AMBISONIC-DML: Higher-Order Ambisonic Music Dataset for Spatial AI Generation

Seungryeol Paik
Dept. of Intelligence, Seoul National University
paik402@snu.ac.kr

Kyogu Lee
Dept. of Intelligence, Seoul National University
kyogu.lee@snu.ac.kr

Abstract

We present AMBISONIC-DML, a dataset of 120 musical excerpts rendered in fifth-order Ambisonics (HOA5) with synchronized motion trajectories sampled at 50,fps. Despite its compact size, the dataset offers high informational density through 36-channel HOA encoding, 50,fps motion capture, and structured stem-level annotations, providing the first open and reproducible resource for dynamic Ambisonic music. The dataset was recorded under controlled studio conditions with composer-defined motion aligned to phrasing and rhythm. Objective and perceptual analyses confirm accurate HOA5 encoding, balanced spatial energy, and perceptual improvements in localization and immersion. AMBISONIC-DML enables reproducible research on spatial signal processing and generative modeling.[1]

1 Introduction

Spatial audio is a key component of immersive media such as virtual and augmented reality. Ambisonics provides a scene-based, playback-independent representation of 3D sound fields, and higher orders (HOA) enhance spatial resolution and localization accuracy over first-order Ambisonics (FOA) [1, 2, 3, 4]. Despite rapid progress in AI-driven music generation [5, 6, 7, 8, 9], most research remains focused on symbolic scores, stereo audio, or text-conditioned models. These approaches overlook the spatial dimension of music, where motion and perspective evolve dynamically, yet datasets coupling musical structure with spatial trajectories are scarce. Existing corpora such as TAU-NIGENS and STARSS23 [10, 11] focus on environmental FOA recordings, while available HOA datasets capture static musical scenes without stem separation or motion tracking. This lack of structured HOA data constrains reproducible research on dynamic spatial music, where motion and phrasing are inherently linked. AMBISONIC-DML addresses this gap with music excerpts rendered in fifth-order Ambisonics (HOA5) and synchronized motion trajectories sampled at 50 fps. The trajectories are composer-defined, temporally aligned with phrasing and rhythm, and paired with instrument-role annotations. This design facilitates research on motion-informed rendering, spatial signal processing, and generative modeling of immersive music, extending beyond the stereo or symbolic domains of prior work.

[1] Released under CC-BY-NC 4.0: https://drive.google.com/drive/folders/1wLmoDqqYOStCFt_WTny7jSUxrBjV5C56?usp=sharing

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

2 Dataset

2.1 Overview

The AMBISONIC-DML dataset comprises 120 music excerpts rendered in fifth-order Ambisonics (36 channels, ACN/SN3D) from the author's spatial audio installations. Each entry provides three synchronized modalities: (1) dry stems, (2) HOA5 spatialized mixes, and (3) XYZ motion trajectories sampled at 50 fps. Excerpts range from short fragments to complete pieces and are organized for multimodal access. The dataset prioritizes motion precision and musical alignment, offering reproducible ground truth for dynamic spatial audio research.

2.2 Structure and Format

The dataset is organized into three top-level directories: `RAW_DATA`, `REFINED_DATA`, and `CODE`. `RAW_DATA` contains the original multitrack recordings and motion control logs for two projects (`EXPEDITION` and `DIALOGUE`), each with subfolders for OSC data, Ambisonic renders, and REAPER sessions containing compressed dry stems and reference screenshots.

`REFINED_DATA` provides the synchronized research release, preserving the same projects but divided into *ORIGINAL* and *PRE_PROCESSED* stages. *ORIGINAL* includes selected dry stems, HOA files, and motion logs derived directly from `RAW_DATA`. *PRE_PROCESSED* contains normalized materials (dry, full mix, HOA, motion, and synchronization logs), all mute-trimmed and frame-aligned. A global `annotation.xls` file lists instrument, role, and trajectory metadata.

`CODE` includes utilities for OSC capture, Ambisonic playback, preprocessing, and conversion between Ambisonic orders (FOA, HOA3, HOA5, and binaural). Motion logs are stored as CSV files with timestamps and XYZ coordinates at 50 fps. The structure corresponds exactly to the public release and resolves prior inconsistencies between documentation and distribution.

2.3 Recording and Spatialization

All stems were recorded in acoustically treated studios using high-quality microphones and DI paths. A 5×5 m room (RT60 = 0.32 s) was used for vocals and winds, and a 6.5×7 m room (RT60 = 0.38 s) for percussion. Signals were captured through Universal Audio Apollo X interfaces using OC818, TLM103, and SM57 microphones. Spatialization was performed in SPAT Revolution with OSC-driven motion exported from a DAW. Each trajectory followed phrasing and rhythm, providing a deterministic mapping between musical structure and motion.

2.4 Audio Statistics

Excerpts range from 3.2 s to 5.1 min in duration (mean = 119 s). Dry stems were loudness-normalized to –35 LUFS following ITU-R BS.1770-4, preserving natural dynamics while ensuring perceptual consistency. HOA5 renderings maintain balanced spatial energy across channels, while motion trajectories exhibit a mean displacement amplitude of 0.8 m (SD = 0.5 m), capturing both localized gestures and wide spatial movements characteristic of dynamic musical motion.

2.5 Motion and Annotation

Each stem in AMBISONIC-DML is annotated with its instrument class and functional role (melody, rhythm, chords, ambience, pad), enabling structured conditioning for generative and analytical tasks. Role distribution is melody (44%), rhythm (33%), chords (11%), ambience (6%), and pad (6%). Instrument classes include analog and digital synthesizers (43.3%), guitars (10.8%), vocals (10.0%), brass and winds (12.5%), percussion (9.1%), traditional instruments (5.0%), and ambience or effect-based sources (9.3%).

Motion trajectories are expressed as XYZ coordinates relative to the listener at the origin. Typical patterns include circular sweeps for melodies, oscillations for pads, frontal stability for bass and percussion, and transient bursts for percussive accents. Melodic stems trace smooth curves aligned with phrasing, while rhythmic and background elements remain localized or static, reflecting musically meaningful spatial organization.

2.6 Comparison with Existing Datasets

Compared with prior spatial datasets, AMBISONIC-DML emphasizes musical motion and multimodal precision. Datasets such as TAU-NIGENS [10] and EigenScape [12] capture acoustic scenes in FOA or HOA formats but focus on ambient or environmental events rather than structured music. STARSS23 [11] extends FOA recordings with synchronized video for sound event localization, while existing HOA datasets capture static ensembles without motion information. Audiovisual corpora such as FAIR-Play [13] and A2B [14] investigate cross-modal alignment or binaural rendering but lack continuous motion trajectories and compositional metadata. Recent music generation datasets [7, 9] remain confined to symbolic or stereo domains.

3 Baseline Evaluation

We conducted baseline evaluations to verify the technical validity of AMBISONIC-DML. The analyses comprise (1) objective signal assessment and (2) perceptual evaluation across Ambisonic orders, providing reproducible references for future research. As a dataset-oriented contribution, the focus is on verifying signal quality, spatial accuracy, and perceptual reliability rather than benchmarking specific models. Given the absence of established evaluation protocols for spatial audio—particularly for fifth-order Ambisonics and musical material—we perform both objective and perceptual analyses to assess the dataset itself.

3.1 Objective Evaluation

Objective analyses verified the signal integrity and spatial accuracy of AMBISONIC-DML. Loudness and crest factor were computed following ITU-R BS.1770-4 with 400 ms gating using `pyLoudnorm`. Dry stems exhibited stable loudness (mean –35 LUFS, SD 1.5) and balanced dynamics (mean crest factor 13.7 dB), confirming a natural dynamic range without over-compression. HOA5 renderings preserved amplitude relationships across 36 channels (ACN/SN3D), validating correct encoding. Spatial resolution, defined as $N^2 + 2N$ for order N [2], confirmed full fifth-order representation.

Directional balance, analyzed via active intensity vectors [15], showed even energy distribution across azimuth and elevation (ratio ≈0.33/0.34/0.33) with minor elevation variability consistent with known perceptual thresholds. Diffuseness, measured using `AmbiX`, averaged 0.89 (SD 0.09), indicating a controlled yet immersive sound field.

Spectral centroid and flatness, evaluated after binaural decoding with MagLS [4], showed a slight centroid reduction (mean shift –0.34 kHz) and marginal flatness increase, consistent with expected high-frequency diffusion in non-individualized HRTFs [16]. Motion precision, computed by comparing rendered trajectories with OSC logs, yielded a mean deviation within ±0.10 mm, confirming tight audio–motion synchronization. Preliminary FOA and HOA3 downmixes followed similar trends, exhibiting lower diffuseness and spatial resolution. All metrics were obtained using open-source toolkits (`pyLoudnorm`, `AmbiX`, `MagLS`) to ensure reproducibility.

3.2 Perceptual Evaluation

A listening test assessed the perceptual quality of HOA5 renderings relative to Stereo, HOA1, and HOA3 versions. Twenty-five participants (10 experts, 15 non-experts) rated five excerpts on four perceptual attributes: localization, immersion, clarity, and rhythm. All excerpts were loudness-normalized and binaurally decoded using the IEM Renderer to ensure consistent playback conditions.

Results revealed consistent perceptual trends across Ambisonic orders. HOA5 achieved the highest mean opinion scores in localization (4.5) and immersion (4.7), confirming perceptual benefits of higher-order reproduction. Stereo received the highest ratings in clarity (4.3) and rhythm (4.6), reflecting sharper transients and reduced spatial diffusion. HOA3 provided a perceptually balanced compromise, maintaining spatial definition without the mild smoothing observed in HOA5. A one-way ANOVA with Tukey HSD indicated significant effects for localization and immersion ($p < 0.01$), demonstrating that higher-order Ambisonics enhances spatial perception at a minor cost in transient precision.

3.3 Discussion

These baseline results demonstrate that AMBISONIC-DML preserves signal integrity and perceptual validity across spatial formats, maintaining stable loudness, correct HOA encoding, spatial uniformity, and precise motion alignment. HOA5 encoding provides clear improvements in localization and immersion, while Stereo remains advantageous for transient clarity, reflecting the known spectral–spatial trade-off in higher-order Ambisonics. Overall, the dataset offers a technically verified and perceptually validated foundation for benchmarking trajectory-informed and generative spatial audio models.

4 Applications and Future Work

AMBISONIC-DML provides a foundation for AI-driven research linking sound, space, and motion. It supports three primary directions: (1) *motion-informed generative modeling*, learning to produce HOA5 audio or synchronized motion trajectories; (2) *spatial parameter estimation*, offering ground truth for trajectory, energy distribution, and order balance; and (3) *trajectory-conditioned processing*, including motion-aware separation, rendering, and decoding across Ambisonic orders.

The dataset includes fully mixed FOA, HOA3, and binaural versions. Ongoing models for HOA generation and motion-parameter estimation using trajectory-conditioned diffusion and transformer architectures reproduce realistic spatial motion and energy coherence, demonstrating the dataset's suitability for dynamic HOA modeling.

As a shared benchmark for the spatial audio community, AMBISONIC-DML enables consistent comparison across Ambisonic orders, playback formats, and motion conditions. Researchers can evaluate localization accuracy, motion smoothness, and perceptual coherence under identical conditions, establishing reproducible benchmarking analogous to DCASE but oriented toward musical and motion-based studies. By providing open stems and spatial renderings, it bridges artistic production and computational modeling, fostering collaboration between creators, engineers, and AI researchers.

Beyond generative applications, the dataset supports research on motion–timbre correlation, spatial salience prediction, and audio–motion synchronization, extending to cross-modal perception and embodied interaction in immersive music. Scripts in the `CODE` directory integrate with PyTorch and Max/MSP for reproducible experimentation.

Future extensions will broaden genre diversity, include longer compositions, and add multimodal metadata (e.g., score and visual context) for cross-domain conditioning. Together, these developments aim to establish AMBISONIC-DML as a standard benchmark for *spatially aware music generation* and *trajectory-informed modeling*.

Although AMBISONIC-DML comprises 120 excerpts, each entry encodes a rich combination of spatial, temporal, and musical information: 36 Ambisonic channels, 50 fps motion trajectories, and detailed instrument-role annotations. This yields high information density comparable to large-scale symbolic or stereo corpora while maintaining the precision required for spatial AI research. By capturing fine-grained motion aligned with compositional intent, the dataset bridges artistic expressivity and computational modeling, offering a compact yet comprehensive foundation for studying spatial structure and generative audio behavior.

5 Conclusion

We introduced AMBISONIC-DML, a dataset of 120 music excerpts rendered in fifth-order Ambisonics with synchronized motion trajectories and stem-level annotations. It bridges environmental FOA corpora and static HOA recordings by providing musically meaningful, dynamic motion aligned with phrasing and rhythm. Baseline analyses confirmed consistent loudness, accurate encoding, and sub-millimeter motion precision, while perceptual tests demonstrated perceptual gains in localization and immersion. By offering reproducible, multimodal ground truth for dynamic Ambisonic music, AMBISONIC-DML establishes a benchmark for trajectory-informed processing and generative spatial modeling. Future extensions will expand genre diversity, incorporate multimodal metadata, and further support research in AI-driven immersive music.

References

[1] M. A. Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, Jan./Feb. 1973.

[2] J. Daniel, S. Moreau, and R. Nicol. Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging. In *114th AES Convention*, Amsterdam, The Netherlands, Mar. 2003.

[3] T. Carpentier, M. Noisternig, and O. Warusfel. Comparing the ambisonic decoding methods: Influence of the order on spatial quality. In *Audio Engineering Society Convention 142*, Berlin, Germany, May 2017.

[4] F. Zotter and M. Frank. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer, Cham, Switzerland, 2019.

[5] S. Ji, X. Yang, and J. Luo. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys*, May 2023. `https://dl.acm.org/doi/10.1145/3597493`.

[6] J. P. Briot, G. Hadjeres, and F. D. Pachet. Deep learning techniques for music generation – a survey. *arXiv preprint arXiv:1709.01620*, Jan. 2017.

[7] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[8] H. Yakura and M. Goto. IteraTTA: An interface for exploring both text prompts and audio priors in generating music with text-to-audio models. *arXiv preprint arXiv:2307.13005*, Jul. 2023.

[9] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, Jan. 2023.

[10] A. Politis, S. Adavanne, and T. Virtanen. TAU-NIGENS Spatial Sound Events 2021. *Zenodo*, Feb. 2021. `https://zenodo.org/record/4568780`.

[11] K. Shimada, A. Politis, P. Sudarsanam, D. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *arXiv preprint arXiv:2306.09126*, Jun. 2023.

[12] M. C. Green. EigenScape: A database of spatial acoustic scenes recorded with a soundfield microphone. *Applied Sciences*, vol. 7, no. 11, p. 1204, Nov. 2017.

[13] R. Gao and K. Grauman. 2.5D visual sound. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 324–333, 2019.

[14] I. D. Gebru, T. Keebler, J. Sandakly, S. Krenn, D. Marković, J. Buffalini, S. Hassel, and A. Richard. A2B: Neural rendering of Ambisonic recordings to binaural. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. `https://ieeexplore.ieee.org/document/10890507`.

[15] J. Merimaa. *Analysis, Synthesis, and Perception of Spatial Sound: Binaural Localization Modeling and Multichannel Loudspeaker Reproduction*. Doctoral Dissertation, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2006. `https://aaltodoc.aalto.fi/items/5f2b1a58-617d-4a25-808f-86a2c2b1b239`.

[16] E. M. Wenzel, F. L. Wightman, D. M. Kistler, and M. Arruda. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993. `https://doi.org/10.1121/1.406784`.