2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3315-1484-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/VRW66409.2025.00352

Extending Gaussian Splatting to Audio: Optimizing Audio Points for Novel-view Acoustic Synthesis

Masaki Yoshida* Ren Togo† Takahiro Ogawa† Miki Haseyama†

* Graduate School of Information Science and Technology, Hokkaido University

† Faculty of Information Science and Technology, Hokkaido University

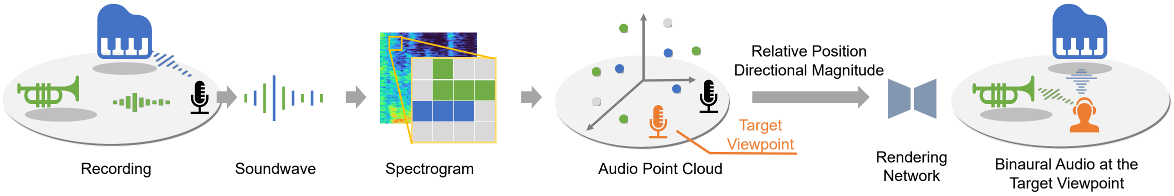

Figure 1: Our framework: The key concept is projecting each spectrogram pixel into 3D space for spatial optimization. The rendering network predicts target view binaural audio based on relative position and directional magnitude.

Abstract

This paper proposes a novel method to extend Gaussian Splatting (3DGS) to the audio domain, enabling novel-view acoustic synthesis solely using audio data. While recent advancements in 3DGS have significantly improved novel-view synthesis in the visual domain, its application to audio has been overlooked, despite the critical role of spatial audio for immersive AR/VR experiences. Our method addresses this gap by constructing an audio point cloud from audio at source viewpoints and rendering spatial audio at arbitrary viewpoints. Experimental results show that our method outperforms existing approaches relying on audio-visual information, demonstrating the feasibility of extending 3DGS to audio.

Index Terms: Gaussian splatting, novel-view acoustic synthesis, binaural audio.

1 Introduction

3D Gaussian Splatting (3DGS) [7] has revolutionized novel-view synthesis, enabling the reconstruction of highly accurate 3D scenes from multiple-view 2D images. While 3DGS has significantly advanced 3D visual reconstruction, extending its capabilities to audio remains largely uncharted. This gap is noteworthy, as spatial audio plays a crucial role in delivering an immersive experience in AR/VR environments. Expanding 3DGS to the audio domain would enable the rendering of target-view audio from multiple recordings at source viewpoints.

3DGS has yet to be applied in the context of spatial audio generation. Chen et al. [3] proposed the task of novel-view acoustic synthesis (NVAS). They employed deep neural networks to generate the binaural audio for target viewpoints based on audio-visual observations from multiple source viewpoints. Subsequently, Liang et al. [8] introduced an approach that leverages NeRF [9] to render a target view image as a condition for audio generation. Recent studies [1] have extended this concept by incorporating acoustic parameters into points constructed by 3DGS, aiming to model sound propagation within the scene. However, these methods rely heavily on visual data and cannot operate without it. Furthermore, point-based approaches do not embed sound information and therefore lack interpretability of acoustic parameters.

As the first attempt to extend 3DGS to audio, we propose a novel approach that treats each pixel in the spectrogram as a point in 3DGS. Our approach enables rendering binaural audio at arbitrary viewpoints by optimizing the parameters of the audio points. Furthermore, our approach eliminates dependence on visual information and embeds the audio information into the points.

2 Methodology

2.1 Extending 3DGS to Audio

The spectrogram, a time-frequency representation obtained by applying short-time Fourier transform to audio waveforms, serves as the foundation for our approach. Spectrograms offer several advantages for this purpose. First, spectrograms can be treated as images, allowing the application of computer vision techniques. Second, they are typically sparse enough that different sound sources' information does not overlap within individual pixels. This is a reasonable assumption, supported by the success of source separation [6] based on spectrogram masking. Based on these characteristics, we treat each pixel in the spectrogram as a point in 3DGS and place these points in 3D space. Optimizing these points ensures rendering from arbitrary target viewpoints. We assign the following parameters to each spectrogram pixel in our method, adopting and extending concepts from 3DGS:

(1) **Position in 3D space:** Each audio point is assigned a position in 3D space to simulate the origin of the sound.

(2) **Magnitude represented as spherical harmonics coefficients:** Following 3DGS, which represents view-dependent appearance with spherical harmonics (SH), our method uses SH to represent view-dependent magnitude. The zeroth-order coefficient indicates view-independent magnitude, equivalent to the magnitude in conventional spectrograms, and higher orders capture directionality.

(3) **Rotation matrix:** The orientation of each point determines the axis of the spherical harmonics.

Since these parameters are adapted from the original 3DGS, their optimization follows the same process as in the original 3DGS. In addition to these parameters, the points retain their T-F positional information in the spectrogram. This enables the inverse transformation from the point structure back to the spectrogram structure ($T \times F \times$ parameters).

To render binaural audio at a target viewpoint, we place a virtual microphone at the target position. This enables the computation of each point's relative position and directional magnitude to the microphone. We then transform parameter representations back into spectrogram structure and feed them into the rendering network. For this purpose, we employed U-Net [10] to predict binaural audio at the target viewpoint. Following the practice of previous methods, the rendering network outputs spectrogram masks instead of directly predicting the left and right channels. This approach leverages the transformation of the left and right channels a_{L} and a_{R}

*e-mail: masaki@lmd.ist.hokudai.ac.jp

†e-mail: {togo, ogawa, mhaseyama}@lmd.ist.hokudai.ac.jp

979-8-3315-1484-6/25/$31.00 ©2025 IEEE
DOI 10.1109/VRW66409.2025.00352

into two components: the monaural signal, $a_{\text{mono}} = a_{\text{L}} + a_{\text{R}}$, which captures distance-related information, and the differential signal, $a_{\text{diff}} = a_{\text{L}} - a_{\text{R}}$, which encodes directional cues. The spectrogram masks are applied to the input audio to predict these components effectively.

2.2 Novel-View Acoustic Synthesis

We optimize the parameters and the network by following the NVAS task. In NVAS, the objective is to generate binaural audio at target viewpoints from audio-visual observation pairs captured from multiple source viewpoints, along with mono audio recorded near the sound source. While camera position information is available, sound source position information is not utilized.

The training process begins with scene-level audio normalization. Conventional audio processing typically normalizes each clip independently. However, relative volume differences within a scene, which reflect distance information, should be preserved. To address this issue, we normalize the audio by using the maximum Root Mean Square value across all clips within the same scene. Following normalization, we initialize point parameters by setting the mono audio recorded near the sound source as the zeroth-order magnitude, with the remaining parameters initialized to zero. During training, the network is optimized across multiple source viewpoints to minimize the L2 distance between the predicted spectrograms of a_{mono}, a_{diff}, and their corresponding ground truth.

3 Experiments

To validate the effectiveness of our method, we evaluate the performance in generating binaural audio within the NVAS framework. We employ the Replay dataset [11], which was proposed alongside the NVAS task. This dataset consists of multiple scenes (e.g., conversation scenes, table tennis scenes), each split into 2-second clips. Each clip includes audio-visual pairs captured from 8 viewpoints. We use a total of 330 seconds from 6 scenes as test data. We use 7 of the 8 viewpoints for training and the remaining one for testing. As a baseline, we use the ViGAS [3] model proposed with the NVAS task. Additionally, we include the AV-NeRF [8] as the state-of-the-art method. This method requires the sound source position as input, which is unavailable and is therefore set to zero. We used the pretrained ViGAS and trained the AV-NeRF based on their official code. We also report results when evaluating the input monaural audio. Accuracy higher than this value indicates the ability to reconstruct binaural effects. We evaluate NVAS performance using the following four metrics. Magnitude Spectrogram Distance (MAG) [4] measures the magnitude distance of spectrograms. Envelope Distance (ENV) [5] is calculated based on the energy envelope of waveforms. Left-Right Energy Ratio Error (LRE) [4] evaluates the ratio of energy between the left and right channels. RT60 Error (RTE) [2] measures reverberation time discrepancies, defined as the time required for sound to decay by 60 dB. These metrics collectively address spectrogram accuracy, spatial sound properties, and reverberation characteristics.

Table 1 shows the evaluation results. We can see that our method outperforms other methods. Remarkably, despite relying solely on audio information, our method achieves higher accuracy than these audio-visual methods. This demonstrates the effectiveness of our approach in leveraging audio information. Figure 2 visualizes the positions of the 1,000 points with the highest and lowest magnitudes after optimization. Lower-magnitude points are clustered near the initialization origin, while higher-magnitude points are more widely distributed. This distribution indicates that the optimization process is performing as designed, embedding sound information into points and adjusting their positions. As part of the discussion, the interpretability of point positions is recognized as a key area for future study. At present, our method focuses on optimizing point positions to improve prediction accuracy. However, improving the alignment between visual and audio points could provide deeper insights into the data representation.

Table 1: Results of NVAS task.

	Audio	Visual	MAG↓	ENV↓	LRE↓	RTE↓
Input Monoral	✓	✗	6.057	0.260	8.301	0.047
ViGAS [3]	✓	✓	1.754	0.185	11.293	0.054
AV-NeRF [8]	✓	✓	3.412	0.201	**4.777**	0.068
Ours	✓	✗	**1.118**	**0.150**	5.071	**0.037**

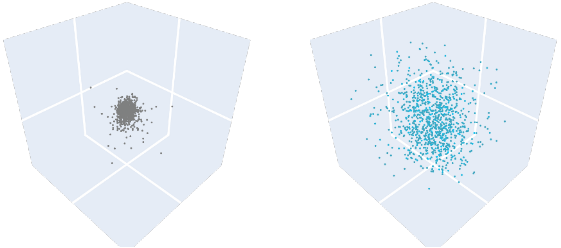

Figure 2: Visualization of points after optimization. The left figure shows the positions of the 1,000 points with the lowest magnitudes, while the right figure shows those with the highest magnitudes.

4 Conclusion

We have proposed a novel approach that treats each spectrogram pixel as an audio point, which achieves high-accuracy NVAS without relying on visual information or sound source positions. Furthermore, the ability of our method to generate target view audio freely from optimized points and the network contributes to AR/VR applications, similar to the advancements brought by 3DGS in visual reconstruction.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP24K02942, JP23K21676, JP23K11141, and JP24KJ0325.

References

[1] 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*, 2024. 1

[2] C. Chen, R. Gao, P. Calamia, and K. Grauman. Visual acoustic matching. In *Proc. CVPR*, pp. 18858–18868, 2022. 2

[3] C. Chen, A. Richard, R. Shapovalov, V. K. Ithapu, N. Neverova, K. Grauman, and A. Vedaldi. Novel-View acoustic synthesis. In *Proc. CVPR*, pp. 6409–6419, 2023. 1, 2

[4] C. Chen, W. Sun, D. Harwath, and K. Grauman. Learning audio-visual dereverberation. In *Proc. ICASSP*, pp. 1–5, 2023. 2

[5] R. Gao and K. Grauman. 2.5D visual sound. In *Proc. CVPR*, pp. 324–333, 2019. 2

[6] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *Proc. ICASSP*, vol. 5, pp. 2985–2988, 2000. 1

[7] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 1

[8] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu. AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis. In *Proc. NeurIPS*, vol. 36, pp. 37472–37490, 2023. 1, 2

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), Dec. 2021. 1

[10] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, pp. 234–241, 2015. 1

[11] R. Shapovalov, Y. Kleiman, I. Rocco, D. Novotny, A. Vedaldi, C. Chen, F. Kokkinos, B. Graham, and N. Neverova. Replay: Multi-modal multi-view acted videos for casual holography. In *Proc. ICCV*, pp. 20338–20348, 2023. 2