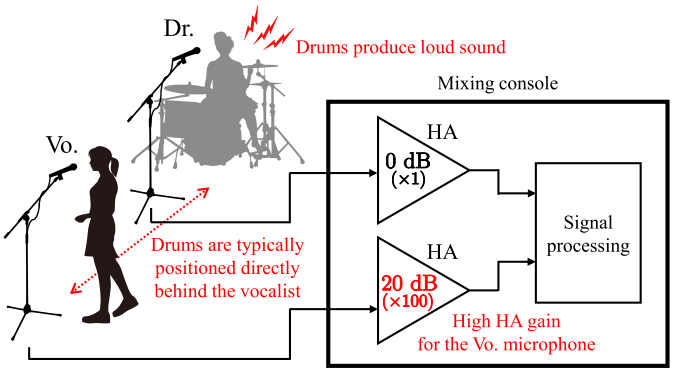


Fig. 2: Typical stage setup in live music performances. The drum set is usually positioned directly behind the Vo. microphone, and drum sounds are inherently loud. Moreover, the HA gain for the Vo. microphone is set high, making the microphone prone to capturing bleeding sounds from the drums.

relatively low energy and therefore have limited impact. This fact will be experimentally verified in Sect. IV-A.

On the basis of the above conditions, we propose a semi-blind source separation approach that specifically focuses on reducing bleeding sounds in the Vo. microphone. The proposed method introduces a semi-blind demixing model into ILRMA. Since this model dramatically reduces the number of parameters to be estimated, it enables robust bleeding-sound reduction even in the presence of spatial aliasing in the observed signals. Furthermore, the proposed method can be interpreted as the estimation of a linear time-invariant demixing filter for the Vo. microphone signal using other reference microphones, which is closely related to echo cancellation algorithms [11], [12]. However, unlike echo cancellation, our method does not require single-talk segments or voice activity detection, as ILRMA estimates the demixing matrix in a fully blind manner.

B. Semi-Blind ILRMA

Let s_{ft1} $(n=1)$ and x_{ft1} $(m=1)$ be the TF components of the Vo. source and the Vo. microphone, respectively, where x_{ft1} includes excessive bleeding sounds from the other sources $s_{ft2},\cdots,s_{ftN}$. The other microphone signals $x_{ft2},\cdots,x_{ftM}$ are assumed to serve as reference channels, i.e., these signals are assumed to contain no bleeding sounds, resulting in $x_{ftm}=s_{ftn}\ \forall m=n\in\{2,\cdots,N\}$. The validity of this assumption depends on the signal-to-noise ratios (SNRs) of each microphone, which will be measured in Sect. IV-A.

On the basis of this assumption, a semi-blind model of the demixing matrix, denoted by $\tilde{\boldsymbol{W}}_f\in\mathbb{C}^{N\times M}$, is given as

$$\tilde{\boldsymbol{W}}_f=\begin{bmatrix}\tilde{w}_{f1}^* & \tilde{w}_{f2}^* & \tilde{w}_{f3}^* & \cdots & \tilde{w}_{fM}^*\\ 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix}=\begin{bmatrix}\tilde{\boldsymbol{w}}_f^{\mathrm{H}}\\ \boldsymbol{e}_2^{\mathrm{H}}\\ \boldsymbol{e}_3^{\mathrm{H}}\\ \vdots\\ \boldsymbol{e}_N^{\mathrm{H}}\end{bmatrix}\quad\forall f. \tag{14}$$

By substituting (14) into the ILRMA cost function (8) and simplifying it using the structure of $\tilde{\boldsymbol{W}}_f$, we obtain the following cost function for semi-blind ILRMA:

$$\mathcal{J}=-T\sum_f\log|\tilde{\boldsymbol{w}}_f^{\mathrm{H}}\boldsymbol{e}_1|^2+\sum_{f,t}\left[\frac{|\tilde{\boldsymbol{w}}_f^{\mathrm{H}}\boldsymbol{x}_{ft}|^2}{\sum_k\tilde{b}_{fk}\tilde{v}_{kt}}+\log\sum_k\tilde{b}_{fk}\tilde{v}_{kt}\right], \tag{15}$$

where $\tilde{b}_{fk}$ and $\tilde{v}_{kt}$ are the nonnegative elements of basis and activation matrices $\tilde{\boldsymbol{B}}\in\mathbb{R}_+^{F\times K}$ and $\tilde{\boldsymbol{V}}\in\mathbb{R}_+^{K\times T}$, respectively. The model spectrogram $\tilde{\boldsymbol{B}}\tilde{\boldsymbol{V}}$ represents the low-rank TF structure of the Vo. source.

In this method, only the first row in $\tilde{\boldsymbol{W}}_f$, denoted as $\tilde{\boldsymbol{w}}_f$, is a spatial variable to be estimated. This vector corresponds to the demixing filter used for reducing bleeding sounds in the Vo. microphone signal. The low-rank TF model $\tilde{\boldsymbol{B}}\tilde{\boldsymbol{V}}$ captures only the power spectrogram of the Vo. source, and the TF models for the other sources are not estimated. As a result, the number of parameters in ILRMA is reduced from $\boldsymbol{W}_f,\boldsymbol{B}_n,\boldsymbol{V}_n\ \forall f,n$ to $\tilde{\boldsymbol{w}}_f,\tilde{\boldsymbol{B}},\tilde{\boldsymbol{V}}\ \forall f$, corresponding to a reduction by a factor of N. This reduction enables robust estimation of $\tilde{\boldsymbol{w}}_f$ even under conditions with spatial aliasing.

The convergence-guaranteed update rules for $\tilde{\boldsymbol{w}}_f$, $\tilde{\boldsymbol{B}}$, and $\tilde{\boldsymbol{V}}$ can be derived in the same manner as in ILRMA, as follows:

$$\tilde{\boldsymbol{U}}_f=\frac{1}{T}\sum_t\frac{1}{\sum_k\tilde{b}_{fk}\tilde{v}_{kt}}\boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^{\mathrm{H}}, \tag{16}$$

$$\tilde{\boldsymbol{w}}_f\leftarrow\tilde{\boldsymbol{U}}_f^{-1}\boldsymbol{e}_1, \tag{17}$$

$$\tilde{\boldsymbol{w}}_f\leftarrow\frac{\tilde{\boldsymbol{w}}_f}{\sqrt{\boldsymbol{e}_1^{\mathrm{H}}\tilde{\boldsymbol{w}}_f}}, \tag{18}$$

$$\tilde{b}_{fk}\leftarrow\tilde{b}_{fk}\sqrt{\frac{\sum_t|\tilde{\boldsymbol{w}}_f^{\mathrm{H}}\boldsymbol{x}_{ft}|^2\tilde{v}_{kt}\left(\sum_{k'}\tilde{b}_{fk'}\tilde{v}_{k't}\right)^{-2}}{\sum_t\tilde{v}_{kt}\left(\sum_{k'}\tilde{b}_{fk'}\tilde{v}_{k't}\right)^{-1}}}, \tag{19}$$

$$\tilde{v}_{kt}\leftarrow\tilde{v}_{kt}\sqrt{\frac{\sum_f|\tilde{\boldsymbol{w}}_f^{\mathrm{H}}\boldsymbol{x}_{ft}|^2\tilde{b}_{fk}\left(\sum_{k'}\tilde{b}_{fk'}\tilde{v}_{k't}\right)^{-2}}{\sum_f\tilde{b}_{fk}\left(\sum_{k'}\tilde{b}_{fk'}\tilde{v}_{k't}\right)^{-1}}}. \tag{20}$$

IV. Experiments

A. Impulse Response Measurement in Live Music Venue

To simulate a realistic mixing system on a live music stage, impulse responses were recorded in an actual live music venue. Fig. 3 shows a top-view schematic of the recording environment. A four-source and four-microphone setup was assumed, and impulse responses were measured from each source to each microphone. A professional SR engineer adjusted the HA gain of each source using actual musical instruments. The HA gains were set to 26 dB, 20 dB, 18 dB, and 0 dB for the Vo., Gt., Ba., and Dr. sources, respectively. We prepared two recording cases, as shown in Fig. 4: Case A, which involved measuring impulse responses on the stage without using any loudspeakers, and Case B, which simulated a more realistic situation in which the FoH loudspeakers emitted a real-time mixture of all the source signals with equal gain, and the

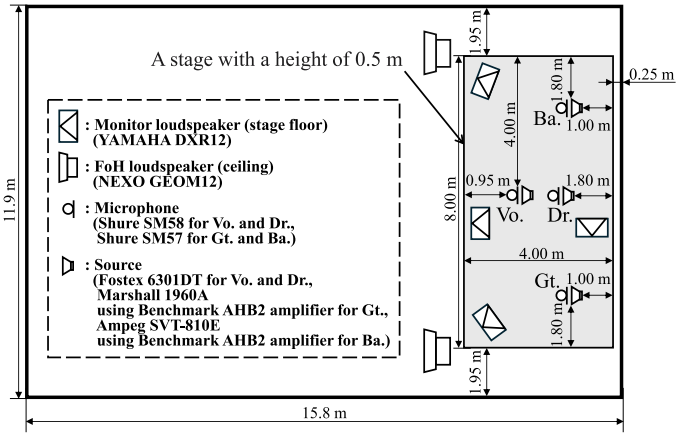
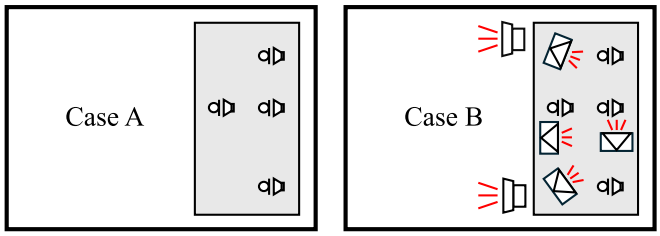


Fig. 3: Top-view schematic of the microphone and loudspeaker arrangement in a live music venue. The gray rectangle represents the stage. Two FoH loudspeakers are suspended from the ceiling, and four monitor loudspeakers are placed on the stage floor. Distances between each source and its corresponding microphone are set to 0.01 m.



(a) Case A (b) Case B

Fig. 4: Two recording conditions for the impulse response measurement: (a) all loudspeakers are muted and (b) all loudspeakers are active. In (b), the FoH loudspeakers emit a real-time mixture of all source signals with equal gain, while the monitor loudspeakers emit a real-time mixture of only the Vo., Gt., and Ba. sources, also with equal gain.

monitor loudspeakers emitted a real-time mixture of only the Vo., Gt., and Ba. sources, also with equal gain. The output gains of FoH and monitor loudspeakers were adjusted by the professional SR engineer. The sounds emitted from the FoH and monitor loudspeakers were also captured as bleeding sounds by each microphone as illustrated in Fig. 1.

The reverberation times, calculated using the impulse responses from the Vo. source to the Vo. microphone, were $T_{60} = 760$ ms and $T_{60} = 775$ ms in Case A and Case B, respectively. Tables I and II show the sourcewise relative energies of the bleeding sounds observed by each close microphone. Since the values are normalized by the energy of the target source observed by each close microphone, the diagonal elements in Tables I and II are zero. From these results, we can confirm that the Vo. microphone suffers from the high-energy bleeding sounds. In particular, the bleeding sound from the Dr. source exceeds 0 dB in the Vo. microphone, namely, the

TABLE I: Observed relative bleeding-sound energy [dB] in Case A

	Vo. source	Gt. source	Ba. source	Dr. source
Vo. microphone	0.0	-17.3	-12.9	1.2
Gt. microphone	-56.3	0.0	-25.3	-29.4
Ba. microphone	-59.7	-34.8	0.0	-31.2
Dr. microphone	-72.7	-44.6	-40.6	0.0

TABLE II: Observed relative bleeding-sound energy [dB] in Case B

	Vo. source	Gt. source	Ba. source	Dr. source
Vo. microphone	0.0	-9.8	-6.9	1.3
Gt. microphone	-31.5	0.0	-16.9	-19.8
Ba. microphone	-32.5	-21.5	0.0	-23.1
Dr. microphone	-43.9	-34.1	-31.4	0.0

Dr. source is louder than the Vo. source itself, even though the microphone is placed in close proximity to the Vo. source. This can be attributed to the factors illustrated in Fig. 2. In contrast, for the Gt., Ba., and Dr. microphones, the bleeding sounds are not as severe. In these microphones, the target source maintains an SNR margin of at least 25 dB and 16 dB in Case A and Case B, respectively. This supports the validity of our assumption that the signals observed by the Gt., Ba., and Dr. microphones can be used as references channels.

B. Performance Analysis with Various Window Lengths in Two-Source Mixture Case

We compared three methods: Gamma-TCNMF [5], simple ILRMA [7], and the proposed semi-blind ILRMA. As the evaluation criterion, we used the source-to-distortion ratio (SDR), calculated using `bss_eval_sources` [13]. SDR reflects the overall separation quality, taking into account both the suppression of bleeding sounds and the absence of artificial distortions. We calculated the SDR of the observed and estimated signals for the Vo. source, and the difference of these values was obtained as an SDR improvement.

As the dry source signals, we used three songs (nos. `022`, `023`, and `040` in the test dataset) randomly selected from the *DSD100* dataset [14]. This dataset consists of full-length music tracks along with their isolated Vo., Ba., Dr., and other signals (Gt. in the case of the three selected songs). For each source, a 20-s segment was extracted from each track and used as the dry source signal.

Since MASS algorithms based on (4) or (6) typically depend on the window length used in the STFT, in this subsection, we evaluate the performance of each method with various window lengths. To clearly observe the performance changes with respect to window length, we simplified the bleeding-sound reduction task, namely, only the Vo. and Dr. sources and microphones were used for producing the two-source and two-channel observed signals. In Case B, all the monitor (and FoH) loudspeakers were active even though only the Vo. and Dr. sources existed on the stage.

The hyperparameters in Gamma-TCNMF were experimentally tuned using the same observed signals to obtain optimal performance. The best values of (κ, θ, α) were found to be

TABLE III: Experimental conditions in two-source mixture case

Parameter	Condition
Window length in STFT	93/186/372/743/1115/ 1486/1858/2229 ms
Window function in STFT	Blackman window
Window shift length in STFT	1/8 of window length
Number of iterations	100
Initial values of off-diagonal elements in mixing matrix for Gamma-TCNMF	Uniform random values in range $(0, 0.2)$
Initial values of activation matrix for Gamma-TCNMF	Uniform random values in range $(0, 1)$
Number of basis vectors for ILRMA and semi-blind ILRMA	$K = 10$
Initial values of demixing matrix for ILRMA and semi-blind ILRMA	Identity matrix
Initial values of basis and activation matrices for ILRMA and semi-blind ILRMA	Uniform random values in range $(0, 1)$

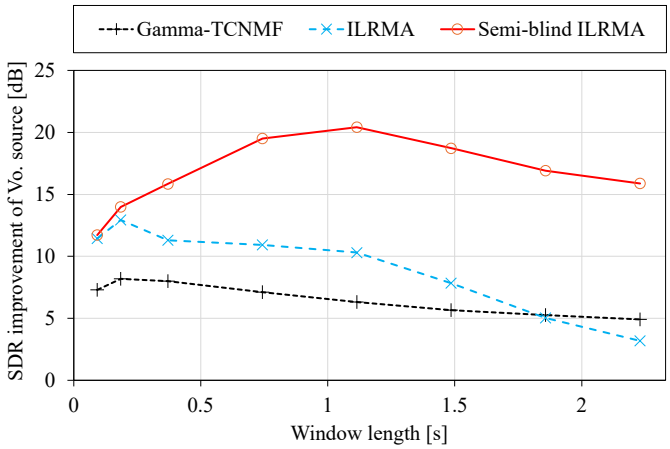
$(1.1, 21.5, 0.00077)$ for Case A and $(2, 1, 0.0001)$ for Case B, where κ and θ are the shape and scale parameters of the gamma distribution, and α is the normalization parameter (see [5]). Other experimental conditions are summarized in Table III. Since all methods require initialization and random values were used as initial values, we ran each method with 30 different pseudorandom seeds and report the average SDR improvement over these 30 trials.

Fig. 5 shows the average SDR improvements of the Vo. source with respect to various window lengths. Simple ILRMA fails to reduce bleeding sounds under longer window conditions, whereas the proposed semi-blind ILRMA consistently achieves significantly better performance. This stable and robust improvement can be attributed to the introduction of the model (14), which reduces the number of parameters to be estimated. Gamma-TCNMF is also effective for bleeding-sound reduction, particularly in Case B, but its optimal performance remains inferior to that of the proposed method.

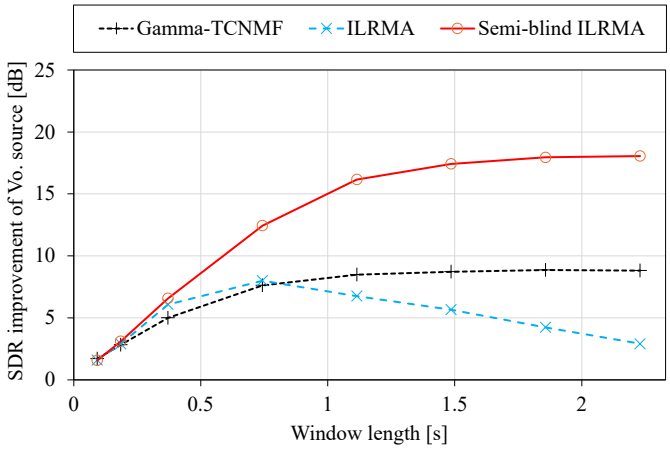
C. Performance Comparison in Four-Source Mixture Case

In this subsection, we compare the performance of each method using the four-source and four-channel observed signals. The window length was set to the optimal value identified in Fig. 5, as summarized in Table IV. All other experimental conditions were the same as those described in Sect. IV-B.

Fig. 6 shows violin plots of SDR improvements, where each violin includes 90 results (30 trials for each of three songs). In Fig. 6 (a), the results for Gamma-TCNMF exhibit no variation with respect to parameter initialization, as previously analyzed in [5], and the three layers visible in the violin plot correspond to the three different songs. However, intriguingly, this initialization robustness is clearly lost in Fig. 6 (b). In Case B, where the FoH and monitor loudspeakers were active, the sounds emitted from these virtual sources appear to disrupt the initialization robustness of Gamma-TCNMF. In both cases, the proposed semi-blind ILRMA clearly outperforms the other methods and achieves significant bleeding-sound reduction. Nonetheless, the performance appears to depend on the choice



(a) Case A



(b) Case B

Fig. 5: Average SDR improvements of the Vo. source in two-source mixture case for (a) Case A and (b) Case B.

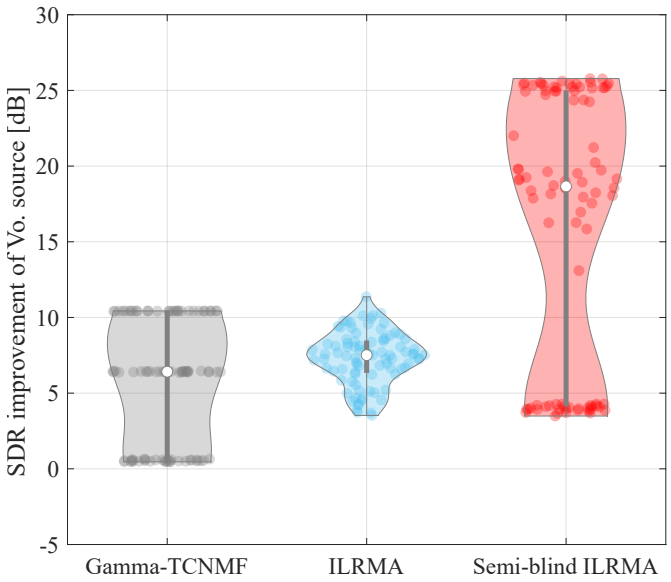
TABLE IV: Best window lengths [ms] for each method

Case	Gamma-TCNMF	ILRMA	Semi-blind ILRMA
Case A	186	186	1115
Case B	1858	743	2229

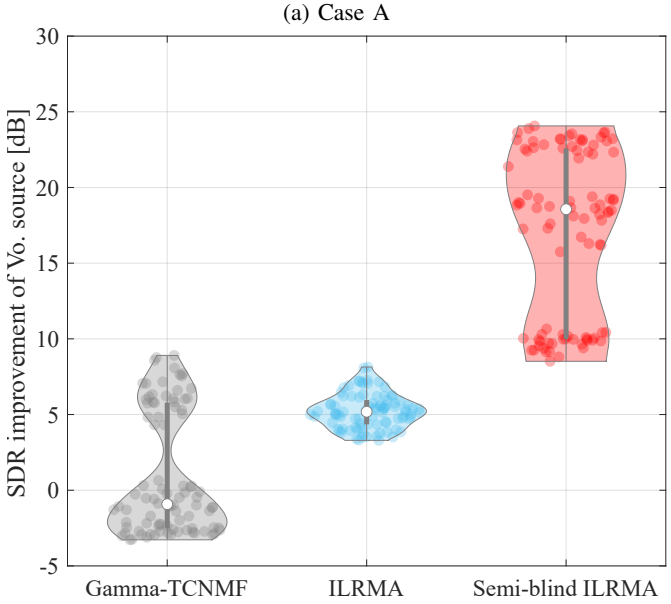
of dry source, indicating that further investigation is required to achieve consistently high performance.

V. Conclusion

In this paper, we proposed an effective bleeding-sound reduction approach for the Vo. microphone. The proposed method exploits the assumption that signals obtained by other microphones on the stage can serve as reference signals for the non-target sources. On the basis of this assumption, we introduced a semi-blind demixing model into ILRMA, a well-established BSS framework. Experiments using impulse responses measured in an actual live music venue demonstrated that the proposed method can accurately estimate a demixing filter for bleeding-sound reduction, even under spatial aliasing conditions.



(a) Case A



(b) Case B

Fig. 6: Violin plot of SDR improvements of the Vo. source in four-source mixture case for (a) Case A and (b) Case B. In each method, a white circle indicates median value, a gray vertical line shows range of 25–75 percentiles, and a violin curve is an estimated distribution.

REFERENCES

[1] 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.

[2] 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.*, pp. 522–525, 2010.

[3] 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*, pp. 203–207, 2014.

[4] 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.*, pp. 651–658, 2021.

[5] Y. Mizobuchi, D. Kitamura, T. Nakamura, N. Takamune, H. Saruwatari, Y. Takahashi, and K. Kondo, "Music bleeding-sound reduction based on time-channel nonnegative matrix factorization," in *APSIPA Trans. Signal Inf. Process.*, 2025 (in press).

[6] 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 Inf. Process.*, vol. 8, no. e12, pp. 1–14, 2019.

[7] 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 Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[8] 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.

[9] 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.*, pp. 189–192, 2011.

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

[11] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer Berlin, Heidelberg, 2001.

[12] R. Cutler, A. Saabas, T. Pärnamaa, M. Purin, E. Indenbom, and N.-C. Ristea, "ICASSP 2023 acoustic echo cancellation challenge," *IEEE Open J. Signal Process.*, vol. 5, pp. 675–685, 2024.

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

[14] 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. Int. Conf. Latent Var. Anal. Signal Separ.*, pp. 323–332, 2017.