
Sci-Phi: A Large Language Model Spatial Audio Descriptor

Xilin Jiang¹, Student Member, IEEE, Hannes Gamper², Member, IEEE
and Sebastian Braun², Senior Member, IEEE

¹Columbia University, New York, NY, USA

²Microsoft Research, Redmond, WA, USA

Work completed during internship at Microsoft.

ABSTRACT Acoustic scene perception involves describing the type of sounds, their timing, their direction and distance, as well as their loudness and reverberation. While audio language models excel in sound recognition, single-channel input fundamentally limits spatial understanding. This work presents *Sci-Phi*, a spatial audio large language model with dual spatial and spectral encoders that estimates a complete parameter set for all sound sources and the surrounding environment. Learning from over 4,000 hours of synthetic first-order Ambisonics recordings including metadata, *Sci-Phi* enumerates and describes up to four directional sound sources in one pass, alongside non-directional background sounds and room characteristics. We evaluate the model with a permutation-invariant protocol and 15 metrics covering content, location, timing, loudness, and reverberation, and analyze its robustness across source counts, signal-to-noise ratios, reverberation levels, and challenging mixtures of acoustically, spatially, or temporally similar sources. Notably, *Sci-Phi* generalizes to real room impulse responses with only minor performance degradation. Overall, this work establishes the first audio LLM capable of full spatial-scene description, with strong potential for real-world deployment. Demo: <https://sci-phi-audio.github.io/demo>

INDEX TERMS Spatial audio, large language model, acoustic scene understanding.

I. Introduction

A spatial acoustic scene is an organic whole of multiple sound events and ambient noise, together with the environment that shapes them. It includes source identity and content; onsets, offsets, and overlaps; direction and distance; loudness and reverberation; and the room’s overall imprint. These aspects are intertwined, and human listeners naturally bind them into a stable, unified representation: psychophysics studies on human hearing [1], [2] have shown that perception groups soundscape into coherent auditory objects and scenes using patterns over time and space, with distance and reverberation shaping where sources seem to be and spatial structure enabling selective listening in clutter. To fully analyze and understand an acoustic scene, one needs to detect multiple sources alongside background, track them along time, localize them in azimuth and elevation, estimate distance and level, and characterize the room.

To solve this task, neural network-based machine listeners have progressed along several strands: sound event detection and localization [3], [4]; automatic speech recognition (ASR) [5], [6]; and general-purpose audio understanding with emergent audio large language models (LLMs) [7],

[8]. Yet despite strong task performance, these models still fall short of perceiving an acoustic scene as an integrated whole: they typically focus on a single (or dominant) foreground source, omit spatial parameters (e.g., direction and distance), and offer little account of the environment (e.g., reverberation, room volume, noise). This gap motivates us to generalize machine listening from recognizing single auditory objects to narrating entire acoustic scenes. Our research goal is therefore twofold: (i) to investigate whether a machine can understand the entire spatial acoustic scene, including *what*, *when*, *where*, and *how* of the sound sources, and the *environment*, analogous to human perception; and (ii) to build a spatial audio understanding model that can be extended to downstream applications including hearing assistants, robotics perception, navigation, and automatic spatial environment monitoring and annotation.

This work introduces *Sci-Phi*, *Spatial-scene comprehension and inference with Phi*, the first spatial audio LLM capable of full spatial-scene description. *Sci-Phi* builds on Phi-4 Multimodal [12], a powerful multimodal LLM for audio understanding and speech recognition that is nevertheless restricted to single-channel

Model	Format	Localization	Speech	Audio	Noise	Acoustics Params	Test on Real RIR
BAT [9]	Binaural	3D + dist.	✗	✓	✗	✗	✗
Can LLM...? [10]	FOA	2D angle	en	✗	✗	✗	✗
SING [11]	Owlet (1ch)	2D angle	en	✗	✗	✗	✗
Phi-4 MM [12]	Monaural	✗	8 lang.	✓	✗	✗	n.a.
Sci-Phi	FOA	3D + dist.	8 lang.	✓	Type, Loudness	Loudness, reverb, room size	✓

TABLE 1. Existing (spatial) audio LLMs vs. *Sci-Phi*, highlighting *Sci-Phi*'s full scene description and generalization ability.

audio input. At a high level, *Sci-Phi* couples a *spatial encoder* with an *audio encoder* and is trained to generate scene metadata from >4,000 hours of synthetic first-order Ambisonics (FOA) mixtures spanning 1—4 sources, background noise, and diverse rooms. Our contributions are threefold: (i) a spatial audio LLM, integrating a spatial encoder, powered by a spatial data and metadata generation pipeline, for comprehensive spatial-scene description; (ii) a permutation-invariant evaluation protocol with 15 metrics that account for multiple sources and environmental attributes; and (iii) extensive experiments demonstrating generalization to real room impulse responses (RIRs), along with careful analyses across SNR, reverberation, source count and other attributes. *Sci-Phi* advances audio foundation models from isolated object recognition to coherent spatial-scene understanding, with promising results towards real-world generalization.

II. Related Works

Research in spatial audio has progressed from sound event detection and localization (SELD) to representation learning and, more recently, to spatial understanding with LLM. Standard SELD systems jointly estimate class labels and locations for multiple sources, with advances in model design [13], [14], training objective [15], and benchmarks [16]. Despite strong ad-hoc performance, they assume a small label set and remain tailored to SELD, limiting open-vocabulary and holistic scene understanding. Another line of work explores self-supervised learning (SSL), including contrastive learning [17] and masked reconstruction [18], [19], without the need for labels and therefore learning embeddings that are naturally generalizable to new labels. However, these SSL models were designed and evaluated by task-specific heads for standard SELD and ASR tasks, limiting zero-shot and task co-learning ability. A related direction learns joint spatial audio-text embeddings [20], [21] via CLIP/CLAP-like cross-modal contrastive learning [22], [23]. While useful for captioning and retrieval, these embeddings are generally limited to single sources and are not directly applicable to open-ended text generation, such as multi-source scenery description.

Audio LLMs, monaural or spatial, define today's standard of audio foundation models with open-vocabulary user

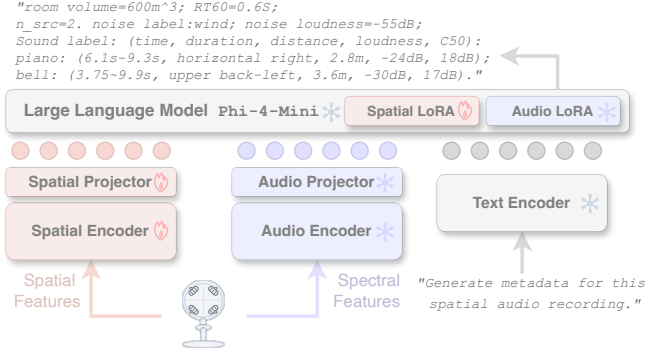


FIGURE 1. *Sci-Phi* architecture, derived from Phi-4-Multimodal (visual components not shown for clarity). Fire and snowflake mark the trainable and frozen components. Light red, blue, and grey colors correspond to spatial, spectral, and textual features, modules, embeddings, and computation flow.

queries and responses. They typically pair an audio encoder with a pretrained transformer-decoder language model (i.e., GPT [24]), leveraging strong linguistic priors and a unified next-token objective across tasks such as SELD, ASR, and more general Q&A. Table 1 reviews current spatial audio LLMs to the best of our knowledge; Phi-4-Multimodal [12] is included as a representative monaural audio LLM [7], [8]. Although heterogeneous spatial formats and limited open-sourcing hinder direct apples-to-apples comparison, most existing spatial audio LLMs are restricted to one audio domain (speech or non-speech), provide only partial localization (2-D or without distance), and omit background and room acoustics completely. In contrast, *Sci-Phi* offers full spatial-scene understanding with a scalable number of directional sources and is the first to demonstrate generalization on real RIRs.

III. Sci-Phi

A. Multimodal Features and Architecture

The overall architecture of *Sci-Phi* is shown in Figure 1. *Sci-Phi* is a spatial audio LLM with two encoders: a spatial encoder for spatial features and an audio encoder for spectral features. Both features are derived from a first-order Ambisonics (FOA) waveform of four channels (W, X, Y, Z), where W is omnidirectional. Concretely, we compute (i) mel spectrograms of all four channels and (ii) intensity vectors (IVs) [25] for (X, Y, Z) relative to W . These seven maps (4 mel + 3 IV) are stacked as the spatial features, while the spectral features are the mel spectrogram of the W channel alone, since the monaural audio encoder only accepts single-channel inputs.

The spatial encoder borrows the architecture and checkpoint of SELDNet¹ [13], [14]. SELDNet contains 3 convolution layers, 2 gated recurrent units, and 2 self-attention layers. Although the encoder was pretrained on sound event

¹Available at https://github.com/partha2409/DCASE2024_seld_baseline

detection and localization, both the amount and coverage of training data, with only 13 sound event labels, and mostly only horizontal spatial direction coverage [16], are insufficient to generalize to more complex acoustic scenes (e.g., our test sets). Therefore, we further finetune the spatial encoder together with the LLM by instruction-tuning on a larger and more diverse training set. While separately pre-training the audio encoder on a larger dataset may help generalization even further, we found the joint training of encoder and LLM to perform well.

We directly use the pretrained monaural audio encoder from *Phi-4 Multimodal*, which consists of 3 convolution layers and 24 conformer blocks [26]. We freeze the monaural audio encoder to maximally preserve its original audio understanding ability trained on monaural audio, particularly its state-of-the-art speech recognition capability in 8 languages. Finally, two separate 2-layer linear projectors (trained from scratch for spatial, and frozen for audio encoder) project the spatial and audio encoder outputs to the same dimension (3072) as the text embedding. The spatial, audio, and text embedding are modeled jointly by the LLM *Phi-4-Mini* [12] (3.8B *small* LLM). The LLM reads input in the following format:

```
"<|user|><|spatial|><|audio|><|question|><|end|>
<|assistant|><|answer|><|end|>"
```

where `<|spatial|>` and `<|audio|>` are variable-length placeholders for spatial and audio embeddings, respectively.

B. Data Generation

Because well-annotated spatial audio corpora large enough to train *Sci-Phi* are not available publicly, we synthesize first-order Ambisonics (FOA) training data and paired metadata at scale. Each 10s sample is created by (i) sampling a room with pre-rendered multi-channel room impulse responses (RIRs), (ii) placing 1–4 directional sound sources distributed in the room and a diffuse background by convolving audio sources with the RIRs, and (iii) randomizing levels, spectral filtering etc. before mixing sources and background. The training set contains 1.6 million 10s mixtures ($\sim 4,444$ h), generated as follows.

Rooms and RIRs. We simulate 10k rooms with the image-source model [27]. Room sizes range from $4 \times 4 \times 3\text{m}^3$ to $25 \times 25 \times 6\text{m}^3$, and the FOA microphone is placed at a random position. For each room we precompute 64 candidate source positions with a roughly spherically uniform direction distribution. We also record room-level attributes such as reverberation time (RT60) and volume.

Sound sources. As diverse sound source corpora we use speech from CommonVoice [28] (8 languages, ~ 385 h) and general audio from Freesound (~ 230 k files) and the BBC sound-effects collection (~ 33 k files). We clean tags and captions with an LLM to remove recording-condition notes and sound-irrelevant text, and divide them into single-source

and multi-source/ambient files. We use files described as multi-source/ambient as background noises and convolve them with all 64 RIRs from one room to simulate diffuse sound.

Metadata and quantization. Each mixture is accompanied by human-readable scene metadata. Room fields include RT60 and room volume; background fields include noise type and its loudness; each source has a caption (also transcription for speech), onset/offset times, direction, distance, level (dBA), and C50. To stabilize generation and evaluation, we quantize: (i) 3-D direction (azimuth and elevation) into 26 regions using 45° angular bins (e.g., “upper back-left”, “horizontal front-right”, “above”), (ii) distance to 0.1m, (iii) RT60 and time to 0.1s, (iv) loudness and C50 to 1dB, and (v) room volume to 100m^3 . These choices hit a balance between these physical (and mostly continuous) acoustic attributes and simple and descriptive language targets. The sound levels also allow calculation of SNR.

Test sets. We generate two test sets with 10k clips each (27 h): a held-out **synthetic-RIR** test set using 100 unseen rooms and unseen audio sources from SoundBible(.com) and speech from VCTK (English only) [29]; a **real-RIR** test set spatializes anechoic sources via real FOA RIRs, and adds real spatial background recordings from 100 real rooms, all from the FOA-MEIR dataset [30]. The FOA-MEIR datasets contains a set of anechoic sound event recordings, and we again use anechoic English speech from VCTK. The FOA-MEIR test set is limited in spatial coverage: *no sources outside the horizontal plane* (above and below $\pm 22.5^\circ$), *no room volume* information, and *only ambient background noise without specific labels*. To test these absent conditions, we have to rely on the synthetic test set only.

C. Training Objective

Sci-Phi is trained to generate a full description of the spatial acoustic scene. We serialize the scene metadata into the `<|answer|>` string with the template below, starting from the environment to the sources:

```
room_volume=<room_volume>; RT60=<rt60>;
n_src=<n_src>. noise_label:<noise_type>;
noise_loudness=<noise_db>.
Sound label:(time, direction, distance, loudness, C50):
<label_1>: (<time_1>, <direction_1>,
           <distance_1>, <loudness_1>, <C50_1>);
<label_2>: (<time_2>, <direction_2>,
           <distance_2>, <loudness_2>, <C50_2>);
...
```

Fields in blue are sample-specific parameters. It is important to note that the **source enumeration order** must be fixed a priori for the LLM to learn and will affect its performance (see Table 3). Unless otherwise noted, we order sources by decreasing **loudness**.

The trainable components are the spatial encoder, the spatial projector, and the spatial low-rank adaptation (LoRA) [31] inside the LLM. We keep the existing mono audio LoRA in *Phi-4-Multimodal* frozen and initialize the

TABLE 2. We evaluate multiple metrics on multiple sources in arbitrary orders via either (1) per-metric optimal permutations $\mathcal{P}_{OM}$ or (2) a single $\mathcal{P}_{OS}$ that maximizes *TupleScore* (joint *What/When/Where*) with respect to the target sources. Scene-level metrics (global attribute, no permutation) are in gray. $\mathcal{P}_{OS}$ scores closely match $\mathcal{P}_{OM}$ per-metric optima. Note: $\mathcal{P}_{OS}$ WER may be slightly better than $\mathcal{P}_{OM}$ because pairs lacking either transcript are skipped.

Protocols	RoomVol ErrLog2	RT60 Err (s)	Noise CLAP	Count Accuracy (%)	Tuple Score	Source CLAP	WER	Direction Accuracy (XYZ XY Z, %)	Zone Err (°)	Distance ErrRatio	Time IoU	Loudness Err (dB)	C50 Err (dB)
<i>On synthetic-RIR test set</i>													
Optimal-Metric $\mathcal{P}_{OM}$	0.590	0.092	0.662	91.5	0.783	0.694	0.464	85.8 92.1 94.0	6.1	0.228	0.815	1.011	1.217
Optimal-Source $\mathcal{P}_{OS}$	0.590	0.092	0.662	91.5	0.783	0.674	0.449	82.9 85.1 92.0	8.4	0.258	0.802	1.253	1.348
<i>On real-RIR test set</i>													
Optimal-Metric $\mathcal{P}_{OM}$	u.a.	0.333	u.a.	75.2	0.765	0.712	0.387	79.7 90.1 88.4	10.4	0.254	0.746	1.642	1.948
Optimal-Source $\mathcal{P}_{OS}$	u.a.	0.333	u.a.	75.2	0.765	0.691	0.371	77.4 84.4 87.5	12.0	0.292	0.737	1.975	2.203

spatial LoRA with the same configuration, i.e., a rank of 320. We optimize the next-token prediction objective below, with Q denoting the question tokens, A the answer tokens, and X_{spatial} , X_{spectral} the spatial and spectral embeddings:

$$\mathcal{L} = - \sum_{i=1}^L \log P_{\theta}(A_i | Q, X_{\text{spatial}}, X_{\text{spectral}}, A_{<i}) \quad (1)$$

Sci-Phi and all baseline models were trained for five epochs with an AdamW optimizer [32], a total batch size of 24, a peak learning rate of 1.0e-4, a linear learning rate warm-up (5% steps) followed by linear decay, on 8 NVIDIA A100 GPUs with bfloat16 precision.

IV. Evaluation Method

While many sentence-level NLP metrics calculate a score between the ground-truth and the generated scene description, they miss precision in specific physical or categorical attributes. Therefore, we extract each attribute (e.g., RT60, direction, distance) from the description and calculate ad-hoc metrics on them. The metrics include the cosine similarity of the audio-aware text embedding from CLAP [23] for source and noise descriptions, **accuracy** of source counting and direction for which we quantize into 26 XYZ (full sphere), 8 XY (azimuth), and 5 Z (elevation) zones, absolute **error** of direction (with respect to the center of the quantized zone), RT60 (s), loudness (dB), C50 (dB), and transcription (i.e., **WER**), or the **error ratio** or log2 of it for distance (m) and room volume (m^3), and finally the intersection-over-union (**IoU**) of estimated vs. ground-truth source active intervals. The IoU is defined as

$$\text{IoU}(g, s) = \frac{\cap(g, s)}{(t_{\text{off}}^g - t_{\text{on}}^g) + (t_{\text{off}}^s - t_{\text{on}}^s) - \cap(g, s)} \quad (2)$$

where $\cap$ denotes the intersection given by

$$\cap(g, s) = \max(0, \min(t_{\text{off}}^g, t_{\text{off}}^s) - \max(t_{\text{on}}^g, t_{\text{on}}^s)) \quad (3)$$

and t_{on} and t_{off} denote the onset and offset time of sound events. A few metrics (room volume, RT60 and background noise type and loudness) are defined and calculated for the entire scene, while others are calculated for each source separately, leading to a critical problem of how to find the

best matched sources from the generated description to the solution.

Permutation-invariant Evaluation. Language models generate tokens autoregressively in a single output stream. While we train *Sci-Phi* to enumerate sources by decreasing loudness, different permutations appear still valid to the human perceivers. Therefore, we argue that evaluation should be **order-invariant** so that correct answers with mismatched or arbitrary orders are not unfairly penalized. We represent each source with six attributes (*label*, *time*, *direction*, *distance*, *loudness*, *C50*) and parse both the generated description and the reference into lists of tuples, $G = [g_1, \dots, g_m]$ and $S = [s_1, \dots, s_n]$. We then seek a permutation matrix $\mathcal{P}$ that reorders G (or S) to calculate $\text{Metric}(\mathcal{P}G, S)$ averaged by all sources in the scene.

We could define the **optimal-metric permutation** $\mathcal{P}_{OM} = \max_{\text{All } \mathcal{P}} \text{Metric}(\mathcal{P}G, S)$ that maximizes a single metric, like label or direction. However, the downside is that it ignores cross-attribute association: e.g., if $G = [(dog, left), (cat, right)]$ and $S = [(cat, left), (dog, right)]$, per-metric matching can yield perfect scores for both *label* and *direction* despite mismatched association. Instead, we define **optimal-source permutation** $\mathcal{P}_{OS}$ which maximizes a composite *TupleScore* of multiple attributes and does not advantage any particular metric.

$$\mathcal{P}_{OS} = \max_{\text{All } \mathcal{P}} \text{TupleScore}(\mathcal{P}G, S) \quad (4)$$

where the *TupleScore* is geometric mean of *what*, *where*, and *when*, with each term and the final score normalized to 0–1:

$$\text{TupleScore}(g, s) = (\text{What} \cdot \text{Where} \cdot \text{When}(g, s))^{1/3} \quad (5)$$

with the specific metrics defined as

$$\text{What}(g, s) = \begin{cases} \frac{\text{CLAPScore}(g, s) + 1}{2} & \text{for audio} \\ \max(0, 1 - \text{WER}(g, s)) & \text{for speech} \end{cases} \quad (6)$$

$$\text{Where}(g, s) = \frac{180 - \text{ZoneError}(g, s)}{180} \quad (7)$$

$$\text{When}(g, s) = \text{IoU}(g, s) \quad (8)$$

where the IoU is given in (2). The *What* metric uses *WER* only if a speech source is detected and transcribed

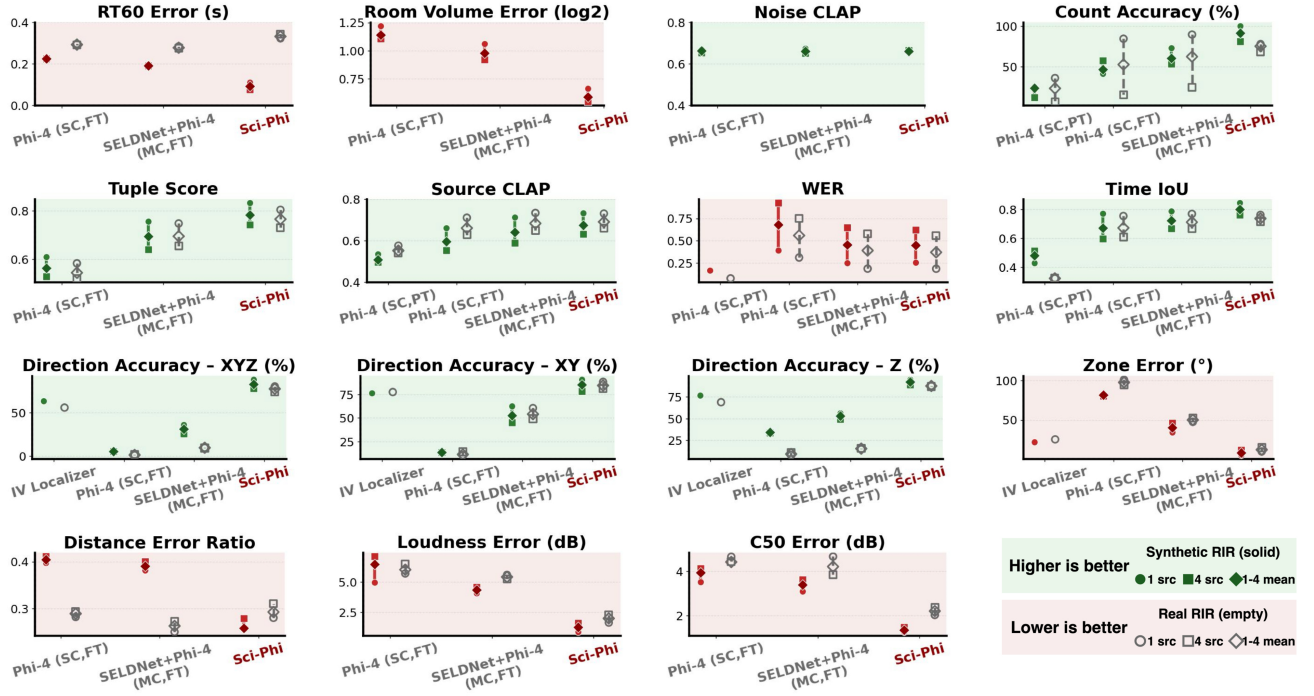


FIGURE 2. Spatial audio analysis results of the synthetic-RIR (solid) and real-RIR (empty) test sets. Each subplot is one evaluation metric. Green/Red indicates higher/lower is better. Note: room volume error and noise CLAP are missing for real-RIR test due to a lack of ground-truths in FOA-MEIR [30].

in the correct language, and the ground-truth transcription is available. Otherwise, we default to CLAP score.

We first determine $\mathcal{P}_{OS}$ and apply it to compute all metrics. By definition, $\mathcal{P}_{OS}$ (and its score) equals $\mathcal{P}_{OM}$ on *TupleScore*; for other metrics, $\mathcal{P}_{OM}$ may choose different permutations that define per-metric upper bounds. Nonetheless, in Table 2, $\mathcal{P}_{OS}$ scores closely match $\mathcal{P}_{OM}$ scores, indicating that *Sci-Phi* learns consistent source-level associations rather than estimating attributes in isolation. At last, our *TupleScore* and $\mathcal{P}_{OS}$ formulation can be easily customized for other problems that require finding the best permutation of multiple sources based on multiple metrics.

V. Results

As reference points to our proposed system, we include the following baselines:

- 1) **IV Localizer**: ad-hoc DSP intensity-vector localizer.
- 2) **Phi-4 (SC, PT)**: pretrained single-channel Phi-4-Multimodal prompted for ASR and multi-sound tagging.
- 3) **Phi-4 (SC, FT)**: finetuned single-channel Phi-4-Multimodal.
- 4) **SELDNet+Phi-4 (MC, FT)**: finetuned multi-channel Phi-4-Multimodal with a frozen SELDNet encoder.

Figure 2 shows the main results for the synthetic RIR test set (solid symbols) and real RIRs (empty symbols). The test data is divided into 4 subsets, depending on the number of directional sources per test clip, i.e. 1-4 sources.

Figure 2 marks the results of the subsets of clips with 1 source ($\circ$) and 4 sources ($\square$), and the average over the whole test set with 1–4 sources ($\diamond$). Across nearly every metric, *Sci-Phi* outperforms all baselines. As expected, the mono Phi-4 baseline fails to localize, and further has poor room volume estimation, source counting, loudness and C50 estimation. Finetuning the mono Phi-4 on our data helps improve counting, source recognition and onset detection. For synthetic RIRs on average, *Sci-Phi* achieves 91.5% count accuracy, 0.783 tuple score, 82.9%, 85.1%, 92.0% accuracy for XYZ, XY, Z with 8.4° zone error, 0.802 time-IoU, and low errors in distance (0.258), loudness (1.25 dB), and C50 (1.35 dB). For real RIRs, *Sci-Phi* averages 75.2% count accuracy, 0.765 tuple score, 77.4%, 84.4%, 87.5% direction accuracy, and 0.737 time-IoU. Reverberation-related metrics degrade more: RT60 error rises from 0.092 s to 0.333 s and distance error from 0.258 to 0.292, suggesting potential overfitting to synthetic RIRs. However, other metrics related to semantic, timing, and direction estimation remain strong in unseen real RIRs and sound sources, indicating strong generalization to real rooms. There are two exceptions aligning with expectations: noise description (CLAP) is on par among all finetuned baselines, implying that predicting diffuse noise type does not demand advanced spatial understanding; for ASR, pretrained single-channel Phi-4 attains lower WER when prompted to only transcribe single speakers, whereas *Sci-Phi*’s unified scene description (including more attributes and potentially up to four speakers) yields slightly higher single-speaker WER, despite beating monaural and spatial

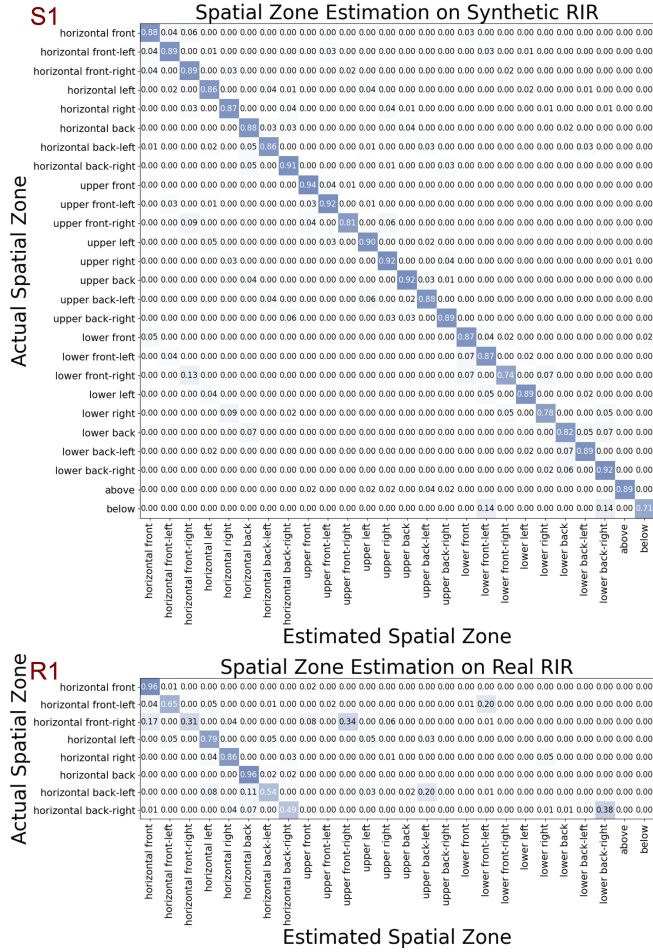


FIGURE 3. Confusion matrices of single-source localization from the synthetic-RIR (**S1**) and real-RIR (**R1**) test sets. Note that we only show a shorter confusion matrix **R1**, because all source elevations of FOA-MEIR are within $[-22.5^\circ, 22.5^\circ]$ (horizontal label by elevation thresholds), although *Sci-Phi* was trained for and may predict all elevations (horizontal, upper, or lower).

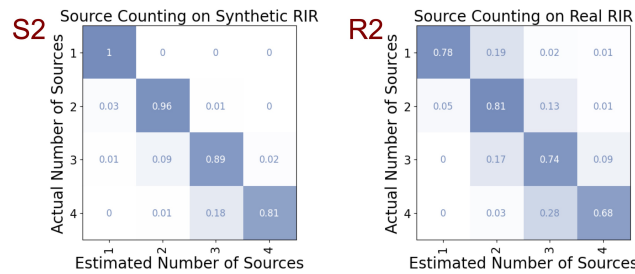


FIGURE 4. Confusion matrices of source counting from the synthetic-RIR (**S2**) and real-RIR (**R2**) test sets.

baselines finetuned on the same speech and audio corpus. It is interesting that WER is slightly lower for the real RIR test set than for the synthetic test set, although the speech source data is the same in both datasets.

Localization Figure 3 **S1** and **R1** show the confusion matrix for *Sci-Phi* on localization for synthetic and real RIRs. We can see that in synthetic RIRs, *Sci-Phi* localizes sound events with minimal confusion. The confusion matrix for real RIRs is only shown for the horizontal plane, as the test set RIRs from FOA-MEIR contain almost no directions above or below our elevation threshold outside the horizontal zone ($\pm 22.5^\circ$). While there is a mild localization accuracy drop visible compared to synthetic RIRs, we can see that confusions are still rare, and most errors are introduced by confusions into adjacent elevation zones (upper/lower etc.).

Source Counting Figure 4 **S2** and **R2** shows source counting confusion on synthetic and real RIRs test sets. It is notable that the strongest confusions happen by mis-estimating only ± 1 source, while larger errors are rare. Manual inspection of examples revealed that many source counting errors actually come from two reasons: 1) the model splits a source (coming from a single audio file) into two events, for example a speech file with a pause, or someone kicking a ball into a window, which may be labelled as one event, but is actually two acoustical events (kicking and glass breaking). 2) the model misses a source which can be hardly audible or masked by other sounds because of too low loudness.

Environmental Robustness. We analyze and demonstrate the robustness of *Sci-Phi* under various challenging environments as shown in Figure 5. **A–F**. As the SNR decreases (**A**), all metrics, including source recognition (CLAP), temporal alignment (IoU), and localization, degrade *monotonically*. Most metrics likewise worsen as reverberation (RT60) grows (**B**). A notable exception is *distance*, *loudness*, and *C50* estimation: performance peaks under *moderate* reverberation, consistent with human psychophysics [33] that modest reflections stabilize distance and clarity cues. Even so, direction accuracy and time IoU remain comparatively high, and zone and distance errors increase only mildly for both lower SNR and higher reverberation.

In **C–E**, *Sci-Phi* continues to detect two distinct sources even when they are similar in *what* (class label), *where* (direction), or *when* (time interval). Two sources of the exactly same class labels leaves metrics nearly unchanged, and as expected, CLAP scores are higher (better) because predicting the same label to both sources is easier. (**C**). Reducing angular separation (**D**) or increasing temporal overlap (**E**) causes small drops in localization (direction, distance) and acoustic estimates (loudness, C50), yet direction accuracy remains near 80% and both CLAP and time IoU stay consistent; We only observe a large performance drop when two sources overlap in time almost 100% (IoU=1). This robustness arises from complementary spatial and spectral (mono) encoders: when two sources are similar along one axis, the other axis can provide discriminative cues.

Lastly, we examine performance versus source duration in **F**. *Sci-Phi* is able to recognize short ($\sim 1s$) sources, maintaining reliable type (CLAP) and timing (IoU) estimation

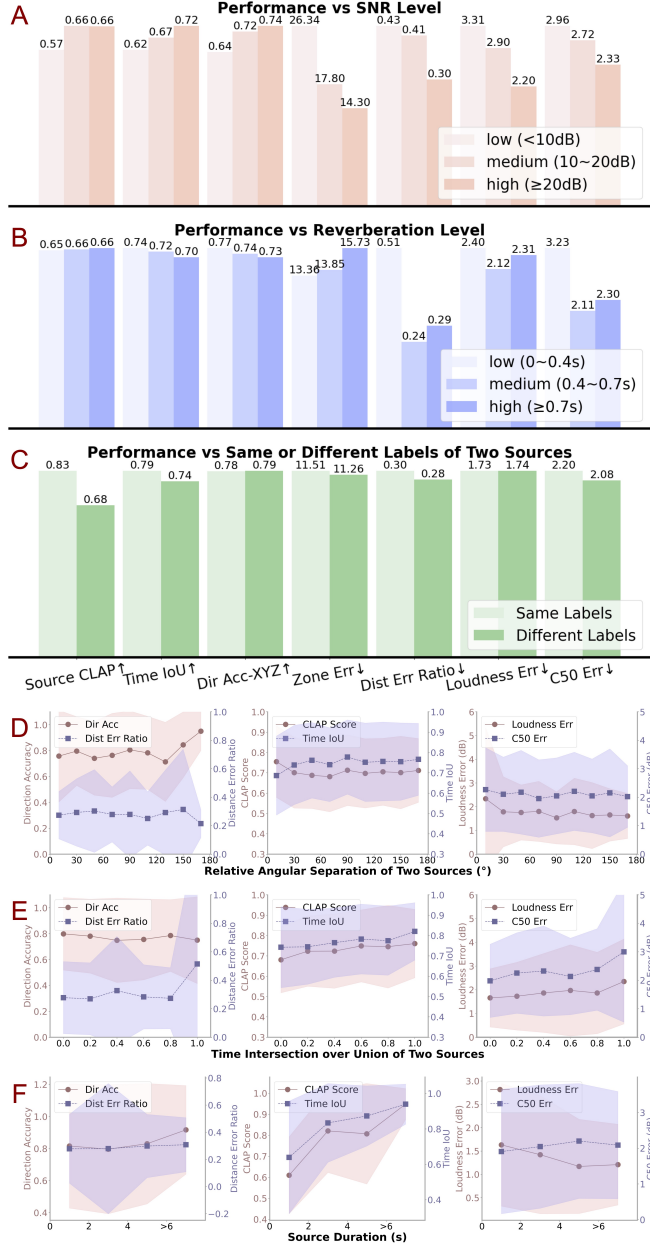


FIGURE 5. Environmental robustness of *Sci-Phi* on the real-RIR test set. A–B: expected behavior across decreasing SNR and increasing reverberation; C–E: consistent performance even when sources are acoustically, spatially, or temporally close; F: reliable recognition of brief (~1 s) sources, with longer durations providing added context. (The marker position and the shaded area correspond to mean $\pm$ std in D–F.)

with competitive direction accuracy, while longer sources naturally provide more context that further lifts these metrics.

Ablations. Table 3 confirms that source enumeration order affects the multi-source description: asking the LLM to order by a given attribute implicitly requires it to be able to estimate that attribute first. Ordering by *loudness* offers the strongest source description performance, yielding the largest number of metric bests. Ordering by zone, distance, or onset

is close behind, whereas ordering by name performs worst because open-vocabulary labels introduce ambiguity. Table 4 zeros either the spatial or spectral features to ablate on their roles. Spatial features determine counting and localization: with spatial-only, count accuracy stays high (87.1%) and 3-D direction accuracy remains strong (81.7%); removing collapses performance toward guessing (25.0% and 9.1%). Meanwhile, spectral features determine ASR. Notably, combining both features boosts performance even on metrics where a single feature excels.

VI. Scalability and Extensibility

Source Count Scaling. While we have only reported *Sci-Phi*’s performance up to 4 sources due to only 4 RIR locations per room from the real-RIR (FOA-MEIR) test set, this is *not a ceiling* of the model capability. We additionally generate 400k/1111h 5-source mixtures (in total 2M/5555h of 1–5 source mixtures) and train a new *Sci-Phi* up to 5 sources. In Figure 6, we evaluated five model checkpoints trained with a maximum of 1–5 sources on the synthetic-RIR test set. Across all metrics, training on more and evaluating on fewer sources matches or outperforms checkpoints trained only up to that count. With an expected minor performance drop on more sources, *Sci-Phi* still performs strongly up to 5 sources, demonstrating scalability to denser auditory scenes.

Spatial Q&A. While *Sci-Phi* is trained to output meta-data for the entire soundscape; users may want targeted information about a specific source or direction. To support this, we further finetune *Sci-Phi* for one additional epoch with diverse Q&A prompts, enabling flexible queries such as direction→label, label→direction, time→label, etc. Several examples are provided on our demo page. When multiple or no sources exist in a queried direction, or when a label appears in multiple directions, *Sci-Phi* returns all matches or reports absence.

VII. Conclusion and Limitations

We introduce *Sci-Phi*, a spatial audio LLM that narrates complete acoustic scenes by jointly modeling spatial and spectral features and generating structured metadata for sources, background, and room acoustics. Trained on synthetic first-order Ambisonics data, *Sci-Phi* generalizes to realistic acoustic conditions with modest performance degradation. Further analysis demonstrates *Sci-Phi*’s robustness and behavior across acoustic conditions such as SNR, reverberation, and the number of directional sound sources. We also propose a consistent evaluation protocol that considers arbitrary ordering of sources and metrics from multiple aspects. This work moves audio foundation models from isolated object recognition toward coherent, multi-attribute spatial scene and object understanding with potential for real-world applications.

Limitations: Our study has not included systematic evaluation on *in-the-wild* recordings, largely due to the absence of reliable ground truth and mismatches in available label

TABLE 3. Ablation on the source enumeration order. Results from the synthetic-RIR test set are averaged across 1-4 sources. Only for this ablation, models were trained with 45% (2k hours) data subset. Environmental metrics are colored in gray, for which the source enumeration order has little effect.

Order By	RoomVol ErrLog2	RT60 Err (s)	Noise CLAP	Count Accuracy (%)	Tuple Score	Source CLAP	WER	Direction Accuracy (XYZ XY Z, %)	Zone Err (°)	Distance ErrRatio	Time IoU	Loudness Err (dB)	C50 Err (dB)
Zone	0.639	0.104	0.667	88.1	0.797	0.665	0.314	77.1 82.0 88.5	11.1	0.279	0.802	1.587	1.606
Distance	0.632	0.104	0.668	87.5	0.800	0.666	0.323	80.7 83.5 90.9	9.5	0.284	0.803	1.613	1.621
Name	0.633	0.104	0.666	87.7	0.796	0.658	0.332	80.5 83.9 90.7	9.5	0.278	0.803	1.665	1.621
Onset	0.632	0.105	0.670	87.7	0.800	0.665	0.315	80.3 84.0 90.6	9.7	0.278	0.804	1.624	1.608
Loudness	0.633	0.103	0.665	87.6	0.798	0.664	0.329	80.8 83.2 91.0	9.3	0.275	0.803	1.579	1.579

TABLE 4. Distinct and combined roles of the spatial and spectral features and encoders. Results from the synthetic-RIR test set are averaged across 1–4 sources.

Features	RoomVol ErrLog2	RT60 Err (s)	Noise CLAP	Count Accuracy (%)	Tuple Score	Source CLAP	WER	Direction Accuracy (XYZ XY Z, %)	Zone Err (°)	Distance ErrRatio	Time IoU	Loudness Err (dB)	C50 Err (dB)
Spatial Only	0.564	0.097	0.647	87.1	0.699	0.553	1.189	81.7 84.3 91.0	9.5	0.269	0.760	1.626	1.507
Spectral Only	0.868	0.654	0.631	25.0	0.562	0.285	0.464	9.1 15.8 46.5	75.5	0.417	0.589	7.893	6.460
Spatial+Spectral	0.590	0.092	0.662	91.5	0.783	0.674	0.449	82.9 85.1 92.0	8.4	0.258	0.802	1.253	1.348

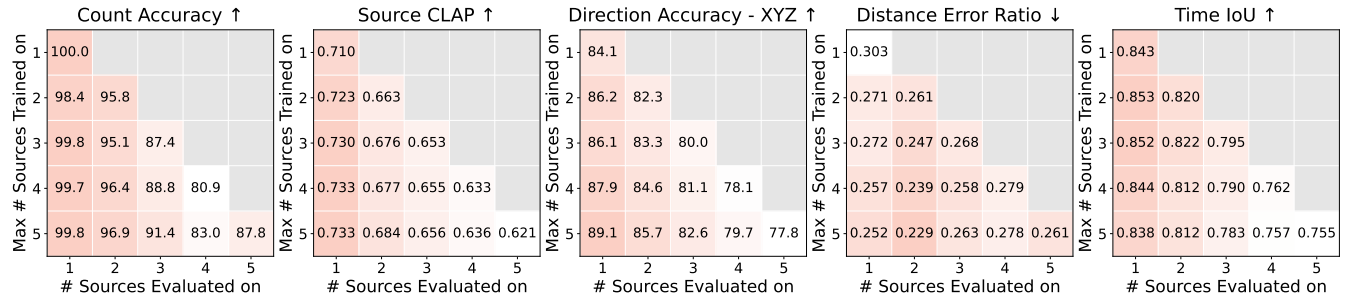


FIGURE 6. Sci-Phi is scalable to the number of sources: still performs strongly up to 5 sources, and preserves performance on fewer sources.

taxonomies. The current framework also assumes stationary sources; generating per-timestep trajectories with an LLM would be computationally very expensive. We leave these directions to future work.

REFERENCES

- [1] Jens Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, 10 1996.
- [2] Andrew J. Kolarik, Brian C. J. Moore, Pavel Zahorik, Silvia Cirstea, and Shahina Pardhan, “Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss,” *Attention, Perception & Psychophysics*, vol. 78, pp. 373 – 395, 2015.
- [3] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [4] Annamaria Mesaros, Sharath Adavanne, Archontis Politis, Toni Heittola, and Tuomas Virtanen, “Joint measurement of localization and detection of sound events,” in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337.
- [5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.
- [6] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Ehsan Variiani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, Ananya Misra, and Chanwoo Kim, “Multichannel signal processing with deep neural networks for automatic speech recognition,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.
- [7] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass, “Listen, think, and understand,” in *Intl. Conf. on Learning Representations*, 2024.
- [8] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2311.07919*, 2023.
- [9] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath, “Bat: Learning to reason about spatial sounds with large language models,” in *International Conference on Machine Learning*, 2024, pp. 61454–61469.
- [10] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, et al., “Can large language models understand spatial audio?,” in *Proc. Interspeech 2024*, 2024, pp. 4149–4153.
- [11] Ayushi Mishra, Yang Bai, Priyadarshan Narayanasamy, Nakul Garg, and Nirupam Roy, “Sing: Spatial context in large language model for next-gen wearables,” in *Forty-second International Conference on Machine Learning*.

- [12] Abdelrahman Abouelenen, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al., "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.
- [13] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.
- [14] Sooyoung Park, Youngho Jeong, and Taejin Lee, "Self-attention mechanism for sound event localization and detection," in *DCASE2021 Challenge — Techn. Reports*, 2021, pp. 1–4, Task 3: Sound Event Localization and Detection with Directional Interference.
- [15] Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji, "Multi-acccdo: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *International Conference on Acoustics, Speech and Signal processing (ICASSP)*. IEEE, 2022, pp. 316–320.
- [16] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al., "Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Adv. in neural information proc. systems*, vol. 36, pp. 72931–72957, 2023.
- [17] Xilin Jiang, Cong Han, Yinghao Aaron Li, and Nima Mesgarani, "Exploring self-supervised contrastive learning of spatial sound event representation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1281–1285.
- [18] Antoni Dimitriadis, Siqi Pan, Vidhyasaharan Sethu, and Beena Ahmed, "Spatial hubert: Self-supervised spatial speech representation learning for a single talker from multi-channel audio," *arXiv preprint arXiv:2310.10922*, 2023.
- [19] Goksenin Yuksel, Marcel van Gerven, and Kiki van der Heijden, "General-purpose audio representation learning for real-world sound scenes," *arXiv preprint arXiv:2506.00934*, 2025.
- [20] Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia, "Learning spatially-aware language and audio embeddings," *Adv. in Neural Information Proc. Systems*, vol. 37, pp. 33505–33537, 2024.
- [21] Jinbo Hu, Yin Cao, Ming Wu, Feiran Yang, and Jun Yang, "SALM: Spatial audio language model with structured embeddings for understanding and editing," 2025.
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *Intl. Conf. on Machine Learning*, 2021, pp. 8748–8763.
- [23] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang, "Natural language supervision for general-purpose audio representations," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 336–340.
- [24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [25] Masahiro Yasuda, Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, and Keisuke Imoto, "Sound event localization based on sound intensity vector refined by dnn-based denoising and source separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 651–655.
- [26] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.
- [27] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
- [28] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.
- [29] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.
- [30] Masahiro Yasuda, Yasunori Ohishi, and Shoichiro Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 226–230.
- [31] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, pp. 3, 2022.
- [32] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Intl. Conf. on Learning Representations*.
- [33] John S Bradley, Hiroshi Sato, and Michel Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoust. Soc. of America*, vol. 113, no. 6, pp. 3233–3244, 2003.