LEVERAGING GEOMETRICAL ACOUSTIC SIMULATIONS OF SPATIAL ROOM IMPULSE RESPONSES FOR IMPROVED SOUND EVENT DETECTION AND LOCALIZATION

Christopher Ick *Brian McFee*

Music and Audio Research Laboratory, New York University
Brooklyn, NY 11201, United States

ABSTRACT

As deeper and more complex models are developed for the task of sound event localization and detection (SELD), the demand for annotated spatial audio data continues to increase. Annotating field recordings with 360° video takes many hours from trained annotators, while recording events within motion-tracked laboratories are bounded by cost and expertise. Because of this, localization models rely on a relatively limited amount of spatial audio data in the form of spatial room impulse response (SRIR) datasets, which limits the progress of increasingly deep neural network based approaches. In this work, we demonstrate that simulated geometrical acoustics can provide an appealing solution to this problem. We use simulated geometrical acoustics to generate a novel SRIR dataset that can train a SELD model to provide similar performance to that of a real SRIR dataset. Furthermore, we demonstrate using simulated data to augment existing datasets, improving on benchmarks set by state of the art SELD models. We explore the potential and limitations of geometric acoustic simulation for localization and event detection. We also propose further studies to verify the limitations of this method, as well as further methods to generate synthetic data for SELD tasks without the need to record more data.

Index Terms— Acoustic Simulation, Localization, Data Augmentation

1. INTRODUCTION

Sound event detection and localization (SELD) is the union of two active fields of research; sound event detection (SED), and localization, or direction-of-arrival (DoA) estimation. Expanding the scene-description capabilities of SED with the spatiotemporal characterization of localization sees applications ranging from autonomous robot navigation [1] and urban monitoring [2], to speaker diarization [3] and immersive experiences in virtual and augmented reality devices.

While earlier techniques for SELD have focused on traditional signal-processing or parametric models such as [4, 5, 6, 7], recent literature is dominated by deep neural network (DNN) approaches, which have been shown superior performance in both pure localization [8, 9] as well as joint SELD tasks [10]. A surge of interest in this field can be attributed to the introduction of SELD as a task in the DCASE2019 challenge [11]. This challenge included the release of several 4-channel audio datasets with spatial and temporal annotations for sound events. The audio was generated by convolving sound events with spatial room impulse responses (SRIRs) recorded in 5 separate rooms at 504 unique azimuth-elevation-distance combinations. This was further iterated upon by the SELD challenge in DCASE2020 [12] with 13 unique rooms. Recently, DCASE2022 reintroduced the challenge with hand-annotated real-world recordings in the STARSS22 dataset [13], providing one of the first datasets with real-world data upon which to evaluate SELD models. This dataset uses a combination of 360° video, and motion capture to extract spatiotemporal annotations that were manually validated. In addition to this, the DCASE2022 dataset also included a release of the SRIRs used to generate the training data for the SELD task, as well as the code for the generator itself, allowing users to generate their own annotated spatial data [14]. This data included SRIRs measured over a wide range of positions over 9 different rooms in Tampere University's campus.

These datasets are unique in the density of SRIR measurements across particular paths, the variety of acoustic enclosures, and the large amount of SRIRs. Because of their scale, visibility, and quality, these datasets have become some of the most cited SRIR datasets for DNN-based approaches to SELD, because of their ability to meet the data requirements of these highly-parametrized models. Despite this, these datasets are still severely limited by the recording procedure of SRIRs, which require time, expertise, and a low-noise environment to produce at a high quality. Increasing the spatial density, variety of trajectories types, and number of trajectory paths becomes multiplicatively time consuming to develop. Furthermore, the range of rooms in which these measurements can be recorded is inherently limited by the limitations of the recording facilities, usually limited to a dozen rooms or so in the best of cases. However, without a wide range of acoustic environments to perform these measurements, generalization to a variety of unseen acoustic environments becomes impossible.

Physical acoustic simulations provide an attractive solution to the limitations of field-recorded SRIRs. Acoustic simulation is typically split into two categories: wave based methods, which simulate the propagation of sound waves through physical media, and geometric modeling methods, which model the transportation of acoustic energy through acoustic rays, mimicking popular methods for modeling optical rays. Geometrical acoustics approximate the wavelength of the propagating sound to have wavelength relatively small compared to the room geometries of interest, and neglects wave effects such as diffraction or scattering. Nevertheless because of ease of implementation and computational efficiency, geometrical acoustic modeling methods have seen wide success in several tasks, including modeling architectural acoustics [15] and room parameter estimation [16].

In this work, we propose utilizing one method of geometrical acoustics modeling, the image source method, to generate simulated SRIRs for training DNN models for SELD. We demonstrate

This material is based upon work supported by the National Science Foundation under NSF Award 1922658.

the effectiveness of this simulated method for SRIR generation using the framework and data provided in previous DCASE SELD challenges. By creating an audio dataset from simulated SRIRs, we train a SELD model with similar performance to one utilizing real-world SRIRs. By directly compares simulated SRIRs with a datasets of recorded SRIRs of the similar size, room geometries, and DoA distributions, we demostrate the downstream effects of simulation in place of recording as being relatively minimal, differing our work from prior studies [17]. Furthermore, we augment a typical SRIR dataset with simulated SRIRs, training models that outperform those trained solely on recorded SRIRs. Finally we propose further experiments to explore the use of simulated SRIRs for training SELD models. The code associated with this work is released in an open-source github repository [1] to further work in using synthetic SRIRs for training DNN models.

2. ACOUSTIC SIMULATIONS

2.1. The Image Source Method

The image source method (ISM) is a technique used in architectural acoustics and room modeling to predict the sound field in enclosed spaces [18]. The ISM considers the primary sound source and virtual images reflected by the room's boundaries. These virtual sources are assumed to emit sound with the same magnitude and phase as the primary source, but with a delay due to the additional path length traveled. Typically, this starts with defining the room geometry, including the positions and shapes of the walls, ceiling, and floor. For each reflecting surface, virtual image sources are "mirrored" across the boundary. The number of virtual sources depends on the order of reflections considered. From here, the interaction between the primary sound source and the virtual image sources is calculated by determining the path lengths, time delays, and attenuation factors associated with each source-receiver combination, as well as the material properties of each surface through which the sound path is reflected upon. By summing the contributions of the primary source and its image sources, the sound field at various locations in the room can be predicted, providing an estimation of the sound pressure level, arrival times, and directivity patterns.

It's important to note several limitations of the ISM model. The ISM implicitly assumes that all surfaces are perfectly reflective and flat, with idealized acoustic properties, which fails to account for acoustic effects such as scattering or diffraction. Furthermore, the order of reflections/virtual sources scales the computational cost exponentially, meaning compute late-stage reflections in an SRIR prohibitively expensive.

Despite its limitations, the image source method is widely used due to its computational efficiency and effectiveness in predicting sound fields in enclosed spaces. Because localization is more reliant on direct sounds/early reflections, the limitations caused by use of the the ISM for computing SRIRs can be expected to be relatively minimal.

2.2. The TAU-SRIR Dataset

To validate the use of ISM-generated SRIRs in a direct-comparison, we take the existing TAU-SRIR database [14] as an example database for which well established metrics for SELD have been measured.

[1] https://github.com/ChrisIck/DCASE_Synth_Data

	Azimuth (ϕ)	Elevation (θ)
M1	45°	35°
M2	-45°	-35°
M3	135°	-35°
M4	-135°	35°

Table 1: Microphone Geometry for TAU-SRIR dataset. Each microphone is 4.2cm from the center, and is modeled with a hypercardioid response.

Room Name	Traj. type	N_t	N_h	N_{SRIRs}
Bomb shelter	Circular	2	9	6480
Gym	Circular	2	9	6480
PB132	Circular	2	9	6480
PC226	Circular	2	9	6480
SA203	Linear	6	3	1594
SC203	Linear	4	5	1592
SE203	Linear	4	4	1760
TB103	Linear	4	3	1184
TC352	Circular	2	9	6480

Table 2: Trajectory information for rooms contained in the TAU-SRIR dataset [14]. Each room contains trajectories across a number of trajectory groups (N_t) and a number of heights (N_h), for a total of $N_t \times N_h$ trajectories per room. Each trajectory is sampled in roughly 1° increments.

The TAU-SRIR database contains SRIRs recorded in 9 different rooms throughout Tampere University's campus. Each SRIR is computed by recording a maximum-length sequence (MLS) played through a loudspeaker, recorded on an Eigenmike spherical microphone array. Each SRIR was downsampled to 24kHz and truncated at 300ms, resulting in 7200 samples per RIR. The data is stored in a 4-channel audio corresponding to a tetrahedral microphone array with the geometry in spherical coordinates (ϕ, θ), specified in Table 1. For each room, the position of the microphone array was provided.

SRIRs were measured along either circular or linear traces at fixed distance from the microphone array along the z-axis at a number of trajectory groups, separated by distance and reflection across the axis of the microphone array in the case of linear traces. Circular trajectory groups had a specified radius of orbit, whereas linear trajectories had a specified start and end point in 3D space. Each trajectory was repeated at a number of different heights, and each trajectory had a fixed number SRIR measurements and corresponding DoA measurements recorded as Cartesian components of a unit vector. The number of SRIR measurements vary across different trajectories/heights, spaced in roughly 1° increments. The total number measurements can be seen in Table 2.

2.3. Room Simulation

We recreate this dataset using the python package *pyroomacoustics* [19], a pythonic implementation of the ISM, that has demonstrated use in implementations of various algorithms for beamforming, direction finding, adaptive filtering, source separation, and single channel denoising.

To replicate the acoustic conditions of each of the rooms in the TAU-SRIR dataset, we randomly sampled the RIRs uniformly in each room until we had a sample of 5 single-channel RIRs. Using the Schroeder method [20], we estimated the RT_{60} of each room by

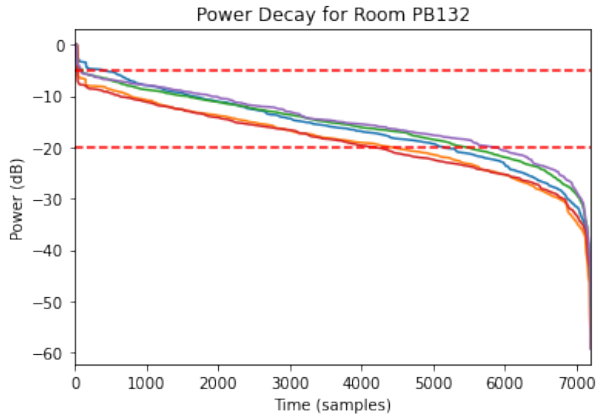

Figure 1: The log-scale energy decay from a sample of RIRs from a singular room. The region in the dashed lines is roughly linear, suggesting it corresponds to mid-late reflections, and is used for an estimate of RT_{60}

hand-selecting the early decay of the energy-decay function of the RIR samples and computing the linear fit. This can be see in in Figure 1. Using the inverse Sabine formula, we used this to estimate the mean absorption coefficient of the rooms and the number of required reflection orders to approximate a room of a similar RT_{60}. We combined these parameters with the geometry estimations from the TAU-SRIR dataset to construct virtual rooms matching those of the 9 rooms in the TAU-SRIR dataset. To this room, we added a virtual tetrahedral microphone with the geometry specified in Table 1, with each virtual microphone using a hypercardioid response pattern centered at the position specified in the TAU-SRIR dataset.

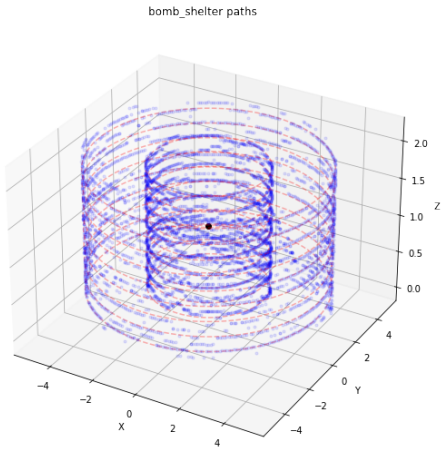

Figure 2: SRIR measurement positions reconstructed from the DoAs provided in the TAU-SRIR dataset (blue), compared to the path specified by the height, radius, and position labels provided (red)

To estimate the positions of the SRIRs in space corresponding to the DoA measurements in the TAU-SRIR dataset, we chose points along the DoA that most closely matched the corresponding path on the line specified by the trajectory in the dataset. Circular datasets specified their height and radius, and were centered along the same z-line as the microphone array. Linear trajectories specified the start point and end points of their traces. These matched points were estimated by projecting the DoA vectors onto a cylinder that matched the radius of the trajectory, after translation by the position of the microphone array and the height of the trajectory of interest (see Figure 2). Once these points were estimated, they were placed into the simulated room in a position relative to the virtual microphone array, and the 4-channel SRIR was computed with the ISM for each point, at the sample rate of 2400kHz. To match the dimensions of the original TAU-SRIR dataset, we truncate the SRIRs to 300ms, providing us with an SRIR with dimensions of 7200×4 for each point.

3. METHODOLOGY

To evaluate the performance of this dataset in SELD tasks, we generated a dataset of audio events consistent with the methodology of dataset generation for training the baseline SELD model in the DCASE 2022 challenge [21]. We generated 3 datasets, one using the original SRIR database, which we will refer to as the TAU-SRIR datset. We also generated a dataset using only the synthetic SRIRs, which we refer to as the SIM-SRIR dataset. Finally, we generated a third dataset that equally samples both the original and simulated SRIRs, which we will refer to as the augmented SRIR dataset, or AUG-SRIR.

3.1. Data Generation

To generate our annotated spatialized audio, we followed the procedure used in DCASE2019-2021, by convolving various sound events with SRIRs.

The sound events were drawn from the FSD50k audioset [22], a subset containing over 20k sound events of 13 classes selected for the DCASE challenge. These sound events were spatialized into virtual recordings, each corresponds to a singular room, allowing for up to three concurrently active sources. The sources can be static or dynamic, with equal probability, and the dynamic sources can move at slow (10°/sec), moderate (20°/sec), or fast (40°/sec) angular speeds. Each sample lasted 60 seconds, 40 of which had at least one active event class.

For each of the 3 SRIR datasets, 1200 recordings were created in separate folds, 900 for training and 300 for validation. The training and validation sets used 6 and 3 rooms respectively, such that none of the same rooms overlapped both folds.

3.2. Model

The model architecture is identical to the one used in the DCASE2022 Task 3 challenge baseline [13]; a SELDnet style CRNN with multi-ACCDOA representation for co-occuring events [23]. The model takes T frames of an STFT time-frequency representation of the multichannel features, and outputs $T/5 \times N \times C \times 3$ vector coordinates, and N is the assumed maximum co-occurring events, in our case 3.

The input features are 4-channel 64-band log-mel spectrograms combined with SALSA-lite spatial features, all of which are truncated to include bins up to 9kHz, without mel-band aggregation following [24]).

Each model was trained on the training folds generated from the SRIR datasets described above. In addition to this, data from the

	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
TAU-SRIR	**0.71**	14.4%	55.1°	39.2%
SIM-SRIR	0.73	13.0%	79.6°	34.8%
AUG-SRIR	0.75	**16.3%**	**52.3°**	**42.3%**

Table 3: The cross-class evaluation metrics for models trained on data generated from different SRIR datasets

Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22) [13] was added for training, using the 54 development sound mixtures for training, but witholding the remaining 52 clips for evaluation of the models, ensuring the results were exclusively on real recorded data from unseen rooms. The models were trained for 100 epochs each,

3.3. Evaluation

Evaluation was completed using join localization-detection metrics established in the DCASE 2020 challenge. The detection metrics used were error-rate and F1 score for a spatial threshold within 20° (ER_{20° and F_{20°). F1 score was macro-averaged to account for class distribution differences in the FSD50k audio subset used. Localization metrics are class dependent localization error and recall (LE_{CD} and LR_{CD}).

4. RESULTS

Despite the coarse physical approximations made by the ISM, the entirely synthetic SRIRs generated with this process performed nearly as well as the SRIRs recorded in real world settings. Furthermore, the dataset of real SRIRs augmented by synthetic SRIRs outperformed both by a narrow margin, showing benefits of geometrical acoustics simulation for data augmentation for SELD tasks.

Regarding the cross-class average performance of the models in Table 3, we can see that for our classification metrics, all three models perform relatively similarly, with a slight performance edge to the AUG-SRIR dataset trained models. Looking into the per-class results in figure 3, we can see that generally, all three models struggle with similar classes (telephone, laughter, door), but the AUG-SRIR dataset outperforms both in certain classes for which both other models perform poorly on (Water tap/Faucet, and Knock).

Looking at the localization based results, it appears that some amount of the performance differences between the SIM-SRIR trained models and the TAU-SRIR trained models can possibly be attributed to model fine-tuning; while SIM-SRIR trained models had poor localization performance on certain results (Water tap/Faucet, and Knock), the AUG-SRIR model outperformed the baseline TAU-SRIR dataset. This suggests that the SIM-SRIR datasets are actually providing beneficial information for these sound classes missing from the TAU-SRIR datasets. With more thorough model tuning, it's possible that the performance for SIM-SRIR trained models it even closer to that of the baseline.

5. CONCLUSION

In this work we demonstrated the potential of using acoustic simulation to generate spatial audio data for training SELD models. We've shown that simulated SRIR data can improve the performance of SELD models as a form of data augmentation. In addition to this, we've shown that simulated SRIRs, while not as effective as those

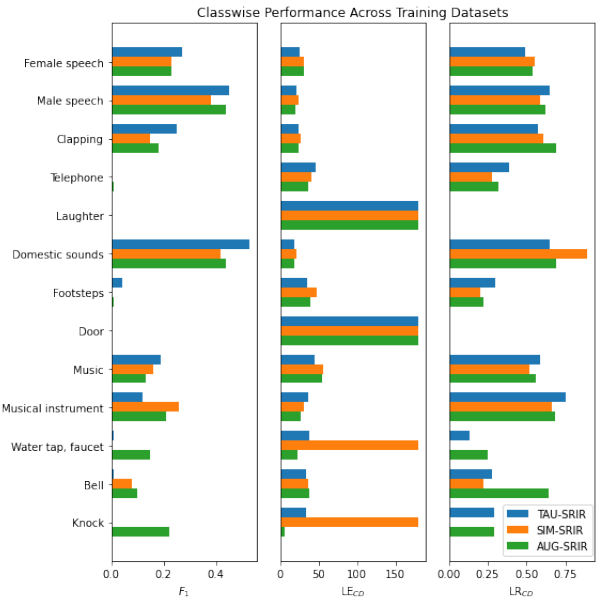

Figure 3: Per-class results of models trained on each dataset for F-measure within 20°, localization error, and localization recall.

recorded in real acoustic environments, can be used to effectively train SELD models, removing the relatively high cost of producing additional data for similarly performing results in a relatively limited setting. Generating larger volumes of SRIRs over a wider range of acoustic conditions could provide even better results than these baselines, potentially demonstrating greater robustness over varying acoustic environments. Furthermore, using a high-volume of simulated SRIRs to train a model, and using a hold-out of limited high-quality real-world data to fine the model could produce SoTA results. This result is promising for future experiments involve SRIRs for use in acoustic simulation data. Understanding the requirements for angular density in dynamic SRIR recordings can help inform future dataset collection practices, as well as the robustness of these models to noise; limited work was done exploring the effect of noise on the models trained with simulated SRIRs. Further ablation studies are necessary to understand the limitations of geometrical acoustic methods for SELD-based tasks, but these early experiments suggest that these can provide a low-resource alternative to real-world SRIR recordings.

6. REFERENCES

[1] L. Bondi, G. Chuang, C. Ick, A. Dave, C. Shelton, B. Coltin, T. Smith, and S. Das, "Acoustic imaging aboard the international space station (iss): Challenges and preliminary results," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5108–5112.

[2] F. Stevens and D. Murphy, "Spatial impulse response measurement in an urban environment," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Aug 2014. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17355

[3] J. Wang, Y. Liu, B. Wang, Y. Zhi, S. Li, S. Xia,

J. Zhang, F. Tong, L. Li, and Q. Hong, "Spatial-aware speaker diarizaiton for multi-channel multi-party meeting," in *Interspeech 2022*. ISCA, sep 2022. [Online]. Available: https://doi.org/10.21437%2Finterspeech.2022-11412

[4] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *2011 19th European Signal Processing Conference*, 2011, pp. 1317–1321.

[5] C. J. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, "Sound based localization and identification in industrial environments," in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, 2017, pp. 6119–6124.

[6] R. Chakraborty and C. Nadeu, "Sound-model-based acoustic source localization using distributed microphone arrays," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 619–623.

[7] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools Appl.*, vol. 75, no. 17, p. 10407–10439, sep 2016. [Online]. Available: https://doi.org/10.1007/s11042-015-3105-4

[8] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," *CoRR*, vol. abs/1710.10059, 2017. [Online]. Available: http://arxiv.org/abs/1710.10059

[9] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[10] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *CoRR*, vol. abs/1807.00129, 2018. [Online]. Available: http://arxiv.org/abs/1807.00129

[11] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," *CoRR*, vol. abs/1904.12769, 2019. [Online]. Available: http://arxiv.org/abs/1904.12769

[12] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 165–169. [Online]. Available: https://dcase.community/workshop2020/proceedings

[13] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129. [Online]. Available: https://dcase.community/workshop2022/proceedings

[14] A. Politis, S. Adavanne, and T. Virtanen, "TAU Spatial Room Impulse Response Database (TAU- SRIR DB)," Apr. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6408611

[15] J. B. Allen and D. 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, 04 1979. [Online]. Available: https://doi.org/10.1121/1.382599

[16] C. Ick, A. Mehrabi, and W. Jin, "Blind acoustic room parameter estimation using phase features," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] F. B. Gelderblom, Y. Liu, J. Kvam, and T. A. Myrvoll, "Synthetic data for dnn-based doa estimation of indoor speech," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4390–4394.

[18] J. B. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 04 1979. [Online]. Available: https://doi.org/10.1121/1.382599

[19] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulations and array processing algorithms," *CoRR*, vol. abs/1710.04196, 2017. [Online]. Available: http://arxiv.org/abs/1710.04196

[20] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965. [Online]. Available: https://doi.org/10.1121/1.1909343

[21] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022.

[22] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k," Oct. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4060432

[23] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," 2022.

[24] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2022. [Online]. Available: https://doi.org/10.1109%2Ficassp43922.2022.9746132