Visually-Informed Multichannel Sound Source Separation Based on 3D Gaussian Primitives

Haruaki Asano* Ryunosuke Nihei* Yoshiaki Bando†‡ Aditya Arie Nugraha‡
Diego Di Carlo‡ Hiroyuki Ueda* Yosuke Ito* Kazuyoshi Yoshii*‡

*Graduate School of Engineering, Kyoto University, Japan
†National Institute of Advanced Industrial Science and Technology (AIST), Japan
asano.haruki.88p@st.kyoto-u.ac.jp

***Abstract*—This paper proposes visually-informed sound source separation for audio-visual understanding of indoor scenes captured by distributed microphone arrays and cameras. Our approach leverages the 3D information of sound-emitting objects, reconstructed via 3D Gaussian splatting (3DGS), to overcome a limitation of modern blind source separation methods like multichannel nonnegative matrix factorization (MNMF). While adaptable and potentially performant, the iterative optimization of MNMF often converges to poor local minima due to the highly-expressive full-rank spatial covariance matrices (SCMs) of sources. Our key idea is to treat the set of 3D Gaussians representing a sizable sound source object as a collection of sub-sources that share an audio signal but have unique emission weights, both of which are to be estimated jointly from an observed mixture. To enforce this structure, we guide MNMF by regularizing the SCM of each source object at each frequency. Specifically, we use a prior that centers the SCM estimate around a weighted sum of theoretical SCMs, which are analytically derived from the 3D Gaussian positions. Experiments with simulated data, featuring two 3D human models, demonstrated the effectiveness of the proposed method. To our knowledge, this is the first work to use 3D Gaussians as a common primitive for joint audio-visual analysis.**

I. Introduction

Multichannel sound source separation forms the basis of various downstream tasks such as speaker diarization [1] and speech recognition [2] in comprehensive understanding of indoor acoustic scenes. Whereas supervised methods based on deep learning have been proven to be effective in offline benchmarks [3], blind source separation (BSS) can still be considered to be useful in real environments due to its robustness against the acoustic variations [4].

Multichannel nonnegative matrix factorization (MNMF) is a major family of BSS methods [5], [6]. It is based on a probabilistic model of multichannel mixture signals that consists of a source model representing the power spectral densities (PSDs) of sources and a spatial model representing the spatial covariance matrices (SCMs) of the sources related to the directions of arrival (DOAs). While adaptable and potentially performant, the iterative optimization of MNMF often converges to a poor local minimum due to the full-rankness of source SCMs. To mitigate this problem, spatial information such as the microphone array geometry [7], [8] and source DOAs or locations [9], [10] have been used for regularizing the SCM estimation.

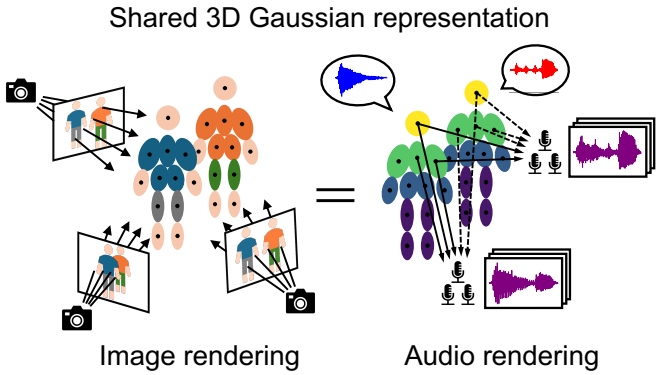

Fig. 1: Audio-visual scene representation based on 3D Gaussian primitives. The 3D shape of each sound source object (speaker) is represented as a set of anisotropically-colored opaque 3D Gaussians that simultaneously emit the same signal with different weights. The images and multichannel signals observed by cameras and microphone arrays at different positions are rendered separately based on the color projection to the image planes and the superimposition of source images, respectively.

When multiple cameras and microphones are distributed in a room as is often the case with realistic scenarios such as smart meeting systems [11], the source SCMs could be effectively regularized based on visual information. In the field of computer vision, novel view synthesis (NVS) techniques such as the neural radiance field (NeRF) [12] or 3D gaussian splatting (3DGS) [13] have been proposed to estimate the 3D shapes of objects from multi-view images. In particular, 3DGS has gained much attention due to the interpretability of the explicit scene representation based on a collection of 3D Gaussians with directional color information. Such NVS techniques have recently been extended for joint audio-visual synthesis at arbitrary positions [14]–[16]. In contrast to the rendering purpose, visual information still remains underutilized for the purpose of sound source separation.

For improved sound source separation, in this paper we propose a visually-informed MNMF with 3DGS-based priors on source SCMs. Our key idea is to treat the set of 3D Gaussians representing a sound source object as a collection of

979-8-3315-7206-8/25/$31.00

sub-sources that emit the same signal with different weights (Fig. 1). The SCM of each source object at each frequency can thus be regularized based on the 3D Gaussian positions. Specifically, we incorporate an inverse-Wishart prior that centers the SCM estimate around a weighted sum of theoretical SCMs, which are analytically derived from the 3D Gaussian positions. Given multichannel mixture signals as observed data. we jointly estimate the MNMF parameters including the PSDs and SCMs of sources and the emission weights of sub-sources in the maximum-a-posteriori (MAP) principle. The source signals are finally estimated via multichannel Wiener filtering.

This is the first work that commonly uses the 3D Gaussian primitives for audio-visual scene understanding. We can effectively integrate 3D information derived from images into spatial models used for joint separation and localization of non-point sources having volumes. This approach reduces initialization sensitivity and improves the stability of MNMF estimation. This work could also contribute to audio-visual understanding for robotics and smart meeting systems [17], [18].

II. Related work

This section reviews related work on BSS, 3DGS, and audio-visual scene analysis.

A. Blind Source Separation

A common approach to BSS is to combine source models representing the time-frequency structures of sources and spatial models representing the inter-channel covariance structures of source images. Frequency-domain ICA [19] and its extensions such as independent vector analysis (IVA) [20], [21] and independent low-rank matrix analysis (ILRMA) [22] are based on rank-1 spatial models in common and have different source models. In contrast, full-rank spatial covariance analysis (FCA) [23] and MNMF [5] use full-rank SCMs, providing richer spatial modeling at the cost of larger sensitivity to local optima and higher computational complexity. FastMNMF [24], [25] assumes the joint diagonalizability of the source SCMs. It still uses a full-rank spatial model while keeping better separation accuracy and accelerating the parameter updates in MNMF.

B. 3D Gaussian Splatting

3DGS [13] is a fast and high-quality NVS technique that can reconstruct a 3D scene as a set of 3D Gaussians from multi-view images. Each Gaussian is parametrized by the mean, covariance, opacity, and directional color for unified modeling of geometry and appearance. In the generative process for volume rendering, each Gaussian is projected onto the image plane. Specifically, a ray is casted from the camera towards the direction of each pixel and Gaussian colors are integrated along the ray, with closer ones attenuating those behind. The color of each pixel is thus determined by light transmission and overlapping contributions along the ray. The parameters are optimized by minimizing the loss between the rendered and observed images from the same viewpoints. Unlike NVS methods that require dense point clouds or voxels, detailed 3D scenes can be reconstructed based on the interpretable compact representation. In general, 3D Gaussians tend to be dense around objects, enabling spatial segmentation to 3D objects. Techniques has been proposed that leverages this property to assign object-wise labels to 3D Gaussians and cluster them [26], [27]. In this study, we thus assume such clustering is precomputed for associating a collection of 3D Gaussians with a sound source object.

C. Audiovisual Integration and Analysis

To enhance spatial understanding by integrating visual and acoustic information, AV-NeRF [14] separately trains a visual model (V-NeRF) and an acoustic model (A-NeRF), where the 3D geometry and material properties estimated by V-NeRF are passed to A-NeRF through audio-visual mapping. This enables the generation of both arbitrary-view images and binaural audio. NeRAF [15] handles cases in which visual and acoustic data are captured separately, extracting voxel-based geometric information from visual input and estimating the room impulse response (RIR) accordingly.

AV-GS [16] extends 3DGS by introducing acoustic parameters representing sound diffusion and absorption. After learning the visual component, the model performs sequential training to incorporate acoustic information, thereby achieving unified audiovisual rendering. All of these approaches demonstrate that jointly modeling audiovisual information improves accuracy and enables more sophisticated scene analysis. In this work, we adopt a similar audiovisual integration framework and apply it to sound source separation.

III. Proposed Method

This section describes the proposed visually-informed source separation method named 3DGS-MNMF.

A. Problem Specification

Suppose that L microphone arrays, each with M channels, are located at known positions and there are N speakers (sound-emitting objects) in a room. Let $\mathbf{r}_{lm} \in \mathbb{R}^3$ be the 3D position of microphone $m \in [1, M]$ in array $l \in [1, L]$. Let $\mathbf{X} \triangleq \{\mathbf{x}_{ft} \in \mathbb{C}^{ML}\}_{f,t=1}^{F,T}$ be the observed multichannel mixture spectrogram over time and frequency, where F and T represent the number of frequency bins and that of time frames, respectively.

Suppose that 3DGS and 3D segmentation (object-wise clustering of 3D Gaussians) are performed in advance using distributed cameras [26]. Let $\mathbf{u}_{ni} \in \mathbb{R}^3$ be the center position of 3D Gaussian $i \in [1, I]$ in speaker n, where I is the number of 3D Gaussians used for representing the 3D shape of speaker n. Let $\mathbf{U} \triangleq \{\mathbf{u}_{ni}\}_{n,i=1}^{N,I}$ be the full set of Gaussian positions.

Our goal is to estimate the sound source *image* $\mathbf{Z}_n \triangleq \{\mathbf{z}_{nft} \in \mathbb{C}^{ML}\}_{f,t=1}^{F,T}$ of speaker n at the microphone positions. For convenience, let $\mathbf{x}_{ftl} \in \mathbb{C}^M$ and $\mathbf{z}_{nftl} \in \mathbb{C}^M$ be the partial observation and image at array l, i.e., $\mathbf{x}_{ft} = [\mathbf{x}_{ft1}^{\mathsf{H}}, \cdots, \mathbf{x}_{ftL}^{\mathsf{H}}]^{\mathsf{H}}$ and $\mathbf{z}_{nft} = [\mathbf{z}_{nft1}^{\mathsf{H}}, \cdots, \mathbf{z}_{nftL}^{\mathsf{H}}]^{\mathsf{H}}$, respectively.

B. Model Formulation

We formulate a probabilistic model of $\mathbf{X}$ with 3DGS-based priors on the SCMs of sources.

1) Likelihood function: The probabilistic model of $\mathbf{X}$ is defined in the same way as MNMF [5]. Assuming the low-rank structure of the PSDs $\{\lambda_{nft}\}_{f,t=1}^{F,T}$ of each source n, we first formulate a source model that represents the generative process of the complex spectrogram $\mathbf{S}_n \triangleq \{s_{nft} \in \mathbb{C}\}_{f,t=1}^{F,T}$ as follows:

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, \lambda_{nft} \triangleq \sum_{k=1}^{K} w_{nkf} h_{nkt}\right), \tag{1}$$

where the PSDs are decomposed by NMF with K bases and $\{w_{nkf}\}_{f=1}^{F}$ and $\{h_{nkt}\}_{t=1}^{T}$ represent the basis and activation of source n and basis k, respectively. We then formulate a spatial model that represents the *image* (multichannel spectrogram) of source n at array l as follows:

$$\mathbf{z}_{nftl} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \mathbf{Y}_{nftl} \triangleq \lambda_{nft}\mathbf{G}_{nfl}\right), \tag{2}$$

where $\mathbf{G}_{nfl} \in \mathbb{C}^{M\times M}$ is the full-rank SCM of source n and array l at frequency f. As in [10], we assume that $\mathbf{z}_{nftl}$ and $\mathbf{z}_{nftl'}(l \neq l')$ are independent reflecting that the distances between arrays l and l' can be measured much less precisely than those between the microphones in each array. Assuming the source additivity, i.e., $\mathbf{x}_{ftl} = \sum_{n=1}^{N} \mathbf{z}_{nftl}$, we have

$$\mathbf{x}_{ftl} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathbf{G}_{nfl}\right). \tag{3}$$

The likelihood of the parameters $\mathbf{W} \triangleq \{w_{nfk}\}_{n,f,k=1}^{N,F,K}$, $\mathbf{H} \triangleq \{h_{nkt}\}_{n,k,t=1}^{N,K,T}$, $\mathbf{G} \triangleq \{\mathbf{G}_{nfl}\}_{n,f,l=1}^{N,F,L}$ to be estimated for the observed mixture $\mathbf{X}$ is given by

$$p(\mathbf{X}|\mathbf{W},\mathbf{H},\mathbf{G}) = \prod_{f,t,l=1}^{F,T,L} \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}_{ftl}\middle|\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft}\mathbf{G}_{nfl}\right). \tag{4}$$

Given $\mathbf{X}$ with $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$, we can estimate the source image $\mathbf{z}_{nftl}$ with a multichannel Wiener filter as follows:

$$\mathbb{E}[\mathbf{z}_{nftl}|\mathbf{x}_{ftl},\mathbf{W},\mathbf{H},\mathbf{G}] = \mathbf{Y}_{nftl}\mathbf{Y}_{ftl}^{-1}\mathbf{x}_{ftl}, \tag{5}$$

where $\mathbf{Y}_{ftl} \triangleq \sum_{n=1}^{N} \mathbf{Y}_{nftl}$.

2) Spatial priors: To regularize the full-rank positive definite matrix $\mathbf{G}_{nfl}$, we put a complex inverse Wishart prior on $\mathbf{G}_{nfl}$ in the same way as [10] as follows:

$$\mathbf{G}_{nfl} \sim \mathcal{IW}_{\mathbb{C}}\left(\nu, (\nu+M)\hat{\mathbf{G}}_{nfl}\right), \tag{6}$$

where $\nu > M-1$ is the degree of freedom and $(\nu+M)\hat{\mathbf{G}}_{nfl} \in \mathbb{C}^{M\times M}$ is a scale matrix such that the mode (the most probable value) of the *actual* SCM $\mathbf{G}_{nfl}$ is the *theoretical* SCM $\hat{\mathbf{G}}_{nfl}$. In most studies on source separation and localization (e.g., [10]), the theoretical SCMs of *point* sources are typically computed from the source and microphone positions. In contrast, we here consider the SCMs of *sizable* sources (objects) represented by 3D Gaussians.

Our key idea is to approximate a sizable source as a collection of a large number of point *sub-sources* (3D Gaussians) that emit the same signal with different weights. As shown in Fig. 2, the theoretical SCM $\hat{\mathbf{G}}_{nfl}$ of source n for frequency f and array l is thus given by a weighted sum of the theoretical SCMs of the I sub-sources as follows:

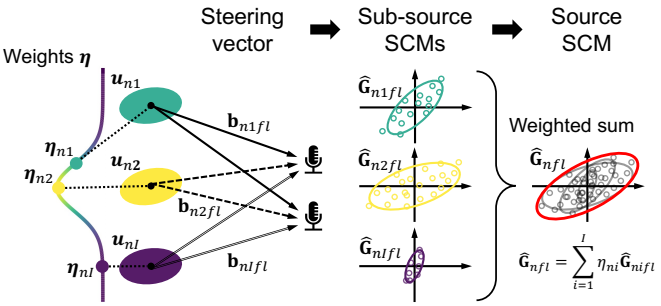

Fig. 2: The theoretical full-rank SCM $\hat{\mathbf{G}}_{nfl}$ of source n given by the weighted sum of the theoretical rank-1 SCMs $\{\hat{\mathbf{G}}_{nifl} = \mathbf{b}_{nifl}\mathbf{b}_{nifl}^{\mathrm{H}}\}_{i=1}^{I}$ of I sub-sources.

$$\hat{\mathbf{G}}_{nfl} = \sum_{i=1}^{I} \eta_{ni}\hat{\mathbf{G}}_{nifl}, \tag{7}$$

where $\hat{\mathbf{G}}_{nifl} \in \mathbb{C}^{M\times M}$ is the theoretical SCM of sub-source i normalized such that $\mathrm{tr}(\hat{\mathbf{G}}_{nifl}) = M$ and η_{ni} is the weight of sub-source i in source n normalized such that $\sum_{i=1}^{I} \eta_{ni} = 1$.

The theoretical SCM $\hat{\mathbf{G}}_{nifl}$ is computed as follows:

$$\hat{\mathbf{G}}_{nifl} \propto \mathbf{b}_{nifl}\mathbf{b}_{nifl}^{\mathrm{H}}, \tag{8}$$

where $\mathbf{b}_{nifl} \triangleq [b_{nifl1}, \cdots, b_{niflM}] \in \mathbb{C}^{M}$ is the theoretical steering vector for sub-source i in source n, which can be computed from the time of arrival of the signal emitted by sub-source i in array l as follows:

$$b_{niflm} = \exp(-\mathring{\jmath}\omega_f \|\mathbf{u}_{ni} - \mathbf{r}_{lm}\|/c), \tag{9}$$

where $\mathring{\jmath}$ is the imaginary unit, ω_f is the angular frequency corresponding to frequency bin f, and c is the speed of sound. Note that although the subsource-level SCM $\hat{\mathbf{G}}_{nifl}$ given by Eq. (8) is a rank-1 matrix, the source-level SCM $\hat{\mathbf{G}}_{nfl}$ given by Eq. (7) can be a full-rank matrix, making the inverse Wishart prior given by Eq. (6) non-degenerate. Specifically, $\hat{\mathbf{G}}_{nfl}$ becomes a full-rank matrix when at least M steering vectors with nonzero weights η_{ni} are linearly independent. This holds true in practice when I is a large number.

To reduce the model complexity, instead of treating all the sub-source weights $\boldsymbol{\eta}_n \triangleq \{\eta_{ni}\}_{i=1}^{I}$ as free parameters, we propose a parametric representation of $\boldsymbol{\eta}_n$ based on the isotropic Gaussian function as follows:

$$\eta_{ni} \propto \exp\left(-\frac{\|\mathbf{u}_{ni} - \boldsymbol{\mu}_n\|^2}{2\sigma_n^2}\right), \tag{10}$$

where $\boldsymbol{\mu}_n \in \mathbb{R}^3$ is the weighting center position and σ_n^2 is the variance that controls the spatial spread of the weights. In speech separation, $\boldsymbol{\mu}_n$ and σ_n would be close to be the mouth position and the head size of speaker n, respectively, such that the weights of sub-sources in the head take larger values. Note that an anisotropic Gaussian function could be used for future extension.

Overall, the source SCMs $\mathbf{G}$ are regularized by the 3DGS-based prior parametrized by the Gaussian positions $\mathbf{U}$, $\boldsymbol{\eta} \triangleq \{\boldsymbol{\mu}_n\}_{n=1}^N$, and $\boldsymbol{\sigma} \triangleq \{\sigma_n\}_{n=1}^N$ as follows:

$$p\left(\mathbf{G}|\mathbf{U},\boldsymbol{\mu},\boldsymbol{\sigma}\right) = \prod_{n,f,l=1}^{N,F,L} \mathcal{IW}_{\mathbb{C}}\Big(\mathbf{G}_{nfl}\,\Big|\,\nu,(\nu+M)\,\hat{\mathbf{G}}_{nfl}\Big). \quad (11)$$

C. Parameter Estimation

We aim to estimate the model parameters $\mathbf{W}$, $\mathbf{H}$, $\mathbf{G}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ that maximize the posterior distribution given by

$$\begin{aligned} p(\mathbf{G}|\mathbf{X},\mathbf{W},\mathbf{H},\mathbf{U},\boldsymbol{\mu},\boldsymbol{\sigma}) &\propto p(\mathbf{X},\mathbf{G}|\mathbf{W},\mathbf{H},\mathbf{U},\boldsymbol{\mu},\boldsymbol{\sigma}) \\ &= p(\mathbf{X}|\mathbf{W},\mathbf{H},\mathbf{G})p(\mathbf{G}|\mathbf{U},\boldsymbol{\mu},\boldsymbol{\sigma}), \end{aligned} \quad (12)$$

where the two terms of the right-hand side are given by Eqs. (4) and (11). Note that $\mathbf{U}$ is assumed to be given in this work.

We use an iterative optimization method that alternately updates the parameters until convergence. Since the NMF parameters $\mathbf{W}$ and $\mathbf{H}$ appear in only the first term of Eq. (12), we use the multiplicative update rules in the same way as MNMF:

$$w_{nkf} \leftarrow w_{nkf}\sqrt{\frac{\sum_{t=1}^T h_{nkt}\mathrm{tr}(\mathbf{G}_{nfl}\mathbf{Y}_{ftl}^{-1}\bar{\mathbf{X}}_{ftl}\mathbf{Y}_{ftl}^{-1})}{\sum_{t=1}^T h_{nkt}\mathrm{tr}(\mathbf{G}_{nfl}\mathbf{Y}_{ftl}^{-1})}}, \quad (13)$$

$$h_{nkt} \leftarrow h_{nkt}\sqrt{\frac{\sum_{f=1}^F w_{nkf}\mathrm{tr}(\mathbf{G}_{nfl}\mathbf{Y}_{ftl}^{-1}\bar{\mathbf{X}}_{ftl}\mathbf{Y}_{ftl}^{-1})}{\sum_{f=1}^F w_{nkf}\mathrm{tr}(\mathbf{G}_{nfl}\mathbf{Y}_{ftl}^{-1})}}. \quad (14)$$

where $\bar{\mathbf{X}}_{ftl} \triangleq \mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}}$.

The source SCMs $\mathbf{G}$ and the hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are updated via a gradient ascent method (e.g., Adam [28]) such that the posterior given by Eq. (12) is maximized.

To solve the scale ambiguity between $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{G}$, we incorporate the normalization constraints $\mathrm{tr}(\mathbf{G}_{nfl}) = M$ and $\sum_{f=1}^F w_{nkf} = 1$. In each iteration, we thus insert the following normalization step:

$$\phi_{nfl} \triangleq \frac{1}{M}\mathrm{tr}(\mathbf{G}_{nfl}), \qquad \begin{cases} \mathbf{G}_{nfl} & \leftarrow \phi_{nfl}^{-1}\mathbf{G}_{nfl} \\ w_{nkf} & \leftarrow \phi_{nfl}w_{nkf} \end{cases}, \quad (15)$$

$$\psi_{nk} \triangleq \sum_{f=1}^F w_{nkf}, \qquad \begin{cases} w_{nkf} & \leftarrow \psi_{nk}^{-1}w_{nkf} \\ h_{nkt} & \leftarrow \psi_{nk}h_{nkt} \end{cases}. \quad (16)$$

Although this step may decrease Eq. (12), it is empirically proven not to affect the convergence in practice.

D. Parameter Initialization

Since the performance of the proposed method is sensitive to initialization in iterative optimization, the model parameters $\mathbf{W}$, $\mathbf{H}$, $\mathbf{G}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ should be initialized appropriately. In particular, the SCM $\mathbf{G}_{nfl}$ of source n, which has a strong impact on the performance, is initialized as follows:

$$\mathbf{G}_{nfl} \leftarrow \frac{1}{I}\sum_{i=1}^I \hat{\mathbf{G}}_{nifl}, \quad (17)$$

where the theoretical SCM $\hat{\mathbf{G}}_{nifl}$ of sub-source i in source n can be computed from the Gaussian positions $\mathbf{U}$ (known in this work) according to Eqs. (8) and (9). The weighting center $\boldsymbol{\mu}_n$ and the variance $\boldsymbol{\sigma}^2$ were initialized as follows:

$$\boldsymbol{\mu}_n \leftarrow \frac{1}{I}\sum_{i=1}^I \mathbf{u}_{ni}, \quad \sigma_n^2 \leftarrow \frac{1}{2}\max_i \|\mathbf{u}_{ni} - \boldsymbol{\mu}_n\|. \quad (18)$$

$\mathbf{W}$ and $\mathbf{H}$ were randomly initialized over the interval $[0, 1)$.

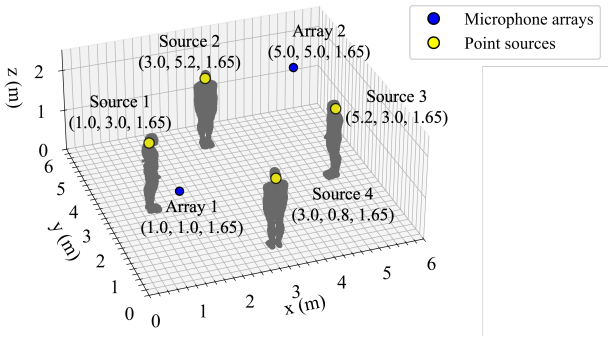

Fig. 3: 3D room layout with four sound source objects (Source1-4) and two microphone arrays(Array1-2). Point sources are positioned at the mouth locations of the objects.

IV. Evaluation

This section reports the experiment conducted for evaluating the effectiveness of the proposed method.

A. Experimental Conditions

We considered a shoebox room of size $6.0\,\mathrm{m} \times 6.0\,\mathrm{m} \times 2.5\,\mathrm{m}$, where four speakers ($N = 4$) with a height of $1.8\,\mathrm{m}$ were located at $(1.0, 3.0)$, $(3.0, 5.2)$, $(5.2, 3.0)$, and $(3.0, 0.8)$, and two four-microphone circular arrays ($L = 2$, $M = 4$) with a radius of $0.1\,\mathrm{m}$ were centered at $(1.0, 1.0)$ and $(5.0, 5.0)$ with a height of $1.65\,\mathrm{m}$ (Fig 3). We created the scene using Blender [29], [30] to render 100 images from different viewpoints for each speaker n. We then applied 3DGS [13] with $I = 10000$ on these images to obtain the 3D Gaussian absolute positions $\{\mathbf{u}_{ni}\}_{i=1}^I$.

As a simplified model of speech signal emission from the mouth, a non-directional point source was placed at a height of $1.65\,\mathrm{m}$ for each source n. The source image $\mathbf{Z}_n$ was then simulated with the image-source model using Pyroomacoustics [31], where an approximately 3-sec speech signal was taken from the CMU ARCTIC corpus [32] as the source signal. The mixture spectrogram $\mathbf{X}$ was obtained by $\mathbf{X} = \sum_{n=1}^N \mathbf{Z}_n$. The STFT used a Hann window of 128 samples (8 ms) with a 128-sample shift and 1024-point FFT at 16 kHz.

The proposed method was configured with $K = 16$ and $\nu = M$ for speech separation. In each iteration, we updated $\mathbf{W}$ and $\mathbf{H}$ once and updated $\mathbf{G}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ ten times. This procedure was iterated 200 times. For comparison, we tested three variants of the proposed method described below.

(I) **MNMF** [5], a BSS method obtained by removing the prior $p(\mathbf{G}|\mathbf{U},\boldsymbol{\mu},\boldsymbol{\sigma})$ from the proposed method.

(II) **3DGS-MNMF-NI**, the proposed method initialized with $\mathbf{G}_{nfl} \leftarrow \mathbf{I}$ in a non-informative manner.

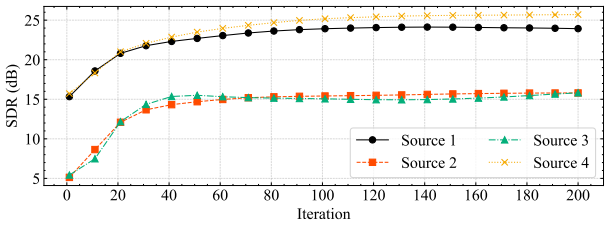

Fig. 4: SDR evolutions over the 3DGS-MNMF-VI iterations.

(III) **3DGS-MNMF-VI**, the proposed method initialized with Eq. (17) in a visually-informed manner.

For updating $\mathbf{G}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$, we used Adam [28]. Learning rates were set to 0.01 for $\mathbf{G}$ in **MNMF** and **3DGS-MNMF-NI**, 0.0001 for $\mathbf{G}$ in **3DGS-MNMF-VI**, and 0.001 for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ in all variants. To investigate the performance stability of each variant, we conducted 30 trials with random initialization of $\mathbf{W}$ and $\mathbf{H}$, and measured the separation performances in terms of the signal-to-distortion ratio (SDR) [33].

B. Experimental Results

Fig. 4 shows the SDR evolution across parameter update iterations for each source estimated using the proposed **3DGS-MNMF-VI** for a single trial. The SDRs improved steadily, reaching $23.93\,\mathrm{dB}$, $15.80\,\mathrm{dB}$, $15.81\,\mathrm{dB}$, and $25.71\,\mathrm{dB}$ for sources 1–4, respectively. The initial SDRs, i.e., at iteration 0, were $15.33\,\mathrm{dB}$, $5.13\,\mathrm{dB}$, $5.42\,\mathrm{dB}$, and $15.72\,\mathrm{dB}$. After 20 iterations, these rose to $20.66\,\mathrm{dB}$, $11.83\,\mathrm{dB}$, $11.83\,\mathrm{dB}$, and $20.84\,\mathrm{dB}$, respectively. We also note that the performances were generally saturated after 80 iterations. These results suggest that the 3DGS-based initialization of $\mathbf{G}$ allows stable optimization with fast convergence.

Fig. 5 displays the average final SDRs (after 200 iterations) along with the 95% confidence intervals (CIs) across 30 trials for the different method variants. The baseline **MNMF** showed poor separation performance and stability with relatively low SDRs and large CIs of $0.27 \pm 1.35\,\mathrm{dB}$, $-5.22 \pm 1.44\,\mathrm{dB}$, $-1.59 \pm 1.13\,\mathrm{dB}$, and $0.34 \pm 1.37\,\mathrm{dB}$ for sources 1–4, respectively. **3DGS-MNMF-NI** demonstrated that incorporating spatial priors, even with a non-informative initialization, significantly improved the separation. It achieved SDRs of $9.16 \pm 0.65\,\mathrm{dB}$, $8.93 \pm 1.42\,\mathrm{dB}$, $3.79 \pm 0.85\,\mathrm{dB}$, and $9.48 \pm 0.38\,\mathrm{dB}$. Furthermore, **3DGS-MNMF-VI** with the visual-based prior initialization achieved even higher SDRs of $22.45 \pm 0.48\,\mathrm{dB}$, $14.67 \pm 0.37\,\mathrm{dB}$, $14.88 \pm 0.64\,\mathrm{dB}$, and $24.29 \pm 0.53\,\mathrm{dB}$. **3DGS-MNMF-VI** surpassed **3DGS-MNMF-NI** in mean SDR across all four sources, and reduced the 95% CIs for three of them, demonstrating both superior separation performance and enhanced estimation stability. These results suggest that incorporating statistical constraints into the spatial model offers a favorable trade-off between separation performance and robustness against initialization.

C. Discussion

Fig. 6 shows the heatmap of the weight vector $\boldsymbol{\eta}$ computed given $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ optimized by **3DGS-MNMF-VI**. The values

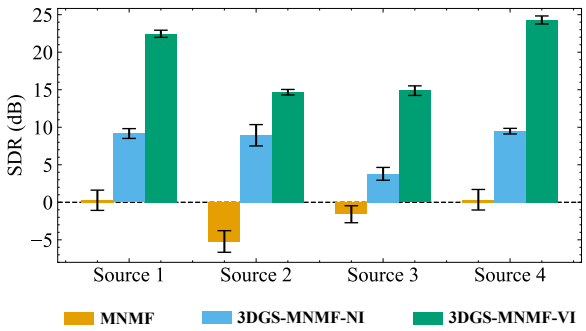

Fig. 5: Average SDRs and 95% confidence intervals.

of $\boldsymbol{\eta}$ are visualized on the corresponding 3D Gaussian centers, with color intensity representing the magnitude of each weight. The figure reveals that $\boldsymbol{\sigma}$ became smaller, i.e., highly-weighted Gaussians were concentrated within a narrow area of each object. The Euclidean distance between the estimated weighted mean and the actual source position was approximately $0.23\,\mathrm{m}$, $0.23\,\mathrm{m}$, $0.23\,\mathrm{m}$, and $0.22\,\mathrm{m}$ for sources 1-4, suggesting that the proposed method can estimate source locations based on visual cues to some extent.

Even though the estimated Gaussians became concentrated within specific regions of each object close to the actual source positions as expected, the estimated weighted mean $\boldsymbol{\mu}$ may fall outside the object. To mitigate this issue, incorporating some spatial constraints will be necessary in future work. The variance $\boldsymbol{\sigma}$ tends to decrease over iterations, resulting in a more accurate estimation. However, it may also degrade the rank of the SCMs, leading to unstable optimization. To prevent numerical instability, a lower bound on $\boldsymbol{\sigma}$ or regularization by adding a scaled identity matrix to $\hat{\mathbf{G}}_{nfl}$ may be beneficial.

V. Conclusion

We proposed a visually-informed multichannel source separation method that introduces priors based on 3D Gaussian positions on source SCMs. By integrating the MNMF-based acoustic model and the visually-informed spatial model, the proposed method achieved higher SDR and stabler results than conventional methods.

Future work includes joint optimization of the separation model and the 3D Gaussian positions and on object-wise 3D Gaussian clustering. We also plan to conduct real-world experiments to evaluate the robustness of the proposed method against reverberation and diffuse noise. In particular, we will examine the validity of the independence assumption across microphone arrays and to assess the impact of possible errors in microphone positions. This could contribute to more comprehensive audio-visual scene understanding with distributed cameras and microphone arrays.

Acknowledgment

This work was partially supported by JST FOREST Grant No. JPMJFR2270 and JSPS KAKENHI Grant Nos. 23K16912, 23K16913, 24H00742, 24H00748, 25H01142, and 25K22841.

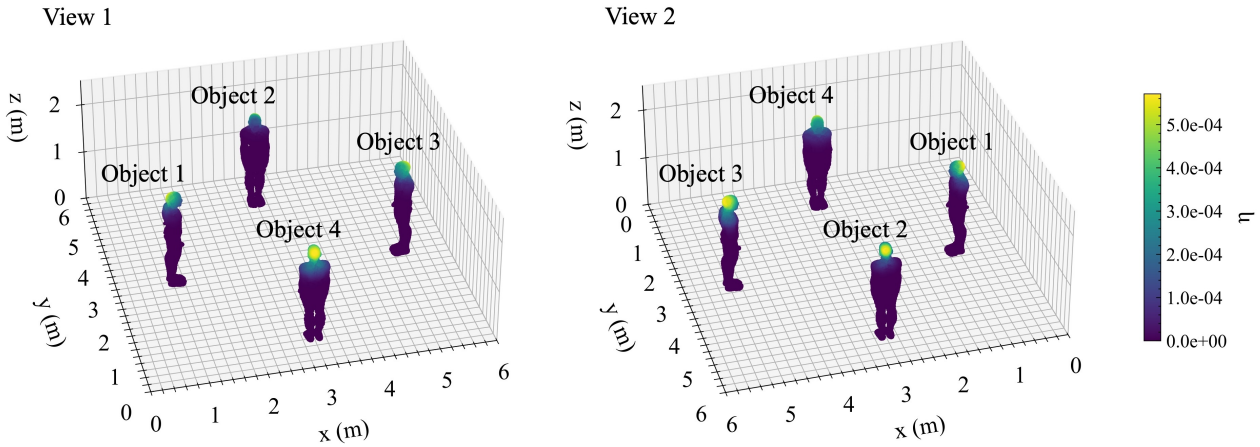

Fig. 6: Heatmap of the optimized weights $\boldsymbol{\eta}$ after convergence. The color represents the weight for each of the 10000 Gaussians estimated by 3DGS.

REFERENCES

[1] T. J. Park *et al.*, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, 2021.

[2] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *ICASSP*, 2018, pp. 5739–5743.

[3] S. Wang *et al.*, "Dasformer: Deep alternating spectrogram transformer for multi/single-channel speech separation," in *ICASSP*, 2023, pp. 1–5.

[4] S. Araki *et al.*, "30+ years of source separation research: Achievements and future challenges," in *ICASSP*, 2025, pp. 1–5.

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

[6] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[7] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel nmf and acoustic tracking," *IEEE/ACM TASLP*, vol. 26, no. 2, pp. 281–295, 2018.

[8] M. Guzik and K. Kowalczyk, "Wishart localization prior on spatial covariance matrix in ambisonic source separation using non-negative tensor factorization," in *ICASSP*, 2022, pp. 446–450.

[9] H. Munakata, Y. Bando, R. Takeda, K. Komatani, and M. Onishi, "Joint separation and localization of moving sound sources based on neural full-rank spatial covariance analysis," *IEEE SPL*, vol. 30, pp. 384–388, 2023.

[10] Y. Sumura, D. D. Carlo, A. A. Nugraha, Y. Bando, and K. Yoshii, "Joint audio source localization and separation with distributed microphone arrays based on spatially-regularized multichannel NMF," in *IWAENC*, 2024, pp. 145–149.

[11] Z. Yu and Y. Nakamura, "Smart meeting systems: A survey of state-of-the-art and open issues," *ACM Comput. Surv.*, vol. 42, no. 2, 2010.

[12] B. Mildenhall *et al.*, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, vol. 12346, 2020, pp. 405–421.

[13] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM TOG*, vol. 42, no. 4, pp. 1–14, 2023.

[14] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, "AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis," in *NeurIPS*, 2023, pp. 1–19.

[15] A. Brunetto, S. Hornauer, and F. Moutarde, "NeRAF: 3D scene infused neural radiance and acoustic fields," in *ICLR*, 2025, pp. 1–24.

[16] S. Bhosale, H. Yang, D. Kanojia, J. Deng, and X. Zhu, "AV-GS: Learning material and geometry aware priors for novel view acoustic synthesis," in *NeurIPS*, 2025, pp. 1–18.

[17] Z. Shi, L. Zhang, and D. Wang, "Audio-visual sound source localization and tracking based on mobile robot for the cocktail party problem," *Appl. Sci.*, vol. 13, no. 10, p. 6056, 2023.

[18] M. E. B. Menai, H. Abbar, L. Bouafif, I. E. Hassani, and A. A. Moussa, "Audio-visual multimodal fusion for human activity recognition in smart meeting rooms," *Appl. Sci.*, vol. 13, no. 10, p. 6056, 2023.

[19] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[20] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Process.*, vol. 87, no. 8, pp. 1859–1871, 2007.

[21] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *ICA*, 2006, pp. 165–172.

[22] 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 TASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[23] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[24] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP*, 2019, pp. 371–375.

[25] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM TASLP*, vol. 28, pp. 2610–2625, 2020.

[26] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3D scenes," in *ECCV*, 2024, p. 162–179.

[27] J. Zhang, J. Jiang, Y. Chen, K. Jiang, and X. Liu, "COB-GS: Clear object boundaries in 3DGS segmentation based on boundary-adaptive gaussian splitting," in *CVPR*, 2025, pp. 19 335–19 344.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015, pp. 1–15.

[29] Blender Foundation, "Blender manual," Version 4.1, Blender Foundation, https://docs.blender.org/manual/en/dev/, 2025.

[30] MB-Lab Community, "MB-Lab: Open-source 3D humanoid character generator for blender," Version 1.7.8, MB-Lab Community, https://mb-lab-community.github.io/, 2025.

[31] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *ICASSP*, 2018, pp. 351–355.

[32] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *SSW5*, 2004, pp. 223–224.

[33] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.