ANALYZING THE EFFECT OF EQUAL-ANGLE SPATIAL DISCRETIZATION ON SOUND EVENT LOCALIZATION AND DETECTION

Saksham Singh Kushwaha[1], *Iran R. Roman*[2*], *Juan Pablo Bello*[2,3]

[1] Courant Institute of Mathematical Sciences, New York University, NY, USA
[2] Music and Audio Research Lab, New York University, NY, USA
[3] Center for Urban Science and Progress, New York University, NY, USA

ABSTRACT

Sound event localization and detection (SELD) models detect and localize sound events in space and time. Datasets for SELD often discretize spatial sound events along the polar coordinates of azimuth (integers from -180º to 180º) and elevation (integers from -90º to 90º). This discretization, known as equal-angle, results in more dense points at the poles (±90º elevation) than at the equator (0º elevation). We first analyzed the effect of equal-angle discretization on the 2022 DCASE SELD baseline model. Since the STARSS 2022 dataset that accompanies the model shows unbalanced sampling of spatial sound events along the elevation axis, we created a synthetic dataset. Our dataset has spatial sound events uniformly distributed along the elevation axis. We created two versions: one with targets spatially discretized using equal-angle, and another one with a uniform spatial discretization (both versions had the same audio). The model trained with equal-angle showed a greater angular localization error for targets around the equator compared to the poles, while the model trained with uniform spatial discretization showed a uniform localization error along the elevation axis. To train the model with the STARSS2022 dataset and reduce the effect of its equal-angle-discretized targets, we modified the model's loss function to penalize localization errors above an angular distance threshold around each target. Using this loss we fine-tuned a model trained with the original loss, and also trained the same model from scratch. Results showed improved localization metrics in both models compared to baseline, while retaining classification metrics. Our results show that equal-angle discretization yields models with non-uniform localization errors for targets along the elevation axis. Finally, our proposed loss function penalizes the SELD model's angular localization errors, regardless of which spatial discretization was used to annotate the dataset targets.

Index Terms— sound event localization and detection, spatial sampling, activity-coupled Cartesian direction of arrival, DCASE

1. INTRODUCTION

Sound event localization and detection (SELD) consists of localizing sound events in space and time while also assigning them to a class label [1]. SELD can be applied for environmental sound classification [2], simultaneous localization and mapping for navigation without visual input or with occluded targets [3, 4], tracking of sound sources of interest [5], audio surveillance [6], and acoustic imaging [7]. As a result, there has been an increased interest in SELD modeling, and research communities have organized challenges to centralize efforts and advancements [8, 9, 10].

*corresponding author email: roman@nyu.edu

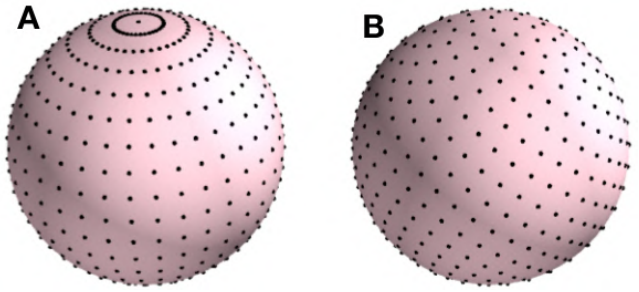

Figure 1: Two types of spatial discretization of points on a sphere. (A) equal-angle and (B) Fibonacci. (A) results in denser points at the poles than at the equator, while (B) does not.

The Detection and Classification of Acoustic Scenes and Events (DCASE) community introduced its annual SELD challenge in 2019 [10]. The DCASE SELD challenge provides participants with multichannel audio recordings of categorical sound events (i.e. speech, footsteps, running water, etc.) and their spatiotemporal trajectories on the azimuth and elevation axes. Participants must develop methods to detect, localize, and classify each event. The DCASE SELD challenge also provides a baseline model, which reflects key incremental advancements from the community. For example, a significant advancement has been in the training loss function, which went from separately measuring sound event detection (SED) and direction-of-arrival (DOA) [1] to jointly carrying out these using a mean-squared-error (MSE) regressor that accounts for overlapping sound events categories [11].

SELD datasets often have spatial targets that are discretized in a sphere along the elevation ($\theta = [-90, 90] \in \mathbb{Z}$) and azimuth ($\phi = [-180, 180) \in \mathbb{Z}$) axes in units of degrees [1, 12, 13, 14]. This sampling of points in space, known as equal-angle, is easy to interpret because it yields uniform-looking grids on a 2D projection (ϕ vs θ). However, equal-angle points on the sphere shows a larger density of points at the poles (±90° elevation) than at the equator (0° elevation) [15]. Furthermore, equal-angle sampling results in larger quantization errors around the equator.

As far as we know, current SELD research has not studied how a non-uniform density of points along the elevation axis impacts model performance. We hypothesize that training SELD models with equal-angle discretized data results in non-uniform localization performance along the elevation axis. This paper empirically analyzes the impact of equal-angle spatial discretization on SELD model performance and proposes a practical way to mitigate it.

2. EQUAL-ANGLE DISCRETIZATION IS IRREGULAR

A sphere can be discretized into N points using a sampling function $s(\theta, \phi)$, resulting in the set of vectors $\{x_1, x_2, ...x_N\} \subset \mathbb{S}^2$. This set represent a lattice of directions around the sphere [16]. For the specific case of equal-angle spatial discretization, $N = N_E \times N_A$, where N_E is the number of points uniformly sampled along the elevation axis $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and N_A is the number points along the azimuth axis $\phi \in [0, 2\pi)$ (see supplement S1[1] for a detailed mathematical description of the equal-angle sampling function and its non-uniform density along the elevation axis).

Figure 1A shows how equal-angle spatial discretization, although regular along the axes of azimuth and elevation, results in a lattice with non-uniform distances between points, particularly noticeable along the elevation axis (i.e. poles versus equator).

It is worth noting that humans listen most events close to the equator (i.e. other speakers or ecologically-relevant sound sources on the azimuth). This introduces another sampling bias in realistic SELD datasets. Thus, equal-angle discretization can yield less-than-ideal resolution where most relevant sound sources exist.

3. SELD WITH EQUAL-ANGLE SPATIAL TARGETS

We want to empirically test if equal-angle discretization affects SELD model performance. We hypothesize that the model's localization error will be a function of target elevation. More specifically, we predict that the irregularities of equal-angle discretization will result in larger errors around the equator than at the poles.

3.1. Synthetic dataset with equal-angle spatial discretization

We wanted to do this analysis with the DCASE STARSS2022 dataset [17] (real-world events belonging to thirteen categories, spatially discretized using equal-angle). However, its distribution of events on the elevation axis is not uniform (see supplement S2). Therefore, it will be hard to determine the effect of equal-angle discretization on model performance (since the data shows non-uniform distribution of targets). Moreover, our analysis is not focused on classification, so a single sound event category would be enough. We decided to create a synthetic dataset that controls for uniform target localization along the elevation axis using a single sound category. In contrast to our dataset, DCASE STARSS2022 is more complex, so we expect SELD models to easily learn our synthetic training split and generalize to the test split.

Our synthetic dataset has no moving or overlapping events, and repeats a single alarm sound (5 seconds duration) from FSD50k [18]. We used impulse responses (IRs) from two rooms (No. 3 and No. 4) in the TAU-SRIR database [19] that we convolve with the alarm sound. The rooms were selected because they contain IRs from sources localized at elevations spanning the integers

$$\theta \in [-33..32] \mid \theta \neq -25, -24, -3, -2, 3, 20, 21 \qquad (1)$$

in units of degrees (the missing integers are elevations not present in the two rooms). To generate the data we used the DCASE2022-data-generator that accompanies the TAU-SRIR database [19].

We synthesized two data folds: training and testing with 1600 and 900 tracks, respectively. Each track was a four-channel signal ($fs = 24kHz$) with a duration of 1 minute (sequences of alarm sounds and silence). While both training and test sets had elevation values spanning the range described in Eq. 1, no IRs overlapped across sets. This ensured that the absolute location of simulated sound sources was different between training and test sets. Figure 2 shows the distribution of sound event locations in our dataset.

[1]the supplement and code are available at `https://github.com/sakshamsingh1/dcase_seld_spatial_sampling_analysis`

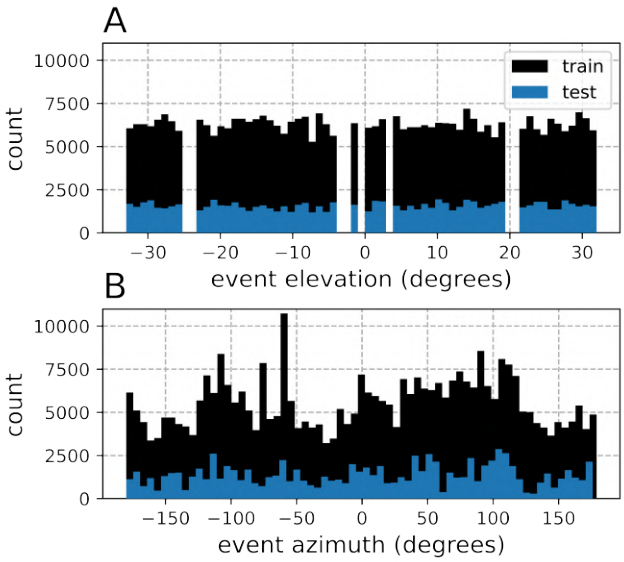

Figure 2: Density of sound events locations in our synthetic dataset along the elevation (A) and azimuth axes (B). Data synthesis controlled for uniform density of sound events along the elevation axis.

3.2. SELD model localization on equal-angle targets

We used our dataset to train the 2022 DCASE SELD baseline model [1], which is a convolutional recurrent neural network that maps multichannel audio features (generalized cross-correlation with phase transform) into sound event locations and classes (see section 6.1 for a description of the model's output format). The trained model detected all test set sound events and showed an average angular localization error of 1.81°. Figure 3A shows a scatter plot with the localization error for each test set prediction as a function of target elevation. To gain intuition about how the model's localization varies as a function of elevation, we fit a line and a second-order polynomial (i.e. parabola) to this plot. The coefficient that multiplies the polynomial's second-order term determines the curvature of the parabola. If its value is close to zero, this indicates that the parabola resembles a line. In contrast, a more negative (positive) coefficient indicates that the parabola is more curved, and we can interpret it as the model's error decreasing (increasing) as a function of elevation away from the equator.

The line was $\hat{y}_l = 0.89 + (3.2 \times 10^{-3})x$, while the polynomial was $\hat{y}_p = 1.03 + (2.9 \times 10^{-3})x - (0.4 \times 10^{-3})x^2$, also shown in Figure 3A. The polynomial's second order coefficient reveals an upside-down parabolic relationship between the model's angular localization error and target elevation. The Pearson's correlation coefficient for the linear regression was $r = 0.01$, and for the polynomial regression was $r = 0.09$. This indicates that the parabola better explains the model's error than the line. These results show that training a SELD model with equal-angle data results in larger localization errors around the equator.

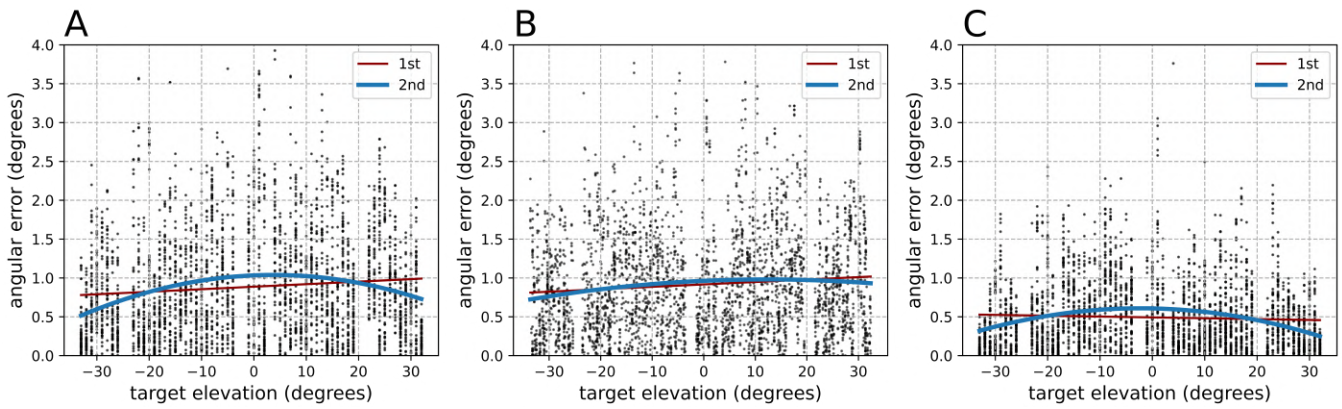

Figure 3: Scatter plots of SELD model angular localization error on our synthetic test data as a function of target elevation. Each dot is a sound event. (A) Trained with equal-angle points. (B) Trained with Fibonacci points. (C) Trained with equal-angle points and fine-tuned using our proposed loss function (see section 6). The red and blue lines are linear and second-order polynomial regressions, respectively.

4. UNIFORM DISCRETIZATION OF THE SPHERE

Alternatives to equal-angle discretization that uniformly sample points on the sphere exist [20, 21]. Perfectly-uniform discretization is limited by using the five Platonic solids, whose vertices can be used as points on the sphere [15]. The Fibonacci lattice [16] is another possible method that results in neighboring points separated by a roughly equal angular distance and is obtained by sampling points along a spiral that links the two poles (see [16] for the Fibonacci lattice formula). Figure 1B shows Fibonacci discretization.

5. SELD WITH UNIFORM SPATIAL TARGETS

We want to analyze SELD model localization when trained with uniformly discretized spatial targets. We hypothesize that this will result in uniform localization error on the elevation axis.

5.1. Synthetic dataset with uniform spatial sampling

Our synthetic dataset generated in section 3 spatially discretized sound events using equal-angle. To generate a uniformly discretized version of our dataset, we took our dataset's equal-angle annotations and transformed them into Fibonacci discretization by converting each equal-angle target into the nearest Fibonacci point. Our Fibonacci discretization had N=32768 points, which is the power of two that yields an angular distance between neighboring points around 1° [16] (similar to the distance between integers in Eq. 1). The audio tracks were exactly the same across equal-angle and Fibonacci versions of the dataset.

5.2. SELD model localization on Fibonacci targets

The 2022 DCASE SELD model trained with Fibonacci targets also detected all sound events, showing an average angular localization error of 1.86° on the test set (Figure 3B shows this model's scatter plot). We also fit a line to this plot, which was $\hat{y}_l = 0.92 + (3.1 \times 10^{-3})x$, and a second-order polynomial, which was $\hat{y}_p = 0.96 + (3.0 \times 10^{-3})x - (0.1 \times 10^{-3})x^2$. Compared to the polynomial for the model trained with equal-angle data, this polynomial's second order term reveals a less pronounced parabola, which is also visible in Figure 3B. The Pearson's correlation coefficient for the linear regression was $r = 0.09$, and for the polynomial regression was $r = 0.11$. This indicates that, compared to equal-angle, the parabola and the line more similarly explain the model's error as a function of target elevation after training with Fibonacci targets. In other words, the model trained with Fibonacci data shows localization errors that are uniform as a function of target elevation.

Our results clearly illustrate how SELD model performance is affected by equal-angle and Fibonacci discretization. However, we recognize that the Pearson and parabolic coefficients we observed show clear trends but are relatively weak indicators. Future work could support our observations using more robust statistical testing.

6. PROPOSED SOLUTION

Our empirical analysis with synthetic data revealed that equal-angle discretization can result in a SELD model with larger localization errors at the equator than at the poles. Substituting the equal-angle discretization with a uniform one (like a Fibonacci lattice) would be a simple solution. In fact, resampling the DCASE STARSS2022 dataset using a Fibonacci lattice and training the model from scratch did result in improved metrics on the test set compared to baseline (see Table 6.1). However, since equal-angle discretization is prevalent in SELD datasets, engineered SELD learning methods that reduce its impact without the need to spatially resample the data are needed. Here we propose a training loss function that, in addition to computing the mean-squared error (MSE) between targets and model predictions, penalizes the model's angular localization error uniformly for all points on the sphere.

6.1. The threshold angular error ADPIT (TAEADPIT) loss

The 2022 DCASE SELD model is trained with the auxiliary duplicating permutation invariant training (ADPIT) loss [11], which uses the multi-class activity-coupled cartesian direction of arrival (multi-ACCDOA) target format $\boldsymbol{P} \in \mathbb{R}^{3\times N\times C\times T}$, where 3 is the dimensionality of 3D cartesian coordinates, N is the maximum number of simultaneous sound events the model is trained to detect, C is the number of classes and T is the number of time frames. A vector $\boldsymbol{P}_{nct} \in \mathbb{R}^3$ has a magnitude of 1, i.e. $||\boldsymbol{P}_{nct}||_2 = 1$ and represents the location of a sound event for a specific track n, a specific

class c and a specific time-frame t. Such a vector may also represent the absence of a sound event if it has a magnitude of 0. A related term, $a \in \mathbb{R}^{N\times C\times T}$, indicates the activities over tracks, classes, and time, and $a_{nct} \in \{0, 1\}$. It is worth noting that the target contains duplicated sound events along the N dimension when the number of simultaneous sound events for each class is less than the maximum N. A complete description of the ADPIT loss can be found in its original publication [11].

The multi-ACCDOA format is permuted for each time-frame and class, and all permutation are compared against the model's output $\hat{\boldsymbol{P}} \in \mathbb{R}^{3\times N\times C\times T}$ using MSE, yielding the ADPIT loss

$$\mathcal{L}^{ADPIT} = \frac{1}{CT}\sum_{c}^{C}\sum_{t}^{T}\min_{\alpha\in\text{Perm}(ct)} l_{\alpha,ct}^{ACCDOA}, \tag{2}$$

$$l_{\alpha,ct}^{ACCDOA} = \frac{1}{N}\sum_{n}^{N}\text{MSE}(\boldsymbol{P}_{\alpha,nct}^{*}, \hat{\boldsymbol{P}}_{nct}), \tag{3}$$

where $\boldsymbol{P}_{\alpha,nct}^{*}$ indicates a permutation of the multi-ACCDOA format, and only the permutation that resulted in the minimum $l_{\alpha,ct}^{ACCDOA}$ term is used to average over classes and time-frames.

Due to the nature of the multi-ACCDOA format, the ADPIT loss function operates over Cartesian coordinates. We propose adding a term that penalizes the model's angular localization error on the sphere where the data is spatially discretized:

$$\begin{aligned} l_{\alpha,ct}^{ALE} &= \max(\boldsymbol{a}_{\alpha,nct}\text{ALE}_{\alpha,nct}, H) \\ \text{ALE}_{\alpha,nct} &= \angle(p(\boldsymbol{P}_{\alpha,nct}), p(\hat{\boldsymbol{P}})_{nct}), \end{aligned} \tag{4}$$

where $p(x)$ is a function that converts from cartesian to polar coordinates denotes, $\angle(a, b)$ is the angular distance between inputs a and b, and H is a threshold. ALE stands for angular localization error. Note that the $\boldsymbol{a}_{\alpha,nct}$ term masks ALE so that only the model's angular localization error related to active targets counts toward the loss. Adding the ADPIT loss gives:

$$\mathcal{L}^{TAEADPIT} = \frac{1}{CT}\sum_{c}^{C}\sum_{t}^{T}\min_{\alpha\in\text{Perm}(ct)} l_{\alpha,ct}^{ACCDOATAE}, \tag{5}$$

$$l_{\alpha,ct}^{ACCDOATAE} = \frac{1}{N}\sum_{n}^{N}\text{MSE}(\boldsymbol{P}_{\alpha,nct}^{*}, \hat{\boldsymbol{P}}_{nct}) + \beta(l_{\alpha,ct}^{ALE} - H), \tag{6}$$

where β is a scale factor on the new term. We call this new loss "thresholded angular error ADPIT" (TAEADPIT) loss. The new term ALE is a regularizer that uniformly penalizes angular localization errors, independent of how targets are spatially discretized.

7. EMPIRICAL EVALUATION OF THE TAEADPIT LOSS

We conducted experiments to assess the TAEADPIT loss. First, we used it to fine-tune the model trained with the equal-angle version of our synthetic dataset. Figure 3C shows the model's angular localization error as a function of elevation. We also fit a line to this plot, which was $\hat{y}_l = 0.49 - (1.1\times10^{-3})x$, and a second-order polynomial, which was $\hat{y}_q = 0.61 - (1.4\times10^{-3})x - (0.31\times10^{-3})x^2$. The parabola's second order coefficient has a value of -0.31×10^{-3}, which is closer to zero compared to the one found before fine-tuning (-0.42×10^{-3}). This indicates that fine-tuning with the TAEADPIT loss flattened the parabola. In other words, the model fine-tuned with the TAEADPIT loss shows more uniform localization errors as a function of target elevation than it did before being fine-tuned.

Loss	ER_{20^o}	F_{20^o}	LE_{CD}	LR_{CD}	SELD
ADPIT-base	0.69	0.24	30.43	0.43	0.55
TAEDPIT-tune	0.71	0.23	28.86	0.47	0.54
TAEDPIT	0.71	0.20	26.42	0.41	0.56
ADPIT-Fib	0.68	0.22	26.11	0.46	0.54

Table 1: Comparison of SELD model performance when training with ADPIT loss versus training with the proposed TAEADPIT loss. ADPIT-base: the baseline 2022 DCASE SELD model. TAEADPIT-tune: fine-tuning the baseline model with TAEADPIT. TAEADPIT: baseline model trained from scratch with TAEADPIT. ADPIT-Fib: baseline model trained with the data spatially resampled to the Fibonacci lattice and the ADPIT loss. The metrics are the DCASE SELD challenge metrics with class-depending macro-averaging are used (see [10]).

We also wanted to assess the TAEADPIT loss using real-world data. First, we ensured that we could replicate the baseline DCASE SELD 2022 model metrics using the four-channel microphone version of the DCASE STARSS2022 dataset [14] and supplemental synthetic data [22] (see Table 1). Next, we ran a couple of experiments to assess whether the TAEDPIT loss could benefit this model's performance. In all experiments, $\beta = 1 \times 10^{-3}$ in Eq. 5 (empirically-found). First, we used the TAEADPIT loss to fine-tune it. Then, we trained it from scratch using the TAEADPIT loss. Table 6.1 shows the results on the DCASE SELD metrics, indicating that using the TAEADPIT loss to fine-tune or train the SELD model from scratch can improve its localization error while retaining or only marginally impacting the classification metrics.

Our results show that the TAEADPIT loss can be used to train a SELD model using equal-angle data and improve localization metrics, and that it does so by reducing the larger localization error around the equator produced by equal-angle discretization.

8. CONCLUSION

In this paper, we studied the irregularities of equal-angle spatial discretization, which results in a larger density of points at the poles than at the equator. No previous studies have shown whether a SELD model's performance is affected by training with equal-angle discretized targets. We have empirically shown that equal-angle data affects SELD model localization on the elevation axis, causing larger localization errors around the equator than at the poles.

We also studied whether discretizing targets using a uniform Fibonacci lattice resulted in the same effect. We found that training a SELD model with Fibonacci data results in more uniform localization errors along the elevation axis compared to equal-angle. We also proposed a loss function to mitigate the effect of equal-angle by adding a thresholded angular localization error term to the ADPIT loss. Empirical results using our proposed loss when training a SELD model with equal-angle showed improved localization metrics compared to when using the ADPIT loss.

Next, we would like to assess whether a thresholded angular localization error in the training loss benefits other SELD models, and whether the benefit depends on audio format (i.e. FOA, HOA, stereo) and/or localization target format (i.e. Cartesian, 3D polar).

9. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation grant no. IIS-1955357. The authors thank the funding source and their grant collaborators. We would also like to thank Charalampos Avraam for his comments to improve Supplement Section S1.

10. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. 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.

[2] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[3] C. Evers and P. A. Naylor, "Acoustic slam," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[4] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *European Conference on Computer Vision*. Springer, 2020, pp. 17–36.

[5] W. Mack, U. Bharadwaj, S. Chakrabarty, and E. A. Habets, "Signal-aware broadband doa estimation using attention mechanisms," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4930–4934.

[6] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.

[7] M. Simeoni, S. Kashani, P. Hurley, and M. Vetterli, "Deepwave: a recurrent neural-network for real-time acoustic imaging," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," in *International evaluation workshop on classification of events, activities and relationships*. Springer, 2006, pp. 1–44.

[9] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[10] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.

[11] 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.

[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, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 125–129. [Online]. Available: https://dcase.community/workshop2021/proceedings

[14] 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," *arXiv preprint arXiv:2206.01948*, 2022.

[15] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[16] B. Keinert, M. Innmann, M. Sänger, and M. Stamminger, "Spherical fibonacci mapping," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–7, 2015.

[17] A. Politis, Y. Mitsufuji, P. Sudarsanam, K. Shimada, S. Adavanne, Y. Koyama, D. Krause, N. Takahashi, S. Takahashi, and T. Virtanen, "STARSS22: Sony-TAu Realistic Spatial Soundscapes 2022 dataset," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6600531

[18] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[19] 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

[20] 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.

[21] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "Improved feature extraction for crnn-based multiple sound source localization," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 231–235.

[22] A. Politis, "[DCASE2022 Task 3] Synthetic SELD mixtures for baseline training," Apr. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6406873