Spatial LibriSpeech: An Augmented Dataset for Spatial Audio Learning

Miguel Sarabia, Elena Menyaylenko, Alessandro Toso, Skyler Seto, Zakaria Aldeneh, Shadi Pirhosseinloo, Luca Zappella, Barry-John Theobald, Nicholas Apostoloff, Jonathan Sheaffer

Apple
miguelsdc@apple.com

Abstract
We present Spatial LibriSpeech, a spatial audio dataset with over 650 hours of 19-channel audio, first-order ambisonics, and optional distractor noise. Spatial LibriSpeech is designed for machine learning model training, and it includes labels for source position, speaking direction, room acoustics and geometry. Spatial LibriSpeech is generated by augmenting LibriSpeech samples with 200k+ simulated acoustic conditions across 8k+ synthetic rooms. To demonstrate the utility of our dataset, we train models on four spatial audio tasks, resulting in a median absolute error of 6.60° on 3D source localization, 0.43m on distance, 90.66ms on T30, and 2.74dB on direct-to-reverberant ratio estimation. We show that the same models generalize well to widely-used evaluation datasets, e.g., obtaining a median absolute error of 12.43° on 3D source localization on TUT Sound Events 2018, and 157.32ms on T30 estimation on ACE Challenge.

1. Introduction
Humans can infer a great amount of information from their acoustic environment, including determining the spatial location of sound sources, estimating the size of a room, and estimating the room reverberance [1,2]. To develop machine learning algorithms with the same level of acoustic spatial awareness, we require a large and diverse dataset with annotations for each task. To the best of our knowledge, no publicly-available dataset has the scale, and diversity needed to train general models for multiple spatial audio tasks. This lack of large-scale datasets limits the applicability of modern machine learning techniques to spatial audio.

In this paper, we introduce Spatial LibriSpeech, a spatially augmented version of LibriSpeech [3] with optional noise from the Microsoft Deep Noise Suppression Challenge 2021 [4]. Spatial LibriSpeech augments the LibriSpeech and Microsoft Deep Noise Suppression samples by simulating how they would be perceived by a microphone array in various synthetic rooms. Spatial LibriSpeech contains over 650 hours of spatial audio with labels for source position, speaking direction, room acoustics, and room geometry (refer to Section 2 for a review of spatial audio tasks for which these labels are useful). Our goal is for Spatial LibriSpeech to be the main training dataset for spatial audio applications.

To create Spatial LibriSpeech, we first generated 8,952 synthetic rooms, which were used to obtain room impulse responses (RIRs) that were convolved and scaled with LibriSpeech samples. The RIRs are modeled on the Zylia microphone array, a 19-channel spherical microphone array. This array facilitates a means for extracting a third-order ambisonics representation up to a frequency of 3080Hz [5]. For completeness, we also provide synthesized, full-bandwidth first-order ambisonics [6], which are aliasing free. Either of these formats may be used to simulate a wide variety of arrays [7]. We describe the dataset generation process in more detail in Section 3.1. A key advantage of convolving existing speech samples with synthetic RIRs for spatial audio dataset generation is that we can ensure our dataset spans a variety of acoustic room properties and room sizes. For instance, full-band T30 values range from 145ms to 2846ms, and room floor area ranges from 13.3m² to 277.4m². Section 3.2 describes diversity statistics of Spatial LibriSpeech.

While Spatial LibriSpeech has the potential to be used to train models for multiple tasks, in this paper we focus on some of the most fundamental spatial audio detection tasks: i) 3D source localization, ii) source distance, iii) third-octave narrow-band direct-to-reverberant ratios (DRRs), and iv) third-octave narrow-band T30s. All models share the same architecture and training regime (presented in Section 4). When training spatial audio models, there is a choice of using the microphone array signals as inputs or converting the microphone array signals to ambisonics to obtain device-agnostic models. In our work, models consume first-order ambisonics.

Our models achieve a median absolute error of 6.60° in 3D source localization, 0.43m in distance estimation, 2.74dB in DRR estimation, and 90.66ms in T30 estimation. Since our training dataset is composed of simulated acoustics, we verify the transferability of our models by testing them on two evaluation datasets: TUT Sound Events 2018 [8], and ACE Challenge [9] (Section 4.1). We find that fine-tuning is beneficial for ACE Challenge but unnecessary for TUT Sound Events 2018, and we plot the responses of our models on both evaluation datasets showing that ACE responses are at the tail of the Spatial LibriSpeech responses (see Section 4.2). Lastly, we show that, for single-task models, using a version of Spatial LibriSpeech made by uniformly sampling 10% of the dataset yields the same performance as the full dataset (cf. Section 4.3).

2. Background
We start by reviewing spatial audio tasks to highlight the large number of potential use cases for Spatial LibriSpeech. We divide audio tasks into three categories: source parameter estimation, environment parameter estimation, and spatial processing. Source parameter estimation tasks are concerned with understanding audio sources. Examples include estimation of: source localization [25], source distance [26], and speaking direction [27]. Environment parameter estimation tasks involve understanding the environment where the audio is produced.
Table 1: Comparison of existing datasets. ROOMS refers to the number of unique acoustic environments in the dataset, ROOM CONFIG. to the number of physical configurations per environment (e.g. different source receiver positions). CHANNELS to the availability of spatial information from a microphone array or other encoding; N - N channels mic array, B - Binaural, Xth OA - Xth order ambisonics. DATA-TYPES: sIR - Simulated IR, rIR - Recorded IR, sA - Simulated Audio, rA - Recorded Audio. LABELS: P - Position, SD - Speaking direction, R - Room acoustics, G - Room geometry, O - Other.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>ROOMS</th>
<th>ROOM CONFIG.</th>
<th>CHANNELS</th>
<th>DATA-TYPES</th>
<th>LABELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>dEchorate [10]</td>
<td>11</td>
<td>180</td>
<td>5</td>
<td>sIR, rIR</td>
<td>P</td>
</tr>
<tr>
<td>Amii [11]</td>
<td>1</td>
<td>21</td>
<td>3\textsuperscript{rd} OA, 4\textsuperscript{th} OA</td>
<td>sIR</td>
<td>P, R</td>
</tr>
<tr>
<td>GIR [12]</td>
<td>1</td>
<td>2.951</td>
<td>1</td>
<td>rIR</td>
<td>R, G</td>
</tr>
<tr>
<td>CoupledRooms [14]</td>
<td>2</td>
<td>101</td>
<td>4\textsuperscript{th} OA</td>
<td>rIR</td>
<td>P, R, G</td>
</tr>
<tr>
<td>DCASE2021 Task 3 [15]</td>
<td>13</td>
<td>1,184-6,480</td>
<td>4, 4\textsuperscript{th} OA</td>
<td>rIR, rA</td>
<td>P</td>
</tr>
<tr>
<td>BUT ReverbDB [16]</td>
<td>8</td>
<td>155</td>
<td>1</td>
<td>rIR</td>
<td>P, R, G</td>
</tr>
<tr>
<td>SBSBRIR [17]</td>
<td>1</td>
<td>180 x 2\textsuperscript{nd}</td>
<td>B</td>
<td>rIR</td>
<td>P</td>
</tr>
<tr>
<td>BIRD [18]</td>
<td>12,500</td>
<td>8</td>
<td>2</td>
<td>sIR</td>
<td>P</td>
</tr>
<tr>
<td>Kemar BIRRs [19]</td>
<td>1</td>
<td>50</td>
<td>B</td>
<td>rIR</td>
<td>P, G</td>
</tr>
<tr>
<td>Motus [20]</td>
<td>1</td>
<td>3,320</td>
<td>32, 4\textsuperscript{th} OA</td>
<td>rIR, rA</td>
<td>P, R, G</td>
</tr>
<tr>
<td>ACE Challenge [9]</td>
<td>7</td>
<td>10</td>
<td>2, 32</td>
<td>rIR</td>
<td>P, R</td>
</tr>
<tr>
<td>DHBRA [22]</td>
<td>2</td>
<td>62</td>
<td>3</td>
<td>rA, sA</td>
<td>P, G</td>
</tr>
<tr>
<td>Voice-Home [23]</td>
<td>12</td>
<td>24</td>
<td>8</td>
<td>rIR, rA</td>
<td>P, G</td>
</tr>
<tr>
<td>Sween-Home [24]</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>rA</td>
<td></td>
</tr>
<tr>
<td>Microsoft DNS 2001 [4]</td>
<td>14,576</td>
<td>1</td>
<td>1</td>
<td>sIR, rIR</td>
<td>R</td>
</tr>
<tr>
<td>TUT Sound Events 2018 [8]</td>
<td>5</td>
<td>487-4,366</td>
<td>8, 4\textsuperscript{th} OA</td>
<td>sA, rA</td>
<td>P, G</td>
</tr>
</tbody>
</table>

Spatial LibriSpeech 8,952 20 19, 4\textsuperscript{th} OA sA P, SD, R, G

Example tasks include estimation of DRR [28], material absorption and scattering [29], and room volume [30]. Finally, spatial processing tasks involve the transformation of audio signals with information extracted from the acoustic environment. Example tasks include de-reverberation [31], beamforming [32], and audio source separation [33].

Table 1 contains a summary of the main differences between 18 published spatial audio datasets in terms of the number of environments, number of physical configurations per environment, recorded/simulated channels, and data-types and labels included in the dataset. We found only two datasets that include over 50k unique configurations across all environments: BIRD [18] and DCASE2021 Task 3 [15]. However, BIRD features only two microphone channels, and DCASE2021 Task 3 includes only position labels. Spatial LibriSpeech is the only dataset to feature over 200k unique configurations, labels for a large number of acoustic tasks, and both mic-domain audio and ambisonics. This lack of diversity and labels limits the applicability of existing datasets to modern machine learning techniques [34] such as multi-task learning or contrastive representation learning [35].

3. The Spatial LibriSpeech Dataset

In this section, we describe the generation of Spatial LibriSpeech and its defining characteristics.

3.1. Generation Methodology

Our pipeline to generate Spatial LibriSpeech consists of three steps: parametric room generation, room impulse response simulations, and mixing.

Parametric room generation. We start by defining a realistic set of conditions for commonly encountered living spaces following [36]. These consist of a predefined distribution of reverberation times and room shapes and sizes. The materials associated with the room surfaces are chosen from a database of typical construction materials, for which the absorption and scattering coefficients as a function of frequency are available.

Room impulse response simulations. We use a geometrical acoustic solver, which includes a high fidelity model of the room that accounts for the directivity of the microphones (including diffraction effects from the array body). The room is populated with several acoustic objects including the recording device placed at a randomized but bounded position, several sources surrounding the device with a randomized looking direction and directivity function. When sources are intended to represent speech, their directivity is computed using a boundary element code using a set of artificially generated models. The direct path from each source to the microphone array, the early reflections, and the late reverberations are used to assemble each impulse response.

Mixing. Once we have computed RIRs for each configuration of the simulated rooms, we mix them with LibriSpeech samples by randomly assigning them to sources in every simulated room. We then remove leading and trailing silence from each sample\(^2\), convolve the sample with the selected RIR and scale the output to a random active speech level (ASL) between 85dB-ASL and 100dB-ASL at mouth reference point\(^3\). Finally, we save the resulting audio samples and store the relevant labels from both the room simulation and LibriSpeech. For the distractor noise we follow the same process, except we assign a random source, different to the main signal source, and set the signal strength to a random signal-to-noise ratio (SNR) between 10dB and 40dB.

3.2. Dataset Characteristics

The main characteristics of Spatial LibriSpeech are summarized in Table 2. The dataset consists of two splits: TRAIN which is derived from LibriSpeech’s train-clean-100 and train-clean-360 subset, and TEST which is derived from test-clean. The audio is sampled at 16kHz.

Spatial LibriSpeech includes the following labels: source localization (azimuth and elevation), speaking direction, room volume, surface and floor area, voice directivity identifier, narrow-band C50, DRR, EDT, T20, and T30 [38]; and all

\(^2\)Silence is defined as any leading or trailing sound before a 100ms segment at more than 1% volume.

\(^3\)Mouth reference point is the point on the reference axis 25 mm in front of the lip plane ITU-T P.581 [37]
original metadata from LibriSpeech, including ids for speaker, books, and chapters, and the reference transcriptions.

We verify that our models do not overfit to the virtual rooms in the training set by checking that the performance of models trained with Spatial LibriSpeech transfer to two existing evaluation datasets (see Section 4.1).

### 4. Training with Spatial LibriSpeech

To demonstrate the utility of Spatial LibriSpeech in real-world situations, we trained several neural networks for: i) 3D source localization, ii) source distance, iii) DRR, and iv) T30. We chose these tasks since they are representative of the source and environment parameter estimation spatial audio tasks and because public third-party evaluation datasets were available to analyse transferability.

For 3D source localization, we report the median absolute error of the 3D angle (cf. [39]):

\[
\alpha = \cos^{-1}\left(\frac{1}{2}\sin(\phi)\sin(\hat{\phi}) + \cos(\phi)\cos(\hat{\phi})\cos(\theta - \hat{\theta})\right),
\]

where \(\phi\) and \(\theta\) are the ground-truth azimuth and elevation respectively, and \(\hat{\phi}\) and \(\hat{\theta}\) are the predicted azimuth and elevation. For distance regression, we report the median absolute error in meters. As DRR and T30 are 20-dimensional vectors representing the third-octave frequencies between 100Hz and 8kHz, we report both the median absolute error as well as the Pearson correlation across frequencies.

Our objective is to show that models trained with Spatial LibriSpeech transfer to two established evaluation datasets: ACE Challenge [9] and TUT Sound Events 2018 [8]. All models share the same architecture. We first transform and normalize segments of spatial audio encoded as 4-channel first-order ambisonics into active and reactive components following [40].

The active and reactive components are fed through two independent branches of four 3D-convolutional layers each (of 2, 4, 8 and 16 channels respectively), with max-pool, batch-norm and exponential linear units between the convolutional layers. Next, the output of both branches is flattened and concatenated, and fed to a 3-layer multi-layer perceptron (MLP). The number of outputs of the last layer depends on the task: two for source localization (one for azimuth, and another for elevation), one for distance regression, 20 for DRR, and 20 for T30. Our model has 101,933 parameters (for 3D source localization and distance) and 747,148 parameters (for T30 and DRR). Models are trained for 20 epochs for 3D source localization and distance, and 50 epochs for T30 and DRR.

We use the Adam optimizer, a weight decay of 0.01, dropout of 50% on the output of the convolutional block, Kaiming uniform initialization, and a learning rate of \(10^{-4}\). Our models completed training in a median 1h15min for 3D Source Localization and Distance Estimation, and 20h for T30 and DRR on a 80-core Intel Xeon CPU with 8 NVIDIA Tesla V100s GPUs, and 375GiB of RAM.

### Table 2: Main characteristics of Spatial LibriSpeech.

<table>
<thead>
<tr>
<th>Total duration</th>
<th>TRAIN</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>573h 13m 12s</td>
<td>85h 29m 20s</td>
<td></td>
</tr>
<tr>
<td>Speech samples</td>
<td>171,951</td>
<td>49,505</td>
</tr>
<tr>
<td>Simulated rooms</td>
<td>8,952</td>
<td>4,970</td>
</tr>
<tr>
<td>Source azimuth</td>
<td>([-180.0^\circ, +180.0^\circ])</td>
<td>([-180.0^\circ, +180.0^\circ])</td>
</tr>
<tr>
<td>Source elevation</td>
<td>([-48.1^\circ, +48.7^\circ])</td>
<td>([-49.5^\circ, +42.4^\circ])</td>
</tr>
<tr>
<td>Source distance</td>
<td>([0.5m, 4.0m])</td>
<td>([0.9m, 4.0m])</td>
</tr>
<tr>
<td>Speaking azimuth</td>
<td>([-180.0^\circ, +180.0^\circ])</td>
<td>([-180.0^\circ, +180.0^\circ])</td>
</tr>
<tr>
<td>Speaking elevation</td>
<td>([-89.3^\circ, +87.7^\circ])</td>
<td>([-74.5^\circ, +68.2^\circ])</td>
</tr>
<tr>
<td>Voice directivities</td>
<td>16 variations</td>
<td>1 variation</td>
</tr>
<tr>
<td>Room floor area</td>
<td>([13.3m^2, 277.4m^2])</td>
<td>([14.3m^2, 277.4m^2])</td>
</tr>
<tr>
<td>Full-band T30</td>
<td>([144.5ms, 2846.0ms])</td>
<td>([157.8ms, 1267.7ms])</td>
</tr>
</tbody>
</table>

*The duration of the segments varies by task, for 3D source localization and distance we use 0.5s, for T30 and DRR we use 4.0s.*

**This represents 80.8M gradient updates for 3D source localization and distance, and 21.2M gradient updates for T30 and DRR estimation.**
Table 4: Performance contextualization for different tasks. Random refers to the performance of the model described in Section 4 without any training. S.L.S indicates the performance of our baseline (see Table 3). Ext. refers to external benchmarks. The comparison with external benchmark may not be fair, as they were trained with different architectures, training regimes, and datasets closer to the evaluation benchmark. *Model was fine-tuned on ACE Dev.

<table>
<thead>
<tr>
<th>TASK</th>
<th>EVAL.</th>
<th>METRIC</th>
<th>RANDOM</th>
<th>S.L.S.</th>
<th>EXT.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Sec.</td>
<td>ANSIM [8]</td>
<td>median abs.</td>
<td>88.49°</td>
<td>8.10°</td>
<td>~4.3° [39]</td>
</tr>
<tr>
<td>Local</td>
<td>RESIM [8]</td>
<td>3D angle</td>
<td>88.55°</td>
<td>5.80°</td>
<td>~4.3° [39]</td>
</tr>
<tr>
<td></td>
<td>REAL [8]</td>
<td>error (eq. 1)</td>
<td>88.53°</td>
<td>12.43°</td>
<td>~4.3° [39]</td>
</tr>
<tr>
<td>T30</td>
<td>ACE [9]</td>
<td>mean error</td>
<td>-7.12±43ms</td>
<td>156.4±43ms</td>
<td>22.1±43ms [28]</td>
</tr>
<tr>
<td>DRR</td>
<td></td>
<td>source localization (Figure 1a) yields overlapping representations for Spatial LibriSpeech and TUT Sounds Events 2018, while the model trained on DRR (Figure 1b) yields representations on ACE at the tail of the Spatial LibriSpeech representations indicating the need for a small amount of fine-tuning of the MLP block on the target data.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1. Performance and transferability

Table 3 shows the performance of our models on all four tasks when evaluated against the test set of Spatial LibriSpeech, TUT Sounds Events 2018 [8], and ACE Challenge Eval set [9]. All models were trained using Spatial LibriSpeech only.

Our results show that the difference in performance between the test set of Spatial LibriSpeech and the performance on TUT Sound Events 2018 is small, +5.83° and +0.17ms worse on REAL than Spatial LibriSpeech, though better on RESIM than on Spatial LibriSpeech (-0.90° and -0.18ms). We find that our models do not exploit spurious correlations between the speech level and the distance labels, instead incorporating information from reverberations to make accurate predictions. This is evidenced by the median absolute error of ANSIM (an anechoic subset of TUT Sound Events 2018) being 4.76ms.

Looking at the performance differences between Spatial LibriSpeech and ACE for DRR and T30 regression, we find that the transfer penalty is higher (+11.76dB for DRR, and +66.66ms for T30). We also investigate fine-tuning the last three layers (the MLP) with the ACE Challenge Dev set for 20 epochs, and find that the median absolute error improves significantly for both T30 and DRR (resulting in 2.07dB and 14.91ms gaps with the Spatial LibriSpeech test set).

For additional context into these results, Table 4 compares the performance of our models against the performance of a randomly initialized version of our model as well as the reported performance of recent benchmarks [28, 39]. In general, benchmarks obtain better results, as they use more complex architectures and train on data specifically modeled after the evaluation sets. We hypothesize that the performance of our model is close to the benchmarks due to the diversity of Spatial LibriSpeech, since we did not carry out an architecture or hyper-parameter search. Furthermore, note that none of these benchmarks published their training sets. In contrast, LibriSpeech was primarily designed for training spatial audio models.

4.2. Visualization of Dataset Representations

Plotting the representations from our models trained on Spatial LibriSpeech further illustrates the transferability to real-world test data seen in Section 4.1. Figure 1 shows UMAP [42] plots of embeddings extracted from representations before the MLP block of the network. We see that the model trained for 3D embeddings are constructed by first partitioning the train and test sets into 50 batches of data, then sampling 10,000 samples from each source localization (Figure 1a) yields overlapping representations for Spatial LibriSpeech and TUT Sounds Events 2018, while the model trained on DRR (Figure 1b) yields representations on ACE at the tail of the Spatial LibriSpeech representations indicating the need for a small amount of fine-tuning of the MLP block on the target data.

4.3. Training with a smaller subset of Spatial LibriSpeech

We explore the performance of models trained with just 10% of Spatial LibriSpeech, which we sample uniformly, maintaining the diversity of the dataset. We find that these models tend to perform worse on the test set of Spatial LibriSpeech than models trained with the full dataset, for instance the best median absolute error on 3D sound localization is 0.39° higher, and 1.36ms higher for T30 regression on the 10% models. However, when looking at performance on external baselines, models trained with 10% of the training data tend to perform better, for example the best median absolute error on 3D sound localization on TUT Sound Events 2018 was 0.88° lower, and the median Pearson correlation on T30 regression on ACE Challenge was 0.12 higher. Still, all performance differences were smaller than the IQR between models. Based on these results, we recommend researchers prototype with the smaller version of the dataset, which will be available as a separate download. Additionally, we release the full Spatial LibriSpeech dataset to enable researchers to explore other tasks, such as ablations of acoustic conditions or representation learning.

5. Conclusion & Further Work

We have presented Spatial LibriSpeech, a spatial audio dataset for multiple spatial audio tasks and representation learning. We have shown the utility of our dataset with a simple convolutional network trained on Spatial LibriSpeech, the performance of which transfers to established baselines with minimal intervention, and that the results are close to the state-of-the-art, despite a less sophisticated architecture.

We intend to use Spatial LibriSpeech for a number of other tasks, such as denoising, or room identification. Another interesting line of research is whether we can further improve performance using representation learning to train a single model to regress to many of the spatial audio tasks with the same embedding. We look forward to the community using Spatial LibriSpeech to accelerate research in spatial audio.
6. References


