MRSAudio: A Large-Scale Multimodal Recorded Spatial Audio Dataset with Refined Annotations

Wenxiang Guo[†1,2], **Changhao Pan**[†1], **Zhiyuan Zhu**[†1], **Xintong Hu**[†1], **Yu Zhang**[†1]
Li Tang[1], Rui Yang[1], Han Wang[1], Zongbao Zhang[1], Yuhan Wang[1], Yixuan Chen[1], Hankun Xu[1]
Ke Xu[1], Pengfei Fan[1], Zhetao Chen[1], Yanhao Yu[1], Qiange Huang[1], Fei Wu[1], Zhou Zhao[‡1,2]
[1]Zhejiang University [2] Shanghai AI Laboratory
{guowx314,panch,zhaozhou}@zju.edu.cn
[†]Equal contribution. [‡]Corresponding author.

Figure 1: Overview of MRSAudio. The dataset comprises four real-world scenarios: MRSSpeech, MRSLife, MRSMusic, and MRSSing, each with multimodal annotations for spatial audio research.

Abstract

Humans rely on multisensory integration to perceive spatial environments, where auditory cues enable sound source localization in three-dimensional space. Despite the critical role of spatial audio in immersive technologies such as VR/AR, most existing multimodal datasets provide only monaural audio, which limits the development of spatial audio generation and understanding. To address these challenges, we introduce MRSAudio, a large-scale multimodal spatial audio dataset designed to advance research in spatial audio understanding and generation. MRSAudio spans four distinct components: MRSLife, MRSSpeech, MRSMusic, and MRSSing, covering diverse real-world scenarios. The dataset includes synchronized binaural and ambisonic audio, exocentric and egocentric video, motion trajectories, and fine-grained annotations such as transcripts, phoneme boundaries, lyrics, scores, and prompts. To demonstrate the utility and versatility of MRSAudio, we establish five foundational tasks: audio spatialization, and spatial text to speech, spatial singing voice synthesis, spatial music generation and sound event localization and detection. Results show that MRSAudio enables high-quality spatial modeling and supports a broad range of spatial audio research. Demos and dataset access are available at `https://mrsaudio.github.io`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/085713-4882

1 Introduction

Humans rely on multisensory integration to perceive and interpret physical environments. With the rapid growth of film, virtual reality (VR), augmented reality (AR), and gaming applications, users increasingly expect not only precise audiovisual alignment but also highly immersive experiences. While recent advances in deep learning have enabled realistic generation of speech, music, and sound effects synchronized with text or video (Kreuk et al., 2022; Yang et al., 2023; Wang et al., 2024), most models focus on monaural audio and neglect spatialized soundscapes that enhance immersion. The human binaural system uses interaural time differences (ITD) and interaural level differences (ILD) to localize sound in three-dimensional space, and spatial audio must remain consistent with visual cues (Yost, 1998; Cohen & Knudsen, 1999; Grothe et al., 2010). Any mismatch can disrupt immersion and weaken the sense of presence. For example, hearing a cat's meow from the left immediately suggests its location, even if it is just off-screen.

Despite the growing importance of spatial audio in these immersive technologies(Xie, 2020; Huiyu et al., 2025), progress in machine learning for spatial audio understanding is limited by the fundamental spatial data constraints. Most existing audio datasets(Gemmeke et al., 2017; Chen et al., 2020; Agostinelli et al., 2023) focus on monaural recordings, which discard vital spatial information, effectively "flattening" the soundscape and preventing models from learning key physical phenomena such as room reverberation, echo patterns, and sound propagation. Moreover, the scarcity of multimodal datasets that align spatial audio with synchronized visual, position geometric, and semantic annotations hinders the development of advanced auditory scene-analysis systems capable of human-like spatial perception.

To address these gaps, we present **MRSAudio**, a 484-hour large-scale multimodal spatial audio dataset designed to support both spatial audio understanding and generation. It integrates high-fidelity spatial recordings with synchronized video, 3D pose tracking, and rich semantic annotations, enabling comprehensive modeling of real-world auditory scenes. As shown in Figure 1, the dataset comprises four subsets, each targeting distinct tasks and scenarios. **MRSLife** (129 h) captures daily activities such as board games, cooking, and office work, using egocentric video and FOA audio annotated with sound events and speech transcripts. **MRSSpeech** (206 h) includes binaural conversations from 44 speakers across diverse indoor environments, paired with video, 3D source positions, and complete scripts. **MRSSing** (80 h) features high-quality solo singing performances in Chinese, English, German, and French by 20 vocalists, each aligned with time-stamped lyrics and corresponding musical scores. **MRSMusic** (69 h) offers spatial recordings of 23 Traditional Chinese, Western and Electronic instruments, with symbolic score annotations that support learning-based methods for symbolic-to-audio generation and fine-grained localization. Together, these four subsets support a broad spectrum of spatial audio research problems, including event detection, sound localization, and binaural or ambisonic audio generation. By pairing spatial audio with synchronized exocentric and egocentric video, geometric tracking, and detailed semantic labels, MRSAudio enables new research directions in multimodal spatial understanding and cross-modal generation. Throughout this paper, unless stated otherwise, we use the term spatial audio to refer to binaural audio.

- We introduce **MRSAudio**, a 484-hour, large-scale multimodal spatial audio dataset explicitly designed to push the boundaries of spatial audio understanding and generative modeling.
- We assemble synchronized binaural and ambisonic recordings with exocentric and egocentric video, geometric source positions, transcripts, scores, lyrics, and event labels, providing one of the most richly annotated multimodal resources for spatial audio research.
- We organize the data into four complementary subsets (MRSLife, MRSSpeech, MRSSing, MRSMusic), each carefully tailored to different real-world acoustic scenarios and equipped with rich, scenario-specific annotations to facilitate downstream task development.
- We establish and release evaluation protocols and baseline implementations for five benchmark tasks: audio spatialization generation, spatial text-to-speech, spatial singing voice synthesis, spatial music generation, and localization and detection of sound events, in order to demonstrate MRSAudio's versatility and to foster reproducible research.

The remainder of this paper is organized as follows. Section 2 reviews existing spatial audio datasets. Section 3 describes the design, collection process, and key statistics of MRSAudio. Section 4 presents extensive benchmark experiments using state-of-the-art methods on five core spatial audio tasks:

audio spatialization, spatial text-to-speech, spatial singing voice generation, spatial music synthesis, and sound event localization and detection. Finally, Section 5 concludes the paper and discusses the limitations and potential risks associated with MRSAudio.

2 Related Work

Deep learning has driven remarkable progress in both audio generation (Huang et al., 2023a; Yang et al., 2023; Kreuk et al., 2022) and audio understanding (Chu et al., 2023; Huang et al., 2023b; Tang et al., 2024) tasks. However, the majority of these advances still rely heavily on monaural audio, which lacks the ability to represent or capture the rich spatial cues that naturally occur in real-world environments. The rapid adoption of VR/AR technologies has concurrently driven growing demand for immersive spatial audio experiences. Researchers have focused on several key technologies including: sound event localization and detection (Adavanne et al., 2019; Wang et al., 2022), mono-to-spatial audio conversion (Gao & Grauman, 2019; Pedro Morgado & Wang, 2018), and end-to-end spatial audio generation (Sun et al., 2024; Kim et al., 2025; Liu et al., 2025).

Despite recent progress, spatial audio generation and understanding remain constrained by the paucity of high-quality datasets. Due to the difficulty and expense of collecting and annotating high-quality spatial audio datasets, most open-source datasets still primarily consist of simulated or web-crawled content. Simulated datasets offer precise annotations but lack perceptual realism, while crawled datasets often provide real-world diversity but are missing critical labels, such as listener and source positions and content-level annotations, thus limiting their utility. For instance, spatial symbolic music generation demands accurate musical scores and corresponding positional metadata. To better understand the current landscape, we survey existing spatial audio datasets. These datasets differ in terms of audio format, including First-Order Ambisonics (FOA), multi-channel arrays, and binaural microphones, as well as in their collection methods (simulated, crawled, recorded) and annotation.

Table 1: Comparison of spatial audio datasets, where T denotes speech transcripts, P represents sound source positions, N indicates natural language prompts, and C stands for sound class tags

Dataset	**Audio Format**			**Collect**	**Hours**	**Type**	**Visual**	**Label**			
	FOA	Multi	Binaural					T	P	N	C
Spatial LibriSpeech	✓	✓	✗	Simulated	650	Speech	-	✓	✓	✗	✗
YT-Ambigen	✓	✗	✗	Crawled	142	ALL	Video	✗	✗	✗	✗
BEWO-1M	✗	✗	✓	Craw+Sim	2800	ALL	Image	✗	✗	✓	✗
FAIR-Play	✗	✗	✓	Recorded	5.2	Music	Video	✗	✗	✗	✗
STARSS23	✓	✓	✗	Recorded	7.5	ALL	Video	✗	✓	✗	✓
BinauralMusic	✗	✗	✓	Crawled	15.2	Music	Video	✗	✗	✗	✓
RealMAN	✗	✓	✗	Recorded	228.2	Speech	Image	✓	✓	✗	✗
Sphere360	✓	✗	✗	Crawled	288	ALL	Video	✗	✗	✗	✗
MRSAudio (Ours)	✓	✗	✓	Recorded	484	ALL	Video	✓	✓	✓	✓

As shown in Table 1, existing datasets vary in their goals and modalities. For instance, Spatial LibriSpeech (Sarabia et al., 2023) simulates spatial speech using the LibriSpeech corpus and is mainly intended for binaural TTS applications. RealMAN (Yang et al., 2024) is a large-scale real-recorded and annotated 32-channel microphone array dataset for multichannel speech enhancement and source localization. YT-Ambigen (Kim et al., 2025) and Sphere360 (Liu et al., 2025) are derived from web-crawled video datasets, but lack explicit spatial annotations, making them suitable primarily for video-to-spatial audio generation. BinauralMusic focuses on musical content, offering instrument class tags. BEWO-1M (Sun et al., 2024) combines crawled content and simulated binaural rendering, and provides image or GPT-generated prompts. STARSS23 (Shimada et al., 2023) features real-world FOA and multi-channel audio alongside synchronized videos and includes sound class and position labels. In contrast to prior datasets, MRSAudio offers a comprehensive, large-scale, and real-world spatial audio corpus, featuring 484 hours of recorded data in both FOA and binaural formats, covering all audio domains including general audio, speech, singing, and music. Uniquely, MRSAudio includes synchronized audio, video, and 3D positional geometry, along with fine-grained cross-modal annotations, such as: transcripts, word and phoneme boundaries and music scores. These features make MRSAudio an ideal resource for a broad range of spatial audio generation and understanding tasks, including spatial speech, music, and singing voice synthesis.

 https://doi.org/10.52202/085713-4882

3 Dataset Description

In this section, we present **MRSAudio**, a freely available multimodal spatial audio corpus with synchronized video, positional data, and fine-grained annotations, released under the CC BY-NC-SA 4.0 license. Figure 2 illustrates our data processing pipeline, with detailed descriptions in subsequent subsections. We then summarize key statistics that demonstrate MRSAudio's scale and diversity.

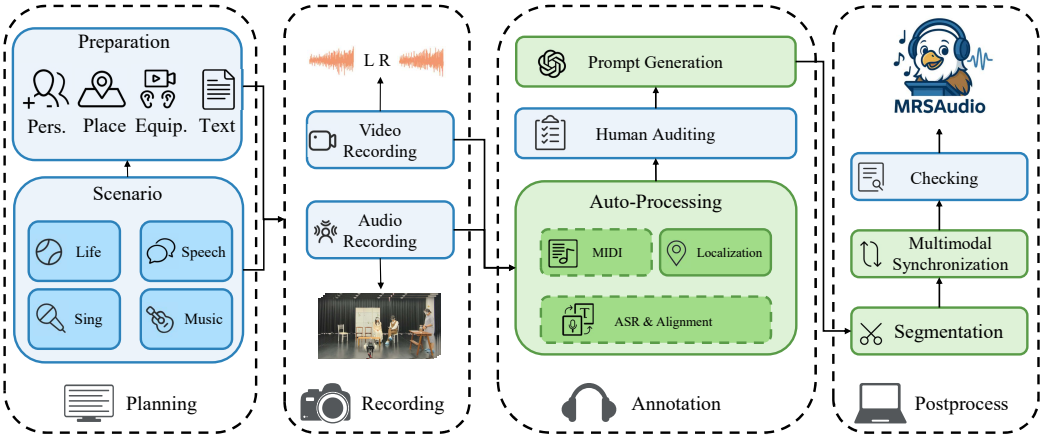


Figure 2: The pipeline of data collection and processing of MRSAudio. The blue boxes indicate steps requiring manual intervention, while the green boxes denote automated processing. In the "auto-processing" section, dashed modules apply to some scenarios, while solid modules apply to all.

3.1 Planning

To ensure that MRSAudio comprehensively covers scenarios from daily life, speech, singing, and music, we conduct systematic and modular planning before recording as follows.

MRSLife: This subset focuses on everyday conversations and environmental sound events. Based on the degree of human vocal interaction, MRSLife is further divided into two parts: MRSDialogue, which captures unscripted conversations that naturally include spontaneous action sounds (e.g., footsteps, door movements), and MRSSound, which focuses on non-verbal sound events primarily caused by physical activities such as cooking, typing, or sports.

MRSSpeech: MRSSpeech targets clean, high-quality conversational recordings for TTS task. All spoken interactions are recorded in controlled indoor environments with minimal noise. We invite speakers to participate in content-driven conversations based on predefined scripts.

MRSSing: MRSSing captures solo vocal performances for singing voice synthesis tasks. It includes professional solo vocal recordings in four languages: Mandarin, English, German, and French, performed by singers covering the full vocal range including soprano, alto, tenor, and bass.

MRSMusic: MRSMusic captures immersive instrumental performances suitable for spatial music generation and analysis. We record solo performances from 45 professional musicians across 23 instruments, including Traditional Chinese, Western and Electronic instruments. Each performance is paired with its corresponding musical score to support score-based music generation.

Details regarding personnel recruitment, venue selection, equipment configuration, and material preparation for each subset are provided in Appendix A.1.

3.2 Recording

Based on the predefined planning, we conduct parallel data collection across all modules. To ensure participant anonymity, masks are worn during recording when necessary. All participants sign an

open-source data release agreement, allowing the dataset to be freely distributed for academic research purposes. The recording details are summarized as follows:

MRSLife: In MRSDialogue, audio is recorded using a professional binaural recording head and high-resolution sound cards, while synchronized video is captured using industry-standard cameras. In MRSSound, in addition to binaural audio and exocentric video, we also captured FOA (Zoom H3-VR) and egocentric video (Gopro camera).To ensure the effectiveness of the egocentric video, participants are asked to remain within the frontal field of view of the binaural recording head.

MRSSpeech: To introduce spatial variability, we select four recording rooms that differ in size and acoustic material. Each speaker reads inflected emotions from the scripts while walking through the space, producing dynamic spatial cues. Recordings include spoken passages, binaural audio and exocentric video, as well as a clean mono audio by a lavalier microphone placed near the speaker.

MRSSing: The professional singers perform according to musical scores. To introduce variation in source–listener geometry, we adjust the position of the head-mounted binaural microphone relative to the singer. In addition to the binaural audio and synchronized video, each session includes a clean vocal track recorded with a studio-grade condenser microphone.

MRSMusic: We record 23 Traditional Chinese, Western and Electronic instruments, performed by 45 professional musicians.To capture rich spatial detail, we vary microphone placement around the instrument. Recordings include binaural audio, monaural audio, exocentric video, and synchronized video that records playing gestures. Full recording details are provided in Appendix A.2.

3.3 Annotation

To maximize MRSAudio's utility across a wide range of tasks, we begin with event-level annotations for all vocal and acoustic content in MRSLife, MRSSpeech, MRSSing, and MRSMusic. However, coarse annotations alone are insufficient for fine-grained tasks such as singing voice modeling and music generation from scores. To bridge this gap, we design a comprehensive annotation pipeline. Full implementation details and detailed annotation guidelines are provided in Appendix A.3.

MRSLife: For MRSDialogue, we apply WhisperX (Bain et al., 2023) for automatic speech recognition and speaker diarization to generate initial transcripts and speaker turns. Human annotators correct recognition errors and speaker attribution mismatches. The audio is then segmented into utterances and the transcripts are converted into phoneme sequences (using pypinyin for Mandarin). A two-stage alignment process follows: we first apply the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) for coarse word/phoneme mapping, then manually refine boundaries in Praat (Boersma, 2001). For MRSSound, we annotate sound event categories and corresponding time intervals.

MRSSpeech: Given the availability of full scripts, we adapt WhisperX for long-form word-to-audio alignment of up to 30 minutes. Each script line is automatically matched to its corresponding audio segment (see Appendix A.3 for more details). Annotators then review these alignments, correcting any omissions or insertions caused by actors' deviations from the script. Finally, phoneme sequences are extracted and aligned with the audio using the same procedure as in MRSDialogue.

MRSSing: We use voice activity detection (VAD) to segment recordings into singing regions, then align pre-existing lyrics using LyricFA's ASR-based dynamic programming algorithm[1]. Phoneme generation is language-dependent: pypinyin for Mandarin, ARPA for English, and MFA's built-in phoneme sets for French and German. Alignment is conducted via MFA, followed by manual refinement. Melody and rhythm of singing are transcribed into MIDI format using ROSVOT (Li et al., 2024). Annotators then label each excerpt with high-level style descriptors, such as emotional tone (happy, sad), tempo (slow, moderate, fast), and pitch range (low, medium, high).

MRSMusic: We use Audio Slicer[2] to segment the music recordings and generate initial symbolic annotations with basic-pitch (Bittner et al., 2022), and then employ professional musicians to verify and adjust note onsets, offsets, and dynamics to match the performance accurately.

Source Localization: For static sources, we manually record 3D positions relative to the capture space. For dynamic scenes, we use the Ultra-Wideband (UWB) system(Aiello & Rogerson, 2003)

[1] `https://github.com/wolfgitpr/LyricFA`

[2] `https://github.com/flutydeer/audio-slicer`

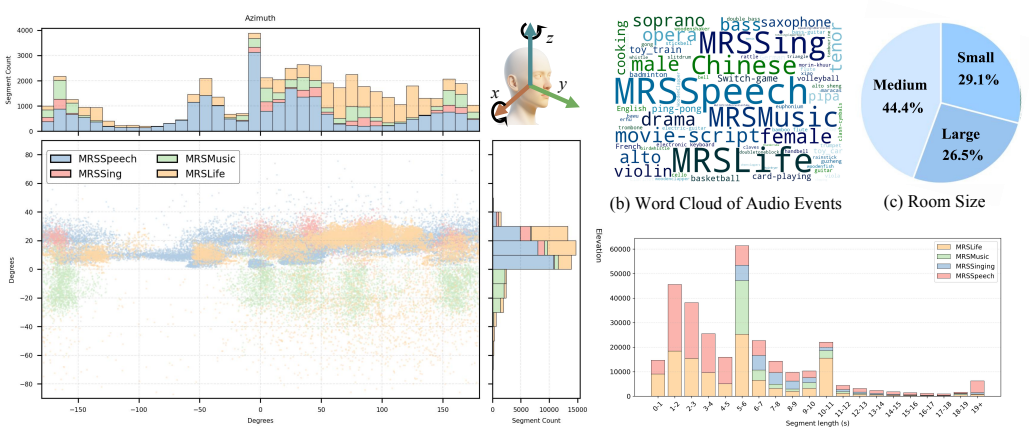
https://doi.org/10.52202/085713-4882

to track the positions of sound sources in real time. Based on the recorded position trajectories, we generate natural language motion descriptions using GPT-4o (Achiam et al., 2023).

3.4 Post-Processing

The post-processing pipeline comprises three key steps to refine raw annotated data. First, segmentation splits continuous recordings into task-oriented clips: utterances for speech and singing are extracted via alignment timestamps, while MRSSound audio is uniformly divided into 10-second segments. Next, multimodal synchronization aligns each clip with auxiliary modalities (text, video, position metadata) with temporal anchors. Static sound sources are annotated with manually measured 3D coordinates, whereas dynamic sources leverage interpolated UWB tracking trajectories. For scenes where participants' faces are visible, anonymization is performed by adding half-face masks. Finally, quality assurance involves domain experts auditing 15% of clips across all modules to verify temporal alignment precision, cross-modal content consistency, and annotation accuracy. Full auditing protocols are documented in Appendix A.4, ensuring reproducibility of this process.

3.5 Statistics



(a) Spatial Map of Human Audiotory Perception in MRSAudio

(b) Word Cloud of Audio Events

(c) Room Size

(d) Distribution of Clip Durations in MRSAudio

Figure 3: Statistical overview of MRSAudio. (a) Spatial distribution of sound sources relative to the listener. Red, green, and blue arrows denote the positive x, y, and z axes; azimuth is measured around the z-axis from the x-axis, and elevation is relative to the xy-plane. (b) Word cloud. (c) Proportions of recording spaces by room size. (d) Distribution of audio segment durations.

To illustrate spatial diversity, Figure 3(a) shows the 3D distribution of sound source positions relative to the listener. The heatmap reveals near-uniform azimuthal coverage, with greater density in the frontal hemisphere due to the prevalence of egocentric video recordings. Elevation angles are concentrated between –40° and 40°, aligning with typical human sound perception patterns. Figure 3(b) summarizes the annotations using a keyword cloud that captures the range of real-world activities recorded. Figure 3(c) presents the proportions of recording spaces by room size: medium-sized rooms are the most common, accounting for approximately 40% of sessions, while small and large rooms each represent around 30%. Recording duration is evenly distributed across room types, ensuring diverse acoustic coverage. Figure 3(d) shows the distribution of segment durations after automatic and manual segmentation. Most audio clips are shorter than 10 seconds, which is suitable for modeling short-duration events and supports efficient downstream training and inference. Overall, MRSAudio offers comprehensive coverage across spatial positions, acoustic environments, scene types, and temporal structures, making it well-suited for a wide range of spatial audio generation and understanding tasks. A more detailed breakdown of per-scenario statistics is provided in Appendix A.

https://doi.org/10.52202/085713-4882

4 Benchmarks

To demonstrate the quality and utility of MRSAudio in real-world scenarios, we evaluate it on five representative spatial audio tasks: (i) audio spatialization, (ii) spatial text-to-speech, (iii) spatial singing voice synthesis, (iv) spatial music generation, and (v) sound event localization and detection (SELD). These tasks cover both generation and understanding, and are critical for applications such as AR/VR, spatial media production, and perceptual scene analysis. All experiments are conducted using state-of-the-art methods on a server equipped with eight NVIDIA RTX 4090 GPUs. For different tasks, we employ distinct objective metrics. For generation tasks, we compute cosine similarity scores for direction (ANG Cos) and distance (DIS Cos) to quantify spatial alignment quality. Additionally, we utilize subjective MOS-Q (Mean Opinion Score for Quality) to evaluate the quality of generated audio and MOS-P (Mean Opinion Score for Position) to assess spatial perception. For implementation details of training and evaluation metrics, please refer to Appendix B.

4.1 Audio Spatialization

Audio spatialization aims to synthesize spatially immersive soundscapes from monaural inputs and source positional information. To evaluate MRSAudio on this task, we adopt BinauralGrad (Leng et al., 2022), a diffusion-based generation model that predicts binaural waveforms conditioned on monaural audio and source coordinates. Since the downstream generation tasks can be formulated as predicting interaural (binaural) representations from mono audio, we employ a single-stage training scheme for all experiments. For comparison, we include a traditional signal processing baseline (DSP), which renders binaural audio using virtual source positions simulated via room impulse responses (RIRs) and head-related transfer functions (HRTFs). We adopt the following objective metrics to evaluate audio quality: (1) W-L2: waveform L2 distance, (2) A-L2: amplitude envelope L2 distance, (3) P-L2: phase difference L2, (4) STFT: multi-resolution STFT loss, (5) PESQ: perceptual speech quality score. Results for the ground truth and DSP baseline are obtained by averaging across all test sets from the four MRSAudio subsets. Further details are provided in Appendix B.3.

Table 2: Audio Spatialization Performance. For MRSlife, we only use the MRSSound subset.

Method	**W-L2** $\times 10^{-3} \downarrow$	**A-L2** $\downarrow$	**P-L2** $\downarrow$	**PESQ** $\uparrow$	**STFT** $\downarrow$	**MOS-Q** $\uparrow$	**MOS-P** $\uparrow$
Ground Truth	–	–	–	–	–	4.69 ± 0.08	4.56 ± 0.10
DSP	1.691	0.048	1.562	**2.830**	**1.246**	3.89 ± 0.09	3.75 ± 0.11
MRSLife	0.076	0.025	0.898	-	2.243	3.91 ± 0.07	3.87 ± 0.10
MRSSpeech	0.460	0.061	0.807	1.929	2.352	3.88 ± 0.08	3.84 ± 0.08
MRSSing	0.647	0.093	1.004	1.723	2.539	3.84 ± 0.09	3.91 ± 0.07
MRSMusic	0.705	0.063	0.835	-	1.724	3.87 ± 0.07	3.93 ± 0.09
ALL	**0.305**	**0.041**	**0.801**	2.352	1.681	$\mathbf{3.93 \pm 0.09}$	$\mathbf{3.94 \pm 0.07}$

As shown in Table 2, BinauralGrad achieves strong performance across all MRSAudio subsets, surpassing the classical DSP baseline on most objective and subjective metrics. The highest performance is observed on MRSLife, likely due to the relatively limited variation in sound sources within these scenes. These results demonstrate that MRSAudio's rich spatial annotations and diverse acoustic environments provide a solid foundation for training and evaluating spatial audio generation models.

4.2 Spatial Text to Speech

Spatial TTS aims to produce high-quality speech enriched with spatial cues, thereby enhancing immersion and realism in AR/VR. Although recent advances in TTS have led to impressive improvements in speech quality, progress in spatialized speech generation remains limited due to the scarcity of spatially annotated recordings with rich, high-quality labels. To evaluate the effectiveness of MRSAudio for this task, we train an end-to-end model following ISDrama (Zhang et al., 2025) to directly generate spatial speech from text and position data. Additionally, we compare with cascaded pipelines that combine monaural TTS models (CosyVoice (Du et al., 2024) and F5-TTS (Chen et al., 2024)) with the audio spatialization module. This allows us to assess both speech generation quality and the spatial fidelity enabled by our dataset. For content evaluation, we report Character Error Rate (CER) and Speaker Similarity (SIM). Further details are provided in Appendix B.4.

https://doi.org/10.52202/085713-4882

Table 3: Spatial TTS Performance on MRSSpeech. "SP" denotes the Audio Spatialization.

Method	Objective				Subjective	
	CER ↓	SIM ↑	ANG Cos ↑	DIS Cos ↑	MOS-Q ↑	MOS-P ↑
Ground Truth	**2.54%**	–	–	–	4.39 ± 0.08	4.16 ± 0.10
Mono + SP	2.56%	**0.98**	0.44	**0.68**	$\mathbf{3.88 \pm 0.08}$	$\mathbf{3.84 \pm 0.08}$
CosyVoice + SP	3.89%	0.96	0.41	0.63	3.75 ± 0.12	3.72 ± 0.09
F5-TTS + SP	3.15%	0.97	0.40	0.62	3.69 ± 0.13	3.67 ± 0.14
ISDrama (speech)	3.35%	0.96	**0.48**	0.65	3.85 ± 0.09	3.82 ± 0.11

As shown in Table 3, the Mono+SP method achieves strong performance across most metrics. The CER remains low and comparable to the ground truth, indicating preserved linguistic accuracy after spatialization. A high SIM score reflects stable timbre learning. ANG Cos and DIS Cos show good spatial alignment with the ground truth, and subjective MOS scores confirm that the generated speech is both natural and spatially coherent. These results demonstrate that MRSSpeech provides high-quality, spatially annotated training data that enables effective and realistic spatial TTS.

4.3 Spatial Singing Voice Synthesis

Spatial SVS aims to produce expressive, high-quality singing voices enriched with accurate spatial cues, thereby enhancing listener immersion. While traditional SVS has advanced considerably, spatial SVS remains underexplored due to the lack of high-quality, spatially annotated datasets. To evaluate MRSAudio for this task, we use the MRSSing subset to train and benchmark models. We adopt the ISDrama architecture to perform end-to-end spatial singing synthesis, incorporating note-level pitch control to enhance prosody accuracy. For comparison, we use the open-source SVS models Rmssinger (He et al., 2023). The outputs are then spatialized with BinauralGrad. We employ objective metrics, including Mel-Cepstral Distortion (MCD) and F0 Frame Error (FFE), to evaluate spectral and pitch similarity between predicted and ground-truth. Details are in Appendix B.5.

Table 4: Spatial SVS Performance on MRSSing. "SP" denotes the Audio Spatialization.

Method	Objective				Subjective	
	MCD ↓	FFE ↓	ANG Cos ↑	DIS Cos ↑	MOS-Q ↑	MOS-P ↑
GT (Ground Truth)	-	-	-	-	4.45 ± 0.10	4.30 ± 0.12
Mono+SP	**3.19**	**0.17**	**0.51**	**0.71**	3.84 ± 0.09	$\mathbf{3.91 \pm 0.07}$
Rmssinger+SP	3.85	0.23	0.45	0.65	3.65 ± 0.08	3.81 ± 0.11
ISDrama(sing)	3.71	0.21	0.47	0.70	$\mathbf{3.86 \pm 0.13}$	3.88 ± 0.09

As shown in Table 4, the Mono + SP approach trained on MRSSing with pitch control achieves the best performance across most metrics. Its low MCD indicates strong spectral fidelity, and high ANG Cos and DIS Cos scores demonstrate effective spatial alignment. Subjective MOS results confirm that the generated singing voices are natural, high-quality, and spatially coherent. These findings validate MRSSing as an effective resource for spatial singing voice generation.

4.4 Spatial Music Generation

This task aims to synthesize spatially immersive music conditioned on symbolic scores. While datasets like FAIR-Play offer high-quality instrument recordings, they lack aligned sheet music, limiting controllable generation. In contrast, our MRSMusic subset includes 23 instruments with aligned scores, enabling fine-grained, score-based spatial music synthesis. We benchmark three systems on MRSMusic. First, we apply BinauralGrad (Leng et al., 2022) to spatialize mono recordings. Second, we utilize Make-An-Audio 2 (Copet et al., 2023) to generate mono music from MIDI scores, which is subsequently spatialized. Third, we adapt ISDrama to accept both score embeddings and spatial poses, enabling end-to-end spatial symbolic music generation. For objective evaluation, we compute Fréchet Audio Distance (FAD) and FFE to evaluate the results. Details are in Appendix B.6.

Table 5: Spatial MG Performance on MRSMusic. "SP" denotes the Audio Spatialization.

Method	Objective				Subjective	
	FAD ↓	FFE ↓	ANG Cos ↑	DIS Cos ↑	MOS-Q ↑	MOS-P ↑
GT (Ground Truth)	–	–	–	–	4.49 ± 0.09	4.34 ± 0.11
Mono+SP	2.88	**0.14**	0.48	**0.74**	3.87 ± 0.07	$\mathbf{3.93 \pm 0.09}$
Make-An-Audio 2+SP	4.39	0.49	0.49	0.41	3.74 ± 0.10	3.73 ± 0.13
ISDrama(music)	**2.45**	0.21	**0.53**	0.68	$\mathbf{3.89 \pm 0.12}$	3.88 ± 0.10

As shown in Table 5, the Mono + SP pipeline achieves strong performance, benefiting from access to ground truth mono audio. The Make-An-Audio 2 + SP exhibits limitations in FAD and pitch accuracy. This suggests that general-purpose audio generation models may struggle to fully capture structured musical information from symbolic prompts. The ISDrama (Music) model generates music directly from the symbolic inputs and spatial cues, achieving better coherence and spatial alignment than the Make-An-Audio 2 + SP. These results highlight MRSMusic's effectiveness in supporting spatially controllable music generation across diverse instruments and spatial conditions.

4.5 Sound Event Localization and Detection

This task evaluates the ability to detect and localize sound events using MRSAudio's spatial annotations. We follow STARSS23 (Shimada et al., 2023) using both audio-only and audio-visual variants on the MRSSound subset. In the audio-only condition, models receive either FOA or binaural waveforms as input and predict sound event classes along with 3D source coordinates. For the audio-visual condition, we extract bounding boxes of visible persons to serve as coarse visual priors, which are fused with the audio representations. We also explore architectural variations by replacing the original convolutional backbone with a Transformer encoder, allowing us to assess the impact of temporal modeling capacity on spatial prediction. We evaluate model performance using four standard joint detection and localization metrics, including location-aware detection (F_{20°, ER_{20°) and class-aware localization (LE_{CD}, LR_{CD}). Details are in Appendix B.7.

Table 6: Sound Event Localization and Detection on MRSSound.

Model	Audio Type	Visual	ER_{20° ↓	F_{20° ↑	LE_{CD} ↓	LR_{CD} ↑
ConvNet	FOA	✓	1.17 ± 0.02	$\mathbf{13.00 \pm 2.72}$	42.44 ± 8.91	82.76 ± 5.19
ConvNet	FOA	✗	1.12 ± 0.02	9.33 ± 0.36	46.47 ± 4.46	85.35 ± 3.80
ConvNet	Binaural	✓	1.11 ± 0.03	5.90 ± 3.86	41.95 ± 9.85	45.81 ± 11.29
ConvNet	Binaural	✗	1.11 ± 0.02	6.00 ± 0.34	45.17 ± 7.92	70.14 ± 10.38
Transformer	FOA	✓	1.01 ± 0.01	7.52 ± 0.42	$\mathbf{35.69 \pm 7.47}$	46.78 ± 7.62
Transformer	FOA	✗	$\mathbf{0.99 \pm 0.05}$	7.95 ± 0.15	48.76 ± 2.66	$\mathbf{87.32 \pm 2.14}$
Transformer	Binaural	✓	1.01 ± 0.02	8.47 ± 0.65	36.18 ± 7.29	41.33 ± 12.73
Transformer	Binaural	✗	1.11 ± 0.04	7.37 ± 0.97	46.74 ± 7.91	43.87 ± 10.74

As shown in Table 6, model performance varies with architecture, input modality, and audio format. Transformer-based models generally outperform ConvNet baselines, particularly in reducing localization error. FOA input consistently yields better results than binaural audio, benefiting from richer spatial representation. For example, the Transformer with FOA and no visual input achieves the lowest error rate (ER_{20° 0.99) and highest localization recall (LR_{CD} 87.32). The addition of visual features improves performance in some ConvNet settings (e.g., F_{20° rises from 9.33 to 13.00 with FOA), but offers limited or inconsistent gains in Transformer models, possibly due to modality mismatch or redundancy. These results highlight the value of MRSAudio's spatial annotations and multimodal streams in supporting flexible evaluation under both unimodal and multimodal configurations.

5 Conclusion and Discussion

We introduce MRSAudio, a large-scale, multimodal spatial audio corpus designed to support a wide range of generation and understanding tasks. MRSAudio comprises four complementary modules, MRSLife, MRSSpeech , MRSSing, and MRSMusic, each captured with binaural/FOA

 https://doi.org/10.52202/085713-4882

audio, synchronized video, precise 3D pose metadata, and richly detailed annotations (event labels, transcripts, phoneme boundaries, lyrics, musical scores, and motion prompts). Through extensive benchmarks on audio spatialization, binaural speech, singing, and music generation , as well as sound event localization and detection, we demonstrate that MRSAudio's scale, diversity, and annotation depth enable state-of-the-art performance and unlock new avenues for spatial audio research.

Limitations and Future Directions: While MRSAudio offers broad multimodal coverage, two limitations remain. First, although synchronized video is provided for all recordings, current benchmarks only explore visual input in a limited subset of tasks. Second, while the dataset includes both binaural and First-Order Ambisonic (FOA) formats, the FOA subset is smaller in scale, and most tasks focus on binaural audio, limiting spatial modeling diversity. Future work will expand the role of visual modalities in tasks such as sound localization and scene understanding. We also plan to increase FOA recordings to balance data distribution and support broader spatial audio research, and develop more FOA-specific benchmarks to better utilize ambisonic spatial cues.

Negative Societal Impact: As with any large-scale audiovisual dataset, MRSAudio carries potential risks if misused. It could be exploited to generate highly realistic yet synthetic spatial audio for deepfakes or disinformation in AR and VR applications. To mitigate such risks, MRSAudio is released under a noncommercial license with clear usage guidelines. We encourage responsible use in accordance with ethical standards, including consent management, data governance, and transparency.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key R&D Program of China (2022ZD0162000), the National Natural Science Foundation of China under Grant No. 62222211, and the National Natural Science Foundation of China under Grant No. U24A20326.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Adavanne, S., Politis, A., and Virtanen, T. A multi-room reverberant dataset for sound event localization and detection. *Proc. DCASE2019*, 2019.

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Aiello, G. R. and Rogerson, G. D. Ultra-wideband wireless systems. *IEEE microwave magazine*, 4 (2):36–47, 2003.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020.

Bain, M., Huh, J., Han, T., and Zisserman, A. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

Bittner, R. M., Bosch, J. J., Rubinstein, D., Meseguer-Brocal, G., and Ewert, S. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, 2022.

Boersma, P. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., and Chen, X. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Cohen, Y. E. and Knudsen, E. I. Maps versus clusters: different representations of auditory space in the midbrain and forebrain. *Trends in neurosciences*, 22(3):128–135, 1999.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Gao, R. and Grauman, K. 2.5d visual sound. *arXiv preprint arXiv:1812.04204*, 2019.

Gao, Z., Li, Z., Wang, J., Luo, H., Shi, X., Chen, M., Li, Y., Zuo, L., Du, Z., Xiao, Z., et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

Grothe, B., Pecka, M., and McAlpine, D. Mechanisms of sound localization in mammals. *Physiological reviews*, 90(3):983–1012, 2010.

He, J., Liu, J., Ye, Z., Huang, R., Cui, C., Liu, H., and Zhao, Z. Rmssinger: Realistic-music-score based singing voice synthesis. *arXiv preprint arXiv:2305.10686*, 2023.

 https://doi.org/10.52202/085713-4882

Huang, J., Ren, Y., Huang, R., Yang, D., Ye, Z., Zhang, C., Liu, J., Yin, X., Ma, Z., and Zhao, Z. Make-an-audio 2: Temporal-enhanced text-to-audio generation, 2023a.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., et al. AudioGPT: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023b.

Huiyu, X., Shuaifan, J., Zhibo, W., Zhongjie, B., and Tao, W. Psa-nerf: Personalized spatial attention neural rendering for audio-driven talking portraits generation. *Chinese Journal of Electronics*, 2025.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Kim, J., Yun, H., and Kim, G. Visage: Video-to-spatial audio generation. In *ICLR*, 2025.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Leng, Y., Chen, Z., Guo, J., Liu, H., Chen, J., Tan, X., Mandic, D., He, L., Li, X., Qin, T., et al. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. *Advances in Neural Information Processing Systems*, 35:23689–23700, 2022.

Li, R., Zhang, Y., Wang, Y., Hong, Z., Huang, R., and Zhao, Z. Robust singing voice transcription serves synthesis, 2024.

Liu, H., Luo, T., Jiang, Q., Luo, K., Sun, P., Wan, J., Huang, R., Chen, Q., Wang, W., Li, X., Zhang, S., Yan, Z., Zhao, Z., and Xue, W. Omniaudio: Generating spatial audio from 360-degree video, 2025. URL `https://arxiv.org/abs/2504.14906`.

Liu, Z., Rosen, E., et al. Autoslicer: Scalable automated data slicing for ml model analysis. *arXiv preprint arXiv:2212.09032*, 2022.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pp. 498–502, 2017.

Pedro Morgado, Nuno Vasconcelos, T. L. and Wang, O. Self-supervised generation of spatial audio for 360deg video. In *Neural Information Processing Systems (NIPS)*, 2018.

Sarabia, M., Menyaylenko, E., Toso, A., Seto, S., Aldeneh, Z., Pirhosseinloo, S., Zappella, L., Theobald, B.-J., Apostoloff, N., and Sheaffer, J. Spatial librispeech: An augmented dataset for spatial audio learning. *arXiv preprint arXiv:2308.09514*, 2023.

Shimada, K., Politis, A., Sudarsanam, P., Krause, D. A., Uchida, K., Adavanne, S., Hakala, A., Koyama, Y., Takahashi, N., Takahashi, S., et al. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in neural information processing systems*, 36:72931–72957, 2023.

Sun, P., Cheng, S., Li, X., Ye, Z., Liu, H., Zhang, H., Xue, W., and Guo, Y. Both ears wide open: Towards language-driven spatial audio generation. *arXiv preprint arXiv:2410.10676*, 2024.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., et al. SALMONN: Towards generic hearing abilities for large language models. *Proc. ICLR*, 2024.

Wang, Q., Chai, L., Wu, H., Nian, Z., Niu, S., Zheng, S., Wang, Y., Sun, L., Fang, Y., Pan, J., et al. The nerc-slip system for sound event localization and detection of dcase2022 challenge. *DCASE2022 Challenge, Tech. Rep.*, 2022.

Wang, Y., Guo, W., Huang, R., Huang, J., Wang, Z., You, F., Li, R., and Zhao, Z. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv e-prints*, pp. arXiv–2406, 2024.

Wei, H., Cao, X., Dan, T., and Chen, Y. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. *arXiv preprint arXiv:2306.15412*, 2023.

Xie, B. Spatial sound-history, principle, progress and challenge. *Chinese Journal of Electronics*, 29 (3):397–416, 2020.

Yang, B., Quan, C., Wang, Y., Wang, P., Yang, Y., Fang, Y., Shao, N., Bu, H., Xu, X., and Li, X. Realman: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization. *Advances in Neural Information Processing Systems*, 37:105997–106019, 2024.

Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Wu, X., et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.

Yost, W. A. Spatial hearing: the psychophysics of human sound localization. *Ear and Hearing*, 19 (2):167, 1998.

Zhang, Y., Guo, W., Pan, C., Zhu, Z., Jin, T., and Zhao, Z. Isdrama: Immersive spatial drama generation through multimodal prompting. *arXiv preprint arXiv:2504.20630*, 2025.

Zheng, Z., Peng, P., Ma, Z., Chen, X., Choi, E., and Harwath, D. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*, 2024.

 https://doi.org/10.52202/085713-4882

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As shown in Section 1, this paper proposes a large-scale multimodal recorded spatial audio dataset along with refined annotation for spatial audio-related tasks.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations and the negative societal impact in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our dataset is available at the `https://huggingface.co/datasets/verstar/MRSAudio`. The source code associated with our work can be accessed in the GitHub repository: `https://github.com/MRSAudio/MRSAudio_Main` and in the supplementary materials. Additionally, training details is shown in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/085713-4882

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our dataset is available at the `https://huggingface.co/datasets/verstar/MRSAudio`. The source code associated with our work can be accessed in the GitHub repository: `https://github.com/MRSAudio/MRSAudio_Main` and in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The settings and details of experiments is shown in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report confidence intervals of subjective metric results in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The required experimental resources are detailed in Section 4 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: 1) The wage details for participants are reported in Appendix A.2, A.3, and B.1. 2) All data were de-identified prior to release, as detailed in Section 3. 3) As described in Appendix A.2, all participants involved in data recording signed consent forms, agreeing to the publication of the dataset under the CC BY-NC-SA 4.0 license. 4) The dataset we have released adheres to the CC BY-NC-SA 4.0 license, and all baselines used in the paper comply with their respective licenses.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

https://doi.org/10.52202/085713-4882

Justification: We discuss the societal impacts from both positive and negative perspectives in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All data were de-identified prior to release, as detailed in Section 3 and AppendixA.4. For participants whose faces were visible, we anonymized the recordings by adding masks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use all models and codes under their Licenses (e.g. CC 4.0, MIT, etc).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed description in Section 3 and Appendix A for the proposed dataset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix B.1 includes the instructions and screenshots of rating systems. The wage for participants is mentioned in Appendix B.1 as well.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    All participants signed the consent form before the recording.

 https://doi.org/10.52202/085713-4882

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components. 1) About LLM: this work only uses pretrained language models to generate natural prompts based on the given trajectory as shown in Section 3; 2) We only use image generators to assist the design of the paper's header figure.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Details of Dataset

A.1 Details of Planning

During the planning phase, we systematically analyze real-world application scenarios for spatial audio and divide them into four major categories. MRSLife focuses on daily environmental sound events, MRSSpeech targets high-quality conversational data, MRSSing captures solo vocal performances, and MRSMusic records immersive instrumental music. These four scenarios collectively cover a broad range of everyday acoustic contexts and are designed to support a wide spectrum of downstream tasks.

Based on predefined scenario requirements, we recruit professionals from relevant fields to participate in the recording process. To ensure spatial diversity, we rent various venues tailored to each scenario type. Following the FAIR-Play protocol, we use a 3Dio Free Space XLR binaural microphone to capture spatial audio, GoPro HERO cameras and mobile phones to record exocentric and egocentric videos, and UWB-based tracking systems to capture motion trajectories. Once personnel, locations, and equipment are secured, we prepare task-specific materials for each subset. For MRSLife, we further divide the content based on the proportion of speech involved, resulting in two subcategories: MRSDialogue (e.g., group games, board games) and MRSSound (e.g., kungfu, kitchens, offices, sports). We predefine common sound events for each, such as clattering dishes, keyboard typing, and table tennis. For MRSSpeech, we compile a large corpus of scripts from movies, TV shows, and crosstalk performances and automatically extract dialogue passages for speaker delivery. For MRSSing, we design lyrics in four languages (Chinese, English, German, and French) and recruit singers across a range of vocal types to maximize diversity. For MRSMusic, we collect solo performances across 23 traditional and modern instruments, covering a wide array of timbres and playing techniques. Specific sound categories for each subset are summarized in Table 7.

Table 7: Examples of predefined content types for each MRSAudio subset.

Subset	Samples of Categories or Keywords
MRSLife	**MRSDialogue**: board games, card games, collaborative tasks **MRSSound**: kungfu, office, maracas, typing, whistle, gong
MRSSpeech	Movie scripts, crosstalk, scripted TV dialogue, multi-speaker conversations
MRSSing	Chinese, English, German, French; soprano, alto, tenor, bass
MRSMusic	violin, electronic keyboard, cello, viola, double bass, trumpet, trombone, euphonium, erhu, pipa, xiao, bawu, trumpet, trombone, and others (23 instruments in total)

A.2 Details of Recording

We recruit a large number of participants with professional backgrounds in singing, music, and language to contribute to the recording process. To protect their identity, participants are asked to wear masks in appropriate scenarios. Prior to participation, all individuals sign consent forms agreeing to the open-source release of their audio and video under the CC BY-NC-SA 4.0 license.

All audio is recorded in WAV format at a sampling rate of 48 kHz. Video is recorded at a minimum resolution of 1080p and 24 frames per second, and is later standardized to this format during post-processing.

MRSLife: We recruit 62 participants to perform daily activities including board games, cooking, exercise, and office work. In MRSDialogue, each participant is compensated $30 per recorded hour. Binaural audio is captured using head-mounted microphones, and third-person video is recorded. In MRSSound scenes such as kung fu or kitchen demonstrations, performers are paid $50 per hour. Both binaural and FOA (First-Order Ambisonics) recordings are collected, along with first-person and third-person video. The binaural dummy head is co-located with the egocentric camera and the Zoom H3-VR (Ambisonic recorder). The egocentric camera is rigidly aligned with the head's gaze to capture first-person visuals, while the Ambisonic recorder is mounted 7.5 cm above the head. An exocentric camera is placed at a surveyed position in the scene to provide a third-person view of the environment and object relationships. Participants receive brief scene descriptions and are asked to

 https://doi.org/10.52202/085713-4882

act naturally while maintaining a single active sound source when possible. The total duration for MRSLife recordings reaches approximately 150 hours.

MRSSpeech: We employ 44 expressive speakers to read from scripted texts, each paid $30 per hour of recorded audio. The total recorded duration reaches 200 hours, with compensation totaling $40,000. During recording sessions, speakers alternate between standing and walking around the recording area to introduce spatial diversity while maintaining speech clarity. Binaural audio is captured using head-mounted microphones, and clean monaural speech is recorded using lavalier microphones. All sessions are filmed from a third-person perspective.

MRSSing: Eighteen professional singers participate in the recordings. Each singer is fluent in at least one of the following languages: Chinese, English, German, or French, and collectively they cover all vocal ranges including soprano, alto, tenor, and bass. Performers are paid $50 per hour, contributing a total of 80 hours of audio. Singing is recorded at a fixed position to ensure spatial consistency. Binaural recordings are captured with head-mounted microphones, and monaural audio is recorded with a studio-grade microphone. Third-person panoramic video is also captured.

MRSMusic: We engage 45 instrumentalists performing 23 Traditional Chinese, Western and Electronic instruments such as violin, erhu, pipa, electric guitar, and keyboard. Each performer receives $60 per recorded hour. A total of 69 hours of solo music performances are recorded. Audio is captured using both head-mounted binaural microphones and reference monaural microphones. Third-person video provides a full-scene view, while first-person cameras are used to capture detailed playing gestures.

A.3 Details of Annotation

We employ a team of domain experts in singing, music performance, and linguistics to carry out and review all annotations, compensating each annotator at a rate of $15 per hour. Prior to beginning their work, every expert receives a clear explanation of how the annotations will be used and agrees to release their annotation results under an open-source license for academic research. For all modules, we first synchronize each audio with its corresponding video, mono-channel reference track, and 3D positional metadata. Annotators then verify and correct this synchronization to ensure perfect alignment across modalities.

MRSLife. In MRSDialogue scenes, we automatically generate initial transcripts and speaker clusters with WhisperX, extracting word-level timestamps and speaker IDs. Experts then load these results in Praat and assign each cluster to the correct speaker, correcting transcription errors as needed. Next, we run Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to produce coarse phoneme-to-audio alignments (exported in TextGrid format), using pypinyin to convert Chinese text into phoneme sequences.[3] Finally, annotators refine word and phoneme boundaries in Praat to achieve millisecond-level precision. In MRSSound segments, annotators additionally label each time interval with the corresponding event category (e.g., "clattering," "typing," "pages turning").

MRSSpeech. Given full dialogue scripts, we perform automatic alignment to 30-minute recordings using a chunk-based extension of WhisperX. This method divides each utterance-level audio segment into fixed-length chunks (e.g., 30 seconds), applies phoneme-level models (e.g., wav2vec 2.0 (Baevski et al., 2020)) for emission prediction on each chunk, and then concatenates emissions to form a complete alignment matrix. We compute alignments via a trellis-based dynamic programming algorithm with backtracking, followed by scaling to restore absolute timestamps. Word- and sentence-level timings are derived by grouping aligned characters using word indices and sentence tokenization. Missing or partial timings are interpolated using a specified method (e.g., nearest). This pipeline enables efficient and accurate alignment of long-form recordings on GPU. We then refine sentence boundaries in Praat and apply the same MFA-plus-Praat workflow used in MRSLife for fine-grained phoneme and word-level alignment.

MRSSing. Each segment contains a solo vocal performance with known lyrics. We first apply voice activity detection (VAD) to segment the recordings. Lyrics are aligned to each segment using LyricFN's ASR-based dynamic programming method. English phonemes follow the ARPABET

[3] `https://github.com/mozillazg/python-pinyin`

standard,[4] while German and French use MFA's phoneme sets.[5] We then apply MFA for initial alignment and refine it manually in Praat to obtain precise word and phoneme boundaries. Annotators assign fine-grained style tags, including emotion, genre, and tempo. To generate score annotations, we extract F0 contours using RMVPE (Wei et al., 2023) and convert them into MIDI format via ROSVOT (Li et al., 2024), followed by expert correction.

MRSMusic. We segment the recordings using AutoSlicer (Liu et al., 2022) and generate preliminary symbolic annotations using basic-pitch (Bittner et al., 2022). Professional musicians then verify and refine the annotations, adjusting note onsets, offsets, dynamics, and articulations to ensure consistency between the score and performance.

A.4 Details of Post-Processing

Segmentation: After annotation, we segment the raw recordings into shorter clips to support spatial audio generation and analysis tasks. For speech and singing, we use alignment timestamps to extract utterances. For MRSSound, each recording is divided into fixed 10-second segments.

Multimodal Synchronization: In addition to binaural audio, each clip is synchronized with the following modalities: audio, text, video, and position data. Using the segment timestamps, we align all streams temporally. For static sources, we attach manually recorded 3D coordinates; for dynamic sources, we interpolate Ultra-Wideband (UWB) tracking data over the segment duration. This process yields fully synchronized multimodal clips ready for downstream spatial audio tasks. Furthermore, for segments where participants' faces are visible, we apply anonymization by using a face detection model to overlay digital masks during post-processing.

Checking: To ensure the reliability of annotations, domain experts perform a random audit on 15% of the segmented clips across all four modules. The audit process involves the following steps: (1) Verifying the temporal synchronization across different modal; (2) Confirming that the assigned event labels accurately reflect the audiovisual content present in each clip; (3) Reviewing the speech segments to check the correctness of word- and phoneme-level alignments; (4) Evaluating singing clips for accurate alignment between lyrics and audio, consistency with musical scores, and correctness of assigned style labels; (5) Assessing MRSMusic excerpts to verify that musical properties, such as key, pitch, and note duration.

A.5 Statistics of MRSAudio

A.5.1 Statistics of MRSLife

MRSDialogue. Figure 4(a) presents the 3D spatial distribution of sound sources in MRSDialogue. In this subset, which features frequent human conversations, most sources are located around the ear-level height of the listener. The azimuthal distribution covers nearly all directions surrounding the listener, offering diverse angular data for training spatial localization models with strong generalization.

Figure 4(b) shows the phoneme distribution across all speech segments. The most common phoneme is 'i', while the least frequent is 'ueng'. This distribution aligns with real-world phonetic patterns and highlights the dataset's linguistic richness. Such broad phoneme coverage is beneficial for downstream tasks in speech synthesis and recognition under spatial settings.

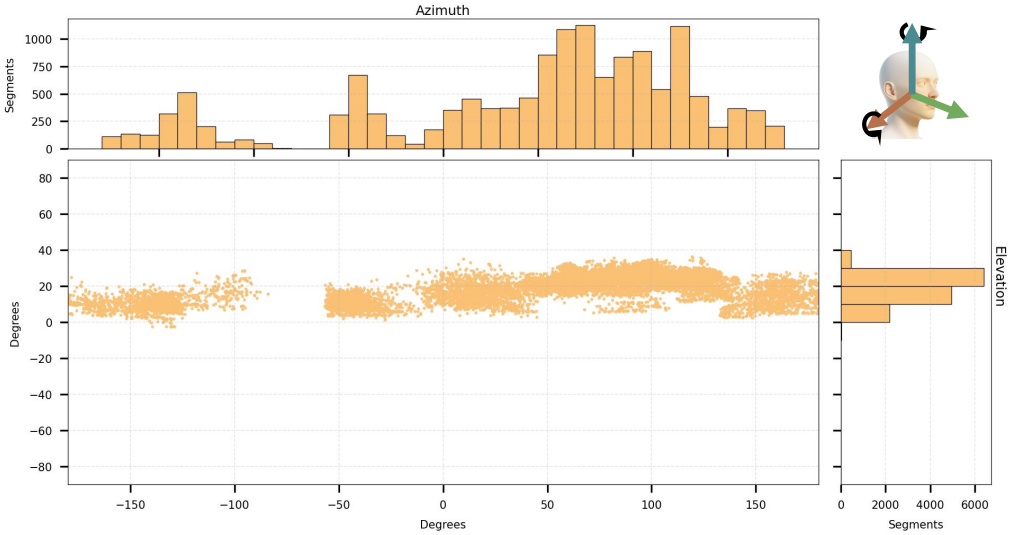
MRSSound. Figure 5(a) illustrates the spatial distribution of sound sources in MRSSound relative to the listener's head-centered coordinate system. The listener's facing direction is at 90 degrees, and most sound events are concentrated in the frontal hemisphere (azimuth from 0° to 180°), which is consistent with the first-person video capture setup. Some events also appear in the rear field (–180° to 0°). In elevation, the majority of sound sources are distributed between –90° and 40°, realistically reflecting everyday human perception in standing scenarios, where sound events typically originate below ear level. This broad spatial coverage supports the training of spatial audio models with strong generalization ability.

Figure 5(b) presents the duration distribution of recorded sound events. MRSSound covers a wide variety of everyday scenarios, including cooking in kitchens, working in office environments, and

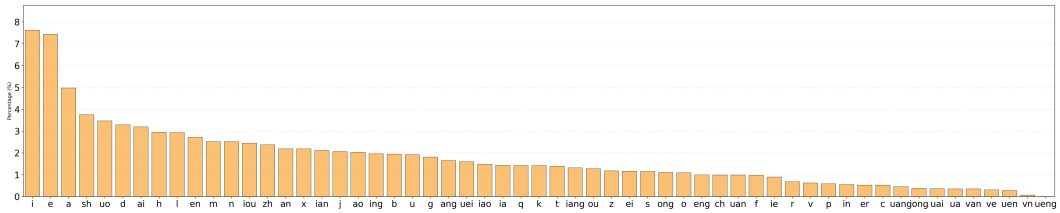
[4] https://en.wikipedia.org/wiki/ARPABET

[5] https://mfa-models.readthedocs.io/en/latest/dictionary/

 https://doi.org/10.52202/085713-4882



(a) Spatial Distribution of Sound Sources Relative to the Listener



(b) Distribution of Phones in MRSDialogue

Figure 4: Statistical overview of MRSDialogue. (a) Spatial distribution of sound sources relative to the listener. Red, green, and blue arrows denote the positive x, y, and z axes; azimuth is measured around the z-axis from the x-axis, and elevation is relative to the xy-plane. (b) Distribution of phones in MRSDialogue.

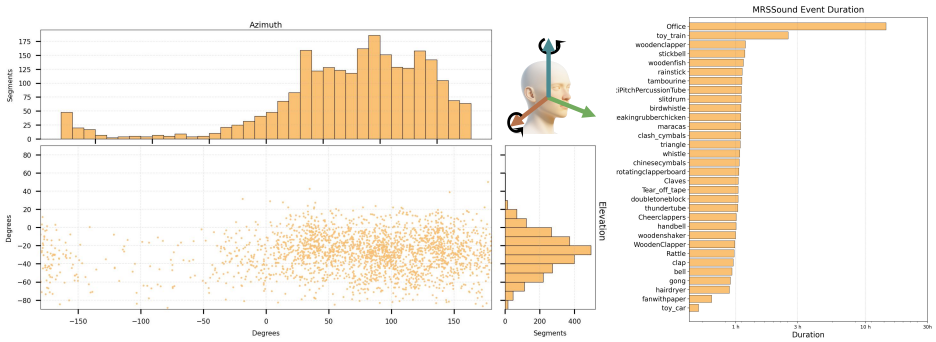
sports-related activities. The diversity of event types and durations makes the subset suitable for training and evaluating models in real-world spatial sound event understanding.

A.5.2 Statistics of MRSSpeech

Figure 6(a) illustrates the spatial distribution of speech sources with respect to the listener. The azimuth angles span the full 360° around the listener, providing comprehensive coverage of spatial directions. Elevation angles are mostly concentrated between 0° and 60°. Smaller elevations reflect scenarios where both the speaker and listener are either standing or seated, while larger elevations (above 30°) simulate common real-world speech situations such as meetings, where a standing speaker addresses a seated listener. This diverse spatial coverage supports generalization in spatial speech modeling.

Figure 6(b) shows a word cloud representing the diversity of dialogue content. The transcripts are sourced from theatrical scripts, films, and other spoken-only scenarios, capturing a wide range of expressive and stylistic variation. Figure 6(c) presents the distribution of room sizes used for speech recordings. Most multi-speaker interactions take place in medium to large rooms, such as meeting or lecture spaces. We include three distinct environments with varying absorption properties and dimensions to simulate different acoustic conditions.

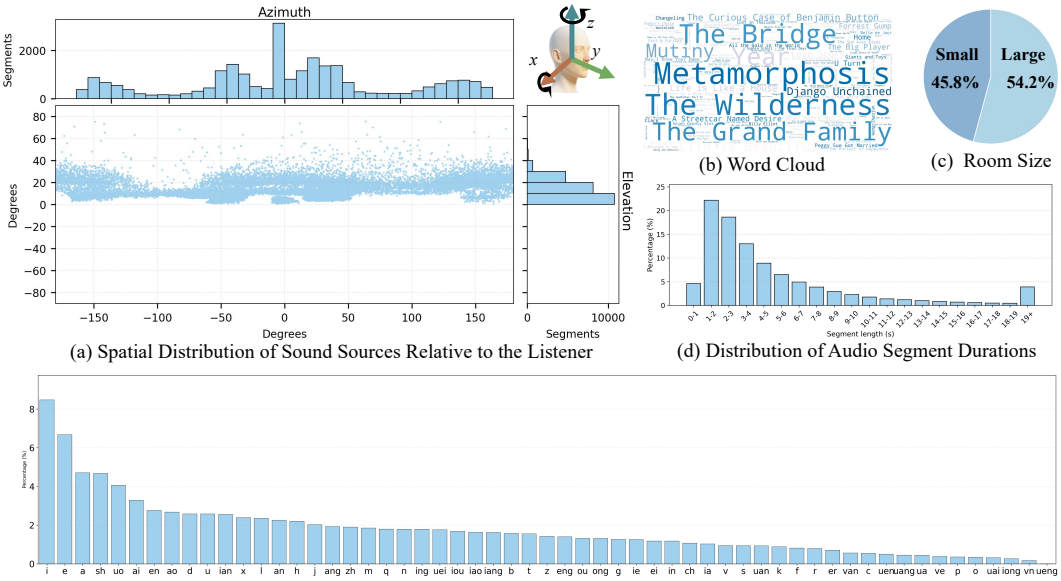
Figure 6(d) displays the distribution of audio segment durations. Most conversational turns are short, but the dataset also includes extended monologues exceeding 20 seconds, allowing models to capture both brief interactions and long-range motion or speaker dynamics.



(a) Spatial Distribution of Sound Sources Relative to the Listener

(b) Distribution of Sound Event Duration in MRSSound

Figure 5: Statistical overview of MRSSound. (a) Spatial distribution of sound sources relative to the listener. Red, green, and blue arrows denote the positive x, y, and z axes; azimuth is measured around the z-axis from the x-axis, and elevation is relative to the xy-plane. (b) Duration distribution of sound event duration in MRSSound.



(a) Spatial Distribution of Sound Sources Relative to the Listener

(b) Word Cloud

(c) Room Size

(d) Distribution of Audio Segment Durations

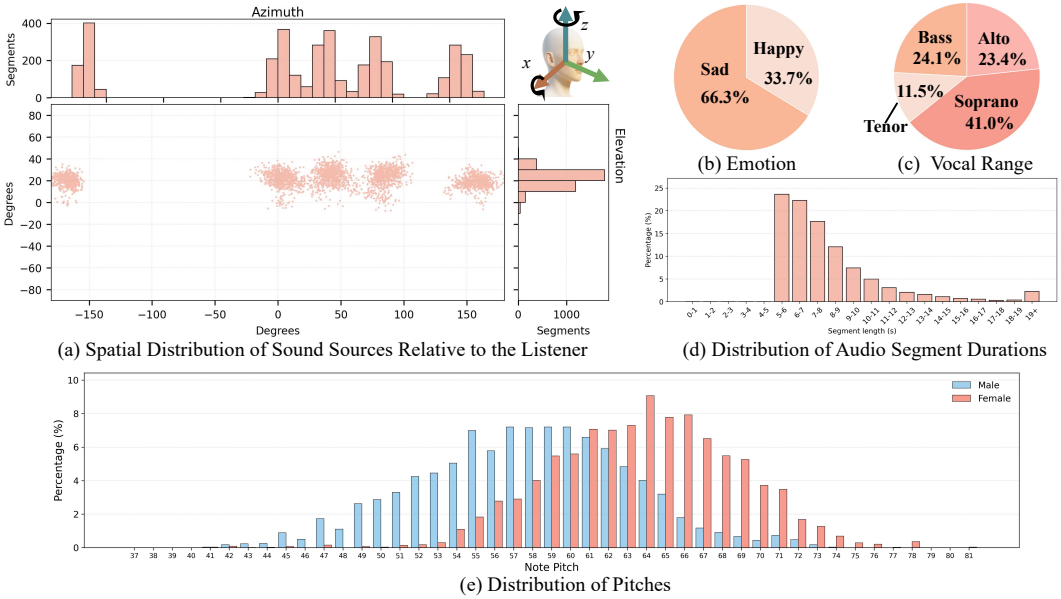
(e) Distribution of Phonemes in MRSSpeech

Figure 6: Statistical overview of MRSSpeech. (a) Spatial distribution of sound sources relative to the listener. Red, green, and blue arrows denote the positive x, y, and z axes; azimuth is measured around the z-axis from the x-axis, and elevation is relative to the xy-plane. (b) Word cloud. (c) Proportions of recording spaces by room size. (d) Distribution of audio segment durations. (e) Distribution of phonemes in MRSSpeech.

https://doi.org/10.52202/085713-4882

Finally, Figure 6(e) presents the phoneme distribution, which covers all common phonetic units in the dataset's target language. This phonetic diversity ensures that MRSSpeech provides strong generalizability for phoneme-aware models in speech synthesis and recognition.

A.5.3 Statistics of MRSSing



(a) Spatial Distribution of Sound Sources Relative to the Listener

(b) Emotion

(c) Vocal Range

(d) Distribution of Audio Segment Durations

(e) Distribution of Pitches

Figure 7: Statistical overview of MRSSing. (a) Spatial distribution of sound sources relative to the listener. Red, green, and blue arrows indicate the positive directions of the x, y, and z axes, respectively. Azimuth is measured around the z-axis from the x-axis, and elevation is relative to the xy-plane. (b) Emotion distribution. (c) Vocal range distribution. (d) Distribution of segment lengths. (e) Distribution of note pitches across segments.

Figure 7(a) shows the 3D spatial distribution of sound sources with respect to the listener. The majority of sources are located in front of the listener, consistent with the setup of solo vocal recordings. However, the coverage also spans surrounding directions in both azimuth and elevation, ensuring spatial variability for training robust spatial audio models.

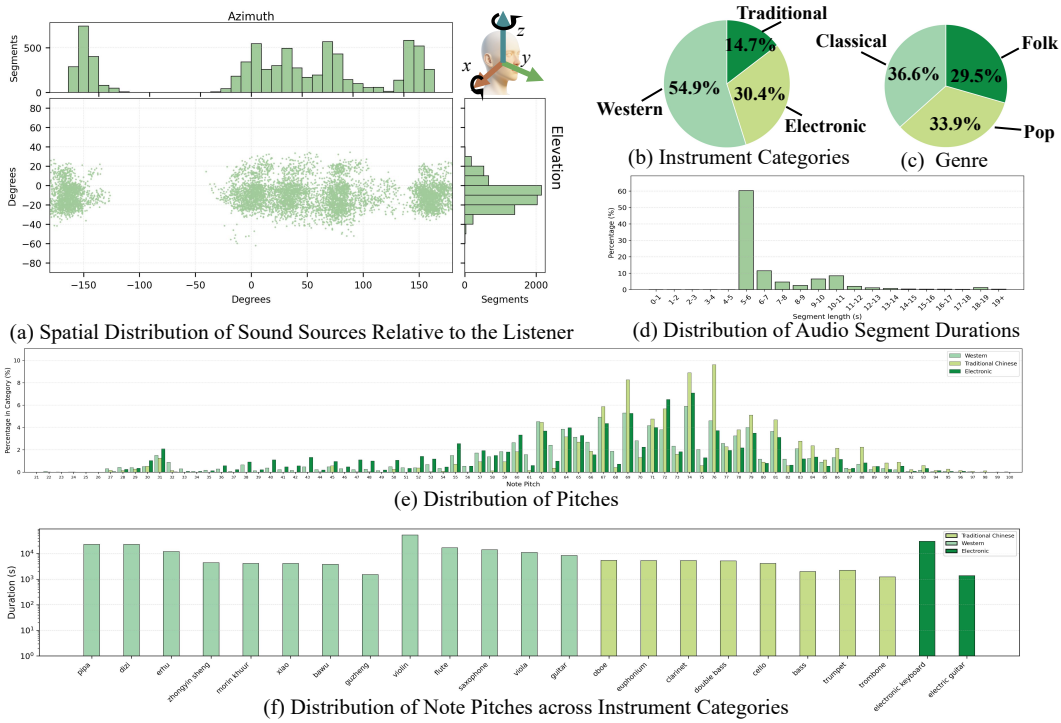
Figures 7 (b) and (c) highlight the diversity in style and content. Emotion annotations in Figure 7(b) include expressive labels such as happy and sad. Figure 7(c) displays the coverage across vocal ranges, including soprano, alto, tenor, and bass, ensuring a wide span of pitch and timbral variation.

Figure 7(d) presents the duration distribution of singing segments, which ranges from approximately 4 to 10 seconds. This aligns with typical input lengths used in training singing voice synthesis models. Figure 7(e) illustrates the distribution of note pitches. The full range of musical notes is well represented, and a clear difference in pitch range is observed between male and female singers, with females generally singing at higher pitches. This confirms that MRSSing captures realistic vocal and musical variability.

In summary, MRSSing provides extensive diversity in spatial positioning, language, emotion, vocal range, segment duration, and pitch. This makes it a strong foundation for research on spatial singing voice synthesis and expressive vocal modeling.

A.5.4 Statistics of MRSMusic

Figure 8(a) illustrates the spatial positions of musical instruments relative to the listener. Most sources are located in front of the listener, consistent with typical music listening scenarios. However, the coverage also includes surrounding directions, contributing to spatial diversity in training and evaluation.



(a) Spatial Distribution of Sound Sources Relative to the Listener

(b) Instrument Categories

(c) Genre

(d) Distribution of Audio Segment Durations

(e) Distribution of Pitches

(f) Distribution of Note Pitches across Instrument Categories

Figure 8: Statistical overview of MRSMusic. (a) Spatial distribution of sound sources relative to the listener. Red, green, and blue arrows denote the positive x, y, and z axes; azimuth is measured around the z-axis from the x-axis, and elevation is relative to the xy-plane. (b) Distribution of instrument categories duration. (c) Distribution of instrument genre duration. (d) Distribution of audio segment durations. (e) Distribution of pitches. (f) Distribution of instrument duration.

Figures 8(b) and (c) highlight the diversity of instrument types and musical genres. The instrument set spans Western, Traditional Chinese, and Electronic categories, while the genre annotations include folk, pop, and classical music. Figure 8(f) further details the recording durations across 23 instruments, showing a relatively balanced distribution that facilitates downstream learning for different instrument types.

Regarding generalization capacity, Figure 8(d) shows that audio segment durations range from approximately 4 to 11 seconds, matching typical training input lengths. Figure 8(e) demonstrates that the dataset covers a full range of musical pitch values, supporting tasks that require robust pitch generalization.

A.5.5 Statistics of Demographic Representation

Performers. Across all four subsets, we recruit students from diverse majors at Zhejiang University. The cohort consists of 161 individuals aged between 18 and 25, with a gender ratio of approximately 3:4 (male to female).

Annotation Team. We employ 25 professional annotators (aged 20–25, gender ratio of about 3:2, male to female) to perform fine-grained annotations of transcripts, phoneme boundaries, music scores, and movement descriptors. All annotators were trained student researchers with relevant domain expertise.

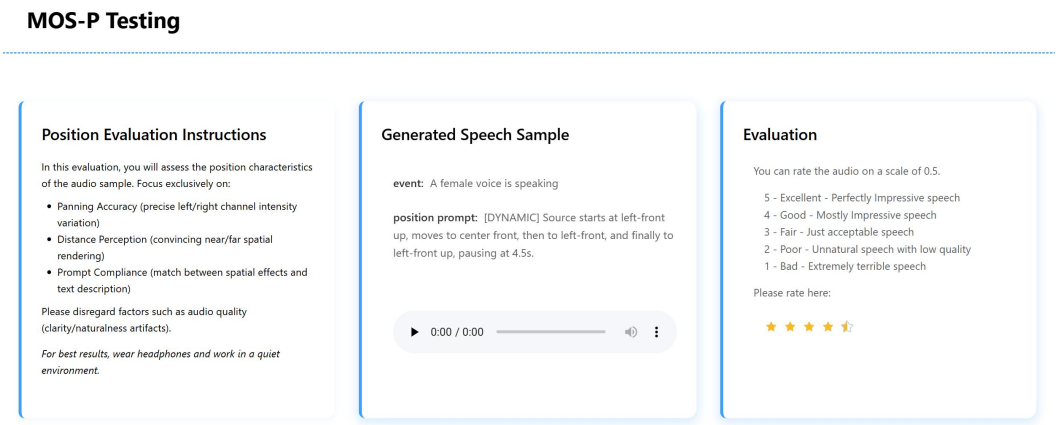
B Details of Experiments

B.1 Subjective Evaluation Metrics

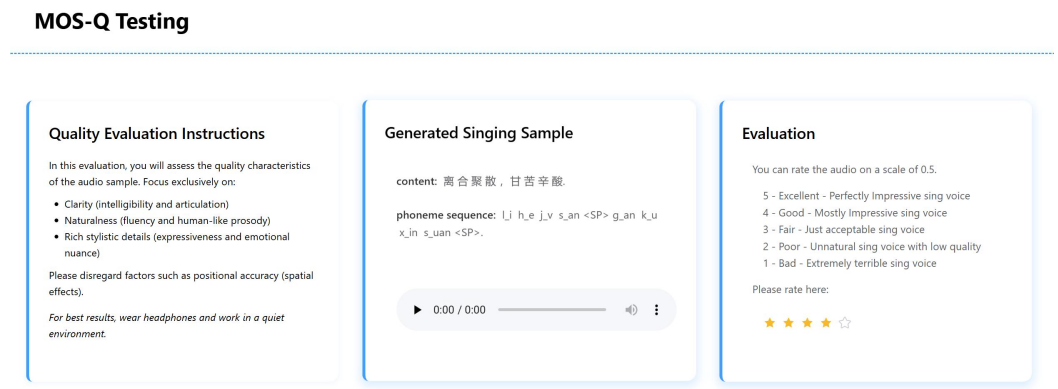
We conduct subjective evaluation of the generation tasks using Mean Opinion Score (MOS). For each task, we randomly sample 40 utterances from the test set. Each utterance is paired with a

https://doi.org/10.52202/085713-4882

corresponding source-position prompt and is evaluated by at least five expert listeners. MOS-Q, Listeners rate each sample on a five-point Likert scale ranging from 1 (bad) to 5 (excellent). For audio quality, we use MOS-Q, where listeners wear headphones and assess the clarity and naturalness of the generated audio. For spatial perception, we use MOS-P, where listeners evaluate the realism of spatial cues and whether the perceived direction and distance of the sound source match the prompt description. All participants are fairly compensated for their time at a rate of $15 per hour, resulting in a total cost of approximately $2000. Participants are informed that their evaluations will be used exclusively for academic research purposes. Instructions for audio evaluations are shown in Figure 9.



(a) Screenshot of MOS-P Testing



(b) Screenshot of MOS-Q Testing

Figure 9: Instructions for audio evaluations. (a) Screenshot of MOS-P Testing. (b) Screenshot of MOS-Q Testing

B.2 Objective Evaluation Metrics

To comprehensively assess spatial audio generation and understanding across multiple tasks, we adopt a set of objective metrics that evaluate signal fidelity, spatial consistency, intelligibility, and speaker or pitch accuracy. We randomly sample 400 data points as the test set.

Audio Spatialization. We measure waveform similarity using Wave L2, the mean squared error (MSE) between the generated and reference binaural waveforms. Amplitude L2 and Phase L2 are computed after applying Short-Time Fourier Transform (STFT), reflecting errors in the magnitude and phase components respectively. MRSTFT loss[6] is also used, combining spectral convergence, log-magnitude, and linear-magnitude terms for better spectral alignment. In addition, we use the perceptual evaluation metric PESQ[7] to assess audio quality for speech-related tasks. Since PESQ is

[6] https://github.com/csteinmetz1/auraloss

[7] https://github.com/aliutkus/speechmetrics

designed for speech, it is omitted for MRSLife and MRSMusic. For all metrics except PESQ, lower values indicate better performance.

To evaluate spatial consistency, we use Spatial-AST(Zheng et al., 2024) to extract angular and distance embeddings from the binaural audio. Since Spatial-AST predicts positions only for static sources, we compute the cosine similarity between the predicted and ground truth embeddings within 1-second segments and average the results to assess overall spatial fidelity.

Spatial Text to Speech. We evaluate speech intelligibility using Character Error Rate (CER), which measures the proportion of character-level differences between ASR transcriptions and reference texts. Transcriptions are generated using the Paraformer-zh model (Gao et al., 2023). To assess speaker consistency, we compute Speaker Identity Matching (SIM) as the cosine similarity between speaker embeddings extracted using a WavLM-based speaker verification model.[8]

Spatial Singing Voice Synthesis. We use Mel Cepstral Distortion (MCD) to assess the spectral similarity between the generated and reference vocals. It is defined as:

$$\text{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(m_t(d) - \hat{m}_t(d))^2}, \tag{1}$$

where $m_t(d)$ and $\hat{m}_t(d)$ represent the d-th Mel-frequency cepstral coefficient (MFCC) at frame t for the ground truth and synthesized signals, respectively, and D is the number of MFCC dimensions.

Additionally, we adopt F0 Frame Error (FFE) to evaluate pitch accuracy by comparing extracted F0 contours between the synthesized and ground truth audio.

Spatial Music Generation. We use Fréchet Audio Distance (FAD) (Kilgour et al., 2018) to assess perceptual similarity between the feature distributions of generated and reference audio. In addition, F0 Frame Error (FFE) is used to evaluate pitch accuracy by comparing the extracted F0 contours against the reference musical scores.

Sound Event Localization and Detection (SELD). Following STARSS23 (Shimada et al., 2023), we adopt four joint detection and localization metrics: F_{20°: location-aware F-score; a prediction is correct if the event class matches and angular error is below 20°. ER_{20°: error rate computed as the sum of insertions, deletions, and substitutions over reference events. LE_{CD}: class-aware localization error, the mean angular difference between predicted and reference directions. LR_{CD}: class-aware localization recall, the percentage of correctly localized events among all instances of each class. We compute all metrics in 1-second non-overlapping segments using macro-averaging across all event classes. Higher values of F_{20° and LR_{CD}, and lower values of ER_{20° and LE_{CD} indicate better performance.

B.3 Audio Spatialization

We build upon BinauralGrad's two-stage diffusion-based framework to convert monaural audio into spatial audio. However, in our subsequent generation experiments, the synthesized monaural audio typically corresponds to a centrally positioned source between the ears, effectively serving as the first stage of BinauralGrad. Therefore, for the spatialization experiments presented here, we use only the second stage of BinauralGrad to validate spatial audio generation. Specifically, we first convert binaural recordings to monaural input by averaging the two channels. Next, we apply a DSP-based method to produce a coarse spatial approximation of the binaural signal. This monaural input and its simulated binaural counterpart are then used as input to the BinauralGrad model, which is conditioned on the object's motion trajectory to generate spatialized binaural audio. We list the architecture and hyperparameters of BinauralGrad in Table 8.

[8] `https://huggingface.co/pyannote/speaker-diarization`

 https://doi.org/10.52202/085713-4882

Table 8: Hyper-parameters of BinauralGrad modules.

Hyperparameter		BinauralGrad
Binaural Encoder	Wave Encoder Layers	2
	Position Encoder Layers	2
	Encoder Conv1D Kernel	3
	Encoder Dropout	0.4
Mel Predictor	Residual Blocks	3
	Bidirectional Layers	3
	Hidden_size	128
	Training Steps	200
	Sampling Steps	6

B.4 Spatial Text to Speech

We fine-tune two pre-trained models, CosyVoice and F5-TTS, using monaural audio obtained by averaging the binaural recordings from the MRSSpeech subset. This allows the models to learn the generation characteristics specific to our dataset. After monaural generation, we apply a pre-trained spatialization model trained on MRSSpeech to convert the output into spatialized binaural audio. For the ISDrama baseline, we adopt the model variant proposed in the original paper that conditions generation on predefined spatial paths to produce binaural spatial speech directly.

Table 9: Hyper-parameters of Rmssinger modules.

Hyperparameter		Rmssinger
Phoneme Encoder	Phoneme Embedding	256
	Encoder Layers	4
	Encoder Hidden	256
	Encoder Conv1D Kernel	9
	Encoder Conv1D Filter Size	1024
	Encoder Attention Heads	2
	Encoder Dropout	0.1
Note Encoder	Pitches Embedding	256
	Type Embedding	256
	Duration Hidden	256
Pitch Predictor	Conv Layers	12
	Kernel Size	3
	Residual Channel	192
	Hidden Channel	256
	Training Steps	100
Mel Predictor	Conv Layers	20
	Kernel Size	3
	Residual Channel	256
	Hidden Channel	256
	Training Steps	100

B.5 Spatial Singing Voice Synthesis

For the ISDrama baseline, we adopt the spatial-path-conditioned generation framework and extend it by incorporating Rmssinger's note encoder, allowing explicit pitch control through musical score input. We use the single stage variation of Rmssinger, a standard singing voice synthesis model, and train it with monaural targets derived from the averaged binaural recordings in the MRSSing subset. This enables the model to generate pitch-accurate audio that reflects the acoustic characteristics. The synthesized monaural audio is then spatialized using the spatialization model trained on MRSSing. We list the architecture and hyperparameters of Rmssinger in Table 9.

B.6 Spatial Music Generation

For spatial music generation, in the ISDrama setting, we follow the path-conditioned generation pipeline but remove the phoneme encoder, using only the note encoder to process symbolic scores as the sole control condition for generating spatialized music. We also adopt Make-An-Audio 2 as our base model and train it using monaural audio derived from averaged MRSMusic binaural recordings. We use the pretrained spectrogram autoencoder. Instrument category and symbolic score information are concatenated into the text prompt to guide the music generation process. The generated audio is subsequently spatialized using the MRSMusic-trained spatialization model. We list the hyper-parameters of Make-An-Audio 2 in Table 10.

Table 10: Hyperparameters of Make-An-Audio 2.

Hyperparameter		Make-An-Audio 2
Autoencoders	Input/Output Channels	80
	Hidden Channels	20
	Residual Blocks	2
	Spectrogram Shape	(80, 624)
	Channel Multipiler	$[1, 2, 4]$
Transformer Backbone	Input shape	(20, T)
	Condition_embedding Size	1024
	Feed-forward Hidden_size	576
	Num of Transformer Heads	8
	Transformer Blocks	8
	Training Steps	1000
	Sampling Steps	100
CLAP Text Encoder	Transformer Embed Channels	768
	Output Project Channels	1024
	Token Length	77

Table 11: Hyperparameters of SELD.

Hyperparameter		SELD
Input	Binaural Channels	3
	FOA Channels	7
	Frequency Bins	128
	Frames	120
Audio Encoder(CNN)	Hidden_size	64
	Conv Blocks	3
Audio Encoder(Vit)	Hidden_size	128
	Conv Blocks	1
	Transformer Blocks	2
	Num of Transformer Heads	4

B.7 Sound Event Localization and Detection

We follow the STARSS23 framework for FOA-based sound event localization and detection, training on the event segments from the MRSSound subset under both audio-only and audio-visual conditions. To enable binaural input, we modify the model architecture to use three input channels and extract interaural phase difference features. This allows us to adapt the SELD model to binaural audio. Additionally, we experiment with replacing convolutional layers with Transformer encoders to explore the effect of different architectures on sound event localization and detection performance. We list the hyper-parameters of SELD in Table 11.

 https://doi.org/10.52202/085713-4882

C Authorship Statement

Zhou Zhao and **Fei Wu** served as the overall project leaders and supervisors, offering full-scale resource support throughout the development of the MRSAudio project.

Data Planning **Wenxiang Guo** is responsible for all aspects of the MRSAudio project, from its initial conception to its final execution. The planning and design of the individual sub-datasets are undertaken by:

- **MRSSpeech**: Yu Zhang, Zhiyuan Zhu, Wenxiang Guo
- **MRSSing**: Changhao Pan, Yu Zhang
- **MRSMusic**: Changhao Pan, Yu Zhang
- **MRSAudio**: Wenxiang Guo, Zhiyuan Zhu, Li Tang

Data Recording

- **MRSSpeech**: **Xintong Hu** serves as the principal coordinator of the dataset, taking charge of venue management, volunteer recruitment, audio recording, and data organization. **Yixuan Chen** and **Pengfei Fan** assist in the recording and organization of audio data, as well as the setup of recording environments. **Zhetao Chen**, **Yanhao Yu**, **Ke Xu**, and **Qiange Huang** participate in the audio recording and data organization. **Wenxiang Guo** provides technical support during the recording process.
- **MRSSing**: **Changhao Pan** serves as the recording lead for this sub-dataset, taking charge of venue management, volunteer recruitment, audio recording, and data organization. **Yuhan Wang** assists with the recruitment and coordination of recording personnel and primarily handles data organization. **Ke Xu** and **Li Tang** participate in the recording and organization of audio data.
- **MRSMusic**: **Changhao Pan** serves as the recording lead for this sub-dataset. **Changhao Pan** and **Hankun Xu** are responsible for the recruitment and coordination of performers. **Han Wang**, **Hankun Xu**, and **Changhao Pan** jointly conduct the audio recording, with **Han Wang** primarily responsible for data organization.
- **MRSLife**: **Wenxiang Guo** serves as the lead coordinator for the MRSLife sub-dataset, which consists of two parts: *MRSSound* and *MRSDialogue*. **Wenxiang Guo**, **Rui Yang**, **Zhiyuan Zhu**, and **Xintong Hu** jointly carry out the recording of MRSSound. **Zhiyuan Zhu** is responsible for the scene design of MRSDialogue, while **Yuhan Wang**, **Rui Yang**, and **Zhiyuan Zhu** conduct the recording of MRSDialogue.

Data Processing and Annotation

- **MRSSpeech**: **Zhiyuan Zhu** is responsible for the complete data processing pipeline, including segmentation, ASR, alignment, and the organization of spatial annotation. **Wenxiang Guo** provides technical support and guidance throughout the process.
- **MRSSing**: **Changhao Pan** and **Zongbao Zhang** process the dataset, covering segmentation, text alignment, note transcription, and the alignment of spatial information. **Yuhan Wang** is responsible for global style annotation, while **Han Wang** organizes the music score data.
- **MRSMusic**: **Changhao Pan** and **Zongbao Zhang** handle data processing, including segmentation, note transcription, and spatial information alignment. **Hankun Xu** is responsible for both global style annotation and music score organization.
- **MRSLife**: **Zhiyuan Zhu** and **Wenxiang Guo** jointly perform the data processing. **Rui Yang** annotates the audio events in the recordings.

Experiments and Benchmarks

- **Sound Event Localization and Detection (SELD)**: **Wenxiang Guo** updates and adapts the baseline models for this task. **Zhiyuan Zhu** is responsible for training and evaluating the baseline models. **Changhao Pan** provides technical guidance for the reproduction process.

- **Audio Spatialization**: **Wenxiang Guo** is solely responsible for the design, implementation, and evaluation of this benchmark.
- **Text-to-Speech**: **Yu Zhang** conducts the experimental work for this benchmark, while **Wenxiang Guo** performs model evaluation.
- **Singing Voice Synthesis**: **Changhao Pan** and **Yu Zhang** carry out the experiments for this benchmark. **Xintong Hu** is responsible for model evaluation.
- **Spatial Music Generation**: **Changhao Pan**, **Wenxiang Guo**, and **Yu Zhang** jointly conduct the experiments and evaluations for this benchmark.

Paper Writing **Wenxiang Guo** is responsible for the overall writing of the paper. **Changhao Pan** contributes to the writing of the experimental sections. **Zhiyuan Zhu** is in charge of the writing of the statistical analysis section. **Xintong Hu**, **Yu Zhang**, and **Zhou Zhao** participate in revising and polishing the manuscript.

 https://doi.org/10.52202/085713-4882