STARSS22: A DATASET OF SPATIAL RECORDINGS OF REAL SCENES WITH SPATIOTEMPORAL ANNOTATIONS OF SOUND EVENTS

Archontis Politis[1], *Kazuki Shimada*[2], *Parthasaarathy Sudarsanam*[1], *Sharath Adavanne*[1], *Daniel Krause*[1]
Yuichiro Koyama[2], *Naoya Takahashi*[2], *Shusuke Takahashi*[2], *Yuki Mitsufuji*[2] *Tuomas Virtanen*[1],

[1] Audio Research Group, Tampere University, Tampere, Finland
[2] Sony Group Corporation, Tokyo, Japan

ABSTRACT

This report presents the Sony-TAu Realistic Spatial Soundscapes 2022 (STARS22) dataset of spatial recordings of real sound scenes collected in various interiors at two different sites. The dataset is captured with a high resolution spherical microphone array and delivered in two 4-channel formats, first-order Ambisonics and tetrahedral microphone array. Sound events belonging to 13 target classes are annotated both temporally and spatially through a combination of human annotation and optical tracking. STARSS22 serves as the development and evaluation dataset for Task 3 (Sound Event Localization and Detection) of the DCASE2022 Challenge and it introduces significant new challenges with regard to the previous iterations, which were based on synthetic data. Additionally, the report introduces the baseline system that accompanies the dataset with emphasis on its differences to the baseline of the previous challenge. Baseline results indicate that with a suitable training strategy a reasonable detection and localization performance can be achieved on real sound scene recordings. The dataset is available in `https://zenodo.org/record/6600531`.

Index Terms— Sound event localization and detection, sound source localization, acoustic scene analysis, microphone arrays

1. INTRODUCTION

Sound event localization and detection (SELD) refers to the task of simultaneously detecting the presence and tracking the location of sound types of interest over time. It relates strongly to the more established tasks of sound event detection (SED) and sound source localization (SSL) but it adds spatial information to the first and semantic information to the second. The SELD task has recently seen increased research interest in part due to its introduction to the DCASE Challenge in 2019 [1]. The challenge dataset was generated with a collection of spatial room impulse responses (SRIRs) from 5 spaces and multiple source positions, convolved with dry isolated sound event recordings [2]. The next iteration of DCASE2020 increased the dataset diversity by including SRIRs of 10 additional rooms with stronger reverberation and, more importantly, by emulating dynamic scenes with both moving and static sound sources [3], while the DCASE2021 dataset introduced additionally directional interfering events out of the target classes [4].

The three SELD datasets of DCASE2019-2021 contributed to the continuous development and improvement of SELD methods by aiming to emulate accurately spatial and acoustical properties of sound scenes and to increase gradually scene complexity towards more realistic conditions. However, there are certain limitations inherent to generating synthetic mixtures. One such limitation is the random presence of target classes and the random sequencing of sound events, discarding natural temporal occurrences or co-occurrences of certain sounds in a real scene. Another limitation is the randomized spatial distribution of sound events ignoring their spatial constraints and connections in a scene. To overcome such limitations, SELD systems should transition to training and evaluation with recordings of real sound scenes. Such datasets require strong event labels provided by human annotators and simultaneous spatial annotations provided by some form of automated tracking. Due to the required annotation effort and complexity, there are no published SELD datasets we know of except for the SECL-UMons one in [5], capturing natural sound events of 11 classes occuring at pre-defined locations in two spaces. However, even though the events have a natural spatial distribution, the dataset is limited to single event recordings in isolation or to combinations of two simultaneous events, ignoring spatio-temporal information linking events in a natural scene. A few more synthetic SELD datasets exist with the same limitations as the DCASE datasets, based on captured SRIRs and targeting certain applications, such as wearable arrays [6] or positional localization in a room with distributed arrays [7].

This report presents the first SELD dataset we are aware of where realistic scenes, loosely acted by multiple actors, are captured and annotated with strong labels temporally and spatially. The challenges of such annotations are dealt with a combination of human listening and optical tracking, employing multiple sensors and modalities. Since the sound scenes are acted naturally, the dataset overcomes the limitations of synthetic datasets discussed earlier. Target sound classes are not combined randomly but are instead constrained by the environment and the participants, while the presence of each class is determined by the natural composition of each scene. Causal and sequential occurrences of sound events, as well as co-occurrences, follow the actions of the actors and their interactions with the environment. The same holds for the location of events and their trajectories in case they are moving; their spatial distributions are naturally constrained by the type of event, while event trajectories can reveal scene information on the agents and their actions. Hence, the dataset opens certain new possibilities for SELD systems apart from evaluation in realistic scenarios.

The STARSS22 dataset serves as the development and evaluation dataset of DCASE2022 Task 3, and it is followed by a suitable baseline and evaluation setup. Changes with respect to the previous DCASE challenges are elaborated. Since the duration of the dataset is limited compared to the synthetic datasets used in previous years, use of external data is allowed in this iteration to improve model training and generalization. An example strategy based on additional synthetic data is presented for the baseline. Finally, results are presented on the development and evaluation set.

2. DATASET

The **Sony-TAu Realistic Spatial Soundscapes 2022** (**STARSS22**) dataset consists of recordings of real scenes captured with a high channel-count spherical microphone array (SMA). The recordings are conducted by two different teams at two different sites, Tampere University facilities in Tampere, Finland, and Sony facilities in Tokyo, Japan. Recordings at both sites share the same capturing and annotation process, organized in sessions corresponding to distinct rooms, human participants, and sound making props with a few exceptions. In each session, various clips are recorded with combinations of that session's participants acting some simple scenes and interacting between them and with the sound props. The scenes are not strongly scripted; they are based on generic instructions on the desired sound events and are otherwise improvised by the participants. The instructions serve as a rough guide to ensure adequate event activity and occurrences of the target sound classes in a clip.

Similarly to the previous challenges, the recordings are converted to two 4-channel spatial formats: first-order Ambisonics (FOA) and tetrahedral microphone array (MIC), both derived from the original 32-channel recordings. Conversion of the Eigenmike recordings to FOA following the SN3D normalization scheme (or ambiX) was performed with measurement-based filters according to [8]. Regarding the MIC format, channels 6, 10, 26, and 22 of the Eigenmike were selected, corresponding to a nearly tetrahedral arrangement. Analytical expressions of the directional responses of each format can be found in the DCASE2020 challenge report [3]. Finally, the converted recordings were downsampled to 24kHz.

The dataset is split into a development set (*dev-set*) and evaluation set (*eval-set*). The development set totals about 4 hrs 52 mins, of which 70 recording clips amounting to about 2 hrs are recorded in 4 different rooms in Tokyo and 51 recordings amounting to about 3 hrs are recorded in 7 different rooms in Tampere. To aid the development process, the development set is further split into a training part (*dev-set-train*, 40+27 clips in 2+4 rooms in Tokyo+Tampere) and a testing part (*dev-set-test*, 30+24 clips in 2+3 rooms in Tokyo+Tampere). The evaluation set is close to 2 hrs, recorded in 2 different rooms in Tokyo (35 clips) and in 3 different rooms in Tampere (17 clips).

2.1. Recording setup and process

Each scene was captured with 4 types of sensors: a) a high resolution 32-channel SMA (Eigenmike em32[1]) recording the main multichannel audio for the challenge, b) a 360°camera (Ricoh Theta V[2]) mounted about 10 cm above the SMA, c) a motion capture (mocap) system of infrared cameras surrounding the scene, tracking reflective markers mounted on the main actors and sound sources of interest (Optitrack Flex 13[3]), and d) wireless microphones mounted on the same tracked actors and sound sources, providing close-miked recordings of the main sound events (Røde Wireless Go II[4]). For each recording session, a suitable position of the Eigenmike and Ricoh Theta V was determined in order to cover the scene from a central position, while taking into account the intended scenarios and the specific room constraints. The origin of the mocap system was then set at ground level on the same position and the height of the Eigenmike was set at 1.5 m, while the mocap cameras were positioned at the boundaries of the room. Tracking markers were mounted to independent sound sources (such as next to the water sink, on a mobile phone on a table, on a hoover, or next to a guitar's soundhole). Head markers were additionally provided to the participants before each scene recording, in the form of headbands or hats. Tracking the head served as the reference point for all human made sounds. Mouth position for *speech* and *laughter* sounds, feet stepping position for *footstep* sounds, and hand position for *clapping* sounds were each approximated with a fixed translation from the head-tracking center close to the top of the head. Regarding clapping, participants were instructed to clap about 20 cm in front of their face to improve the position approximation. Head rotations were also logged during the scene with respect to the global coordinate frame of the mocap system. Finally, the wireless microphones were mounted to the lapel of each actor and to additional independent sound sources. A clapper sound was used to initiate the acting and to serve as a reference signal for synchronization between the different types of recordings.

Target Class	Related Audioset subclasses
Telephone	*Telephone bell ringing*, *Ringtone* (no musical ringtones)
Domestic sounds	*Vacuum cleaner, Mechanical fan, Boiling* (produced by hoover, air circulator, water boiler)
Door, open or close	Combination of *Door* & *Cupboard, open or close*
Music	*Background music* & *Pop music*, (played by a loudspeaker in the room)
Musical instrument	*Acoustic guitar, Marimba, Xylophone, Cowbell, Piano, Rattle (instrument)*
Bell	Combination of sounds from hotel bell and glass bell, closer to *Bicycle bell* & single *Chime*

Table 1: Relation of target classes to specific Audioset classes. Target classes not included in the table have an one-to-one relationship with the similarly named Audioset ones.

2.2. Annotation process

Spatiotemporal annotations of the sound events were conducted manually by the authors and research assistants. Three types of information were required in order to obtain such annotations: a) the subset of the target classes that were active in each scene, b) the temporal activity of such class instances, and c) the position of each such instance when active. (a) was observed and logged during each scene recording. (b) was manually annotated by listening to the wireless microphone recordings. Since each such microphone would capture prominently sounds produced by the human actor or source it was assigned to, onset, offsets, source, and class information of each event could be conveniently extracted. In scenes or instances where associating an event to a source was ambiguous purely by listening, annotators would consult the video recordings to establish the correct association. The temporal annotation resolution was set to 100 msec.

After onset, offset, and class information of events was established for each source and actor in the scene, the positional annotations (c) were extracted for each such event by masking the tracker data with the temporal activity window of the event. Additionally,

[1] https://mhacoustics.com/products#eigenmike1

[2] https://theta360.com/en/about/theta/v.html

[3] https://optitrack.com/cameras/flex-13/

[4] https://rode.com/en/microphones/wireless/wirelessgoii

	Global	**Fem. speech**	**Male speech**	**Clap**	**Phone**	**Laugh**	**Dom. sounds**	**Footsteps**	**Door**	**Music**	**Music. instr.**	**Faucet**	**Bell**	**Knock**
Frame coverage (% total frames)	84.7	20.4	37.6	0.7	1.4	2.7	17.9	1.3	0.6	29.4	4.0	1.7	1.5	0.1
Max. polyphony	5	2	3	2	1	4	1	1	1	1	4	1	1	1
Mean polyphony	1.5	1.04	1.07	1.17	1.00	1.18	1.00	1.00	1.00	1.00	1.86	1.00	1.00	1.00
Polyphony 1 (% active frames)	61.5	96.1	93.3	83.4	100	84.0	100	100	100	100	52.2	100	100	100
Polyphony 2	29.55	3.9	6.5	16.6	0	14.5	0	0	0	0	16.6	0	0	0
Polyphony 3	7.15	0	0.2	0	0	1.1	0	0	0	0	24.2	0	0	0
Polyphony 4	1.6	0	0	0	0	0.4	0	0	0	0	7.0	0	0	0
Polyphony 5	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2: Dataset class activity information. The mean polyphony is computed over active frames only having one or more events present.

class-specific translations to the tracking data were applied if necessary, as mentioned earlier for most human made sounds. Positional information was logged in Cartesian coordinates with respect to the mocap system's origin and subsequently converted to directions-of-arrival with respect to the center of the Eigenmike. Finally, the class, temporal, and spatial annotations were combined and converted to the text format used in the previous DCASE2019-2021 challenges. Validation of the annotations was performed by observing and listening to the 360° videos, overlaid with labeled markers positioned at the DOAs of the annotated events on the 360° video plane.

2.3. Target sound classes

A set of 13 target sound classes are selected to be annotated, based on the sound events captured prominently in the recorded scenes. The class labels are selected to conform to the Audioset ontology [9] and they are: ***female speech/woman speaking, male speech/man speaking, clapping, telephone, laughter, domestic sounds, walk/footsteps, door open or close, music, musical instrument, water tap/faucet, bell, knock***. Since some of these labels correspond to superclasses with a large diversity of sounds and number of subclasses in the ontology (e.g. *domestic sounds* or *musical instrument*) we provide some additional information on the subset of sounds encountered in the recordings for some of the target classes, in the form of more specific audioset-related labels. This information is summarized in Table 1 and it can aid training and testing of methods. Certain directional sound events in the recordings are not annotated and are treated as directional interferers; examples include *computer keyboard*, *shuffling cards*, and *dishes, pots, and pans*. Additionally, there is natural background noise in all recordings, mostly HVAC-related, ranging from low to considerable levels. Based on the annotations, information on the percentage of frames that each class is active and the degree of polyphony globally and of each class separately is presented in Table 2.

3. BASELINE

3.1. Model architecture

The baseline of the DCASE2022 Task 3 challenge is similar to the one used in used in DCASE2021; a SELDnet-inspired CRNN architecture [10] improved with the ACCDOA output representation and loss [11]. However, due to the inability of the original ACCDOA representation to handle co-occuring events of the same class, the baseline adopts the recent *multi-ACCDOA* (mACCDOA) extension [12]. The mACCDOA model receives a sequence of T STFT frames of multichannel features and outputs $T/5 \times N \times C \times 3$ vector coordinates, where C is the number of target classes and N the maximum assumed number of co-occuring events in the recordings. For the current baseline N is set to 3 maximum simultaneous sources, while a value of 0.5 is used as the threshold on the length of the output vectors to indicate track and class activity. Note that a reduction of the STFT temporal resolution by a factor of 5 is performed to match the resolution of the annotations at every 100 msec.

Input features remain similar to the previous challenge [3]; namely, 4 channel 64-band log-mel spectrograms combined with acoustic intensity vectors for the FOA format or combined with generalized cross-correlation (GCC) sequences for the MIC format, following [13]. Additionally, the option of the *SALSA-lite* spatial features for the MIC format is added in the current baseline, recently shown to offer better performance than GCC in multi-source scenarios [14]. In this case, the original STFT spectrograms and the SALSA-lite features are truncated to include bins up to about 9 kHz, without mel-band aggregation, following [14]. A block diagram of the model architecture is presented in Fig. 1.

3.2. Model training

The baseline model is trained and evaluated twice: firstly only on the development set reporting baseline results for the participants to compare against during development. Secondly, it is trained on the development set and tested on the evaluation set, with results reported after the completion of the evaluation phase of the challenge. Since, the amount of training material is insufficient for the complexity of the task, additional material is synthesized for training. Those synthetic mixtures are generated with the same generation method and SRIRs as the DCASE2020-2021 datasets [3, 4]. 1200 one-minute spatial mixtures are synthesized (*synth-set*) using measured SRIRs from 9 rooms in TAU and sound event samples sourced from FSD50K [15]. The samples are chosen to match the target classes on the basis of their annotated labels which follow the Audioset ontology. The synthetic mixtures are made publicly available for reproducibility[5] along with the list of the selected FSD50K sound samples. Additionally, the SRIRs are also publicly shared[6] along with the scene generation code[7], so that participants can generate their own synthetic mixtures for training following the same process if desired. The sets and splits for training and testing of the baseline for each phase are summarized in Table 3.

[5] `https://doi.org/10.5281/zenodo.6406873`

[6] `https://doi.org/10.5281/zenodo.6408611`

[7] `https://github.com/danielkrause/DCASE2022-data-generator`

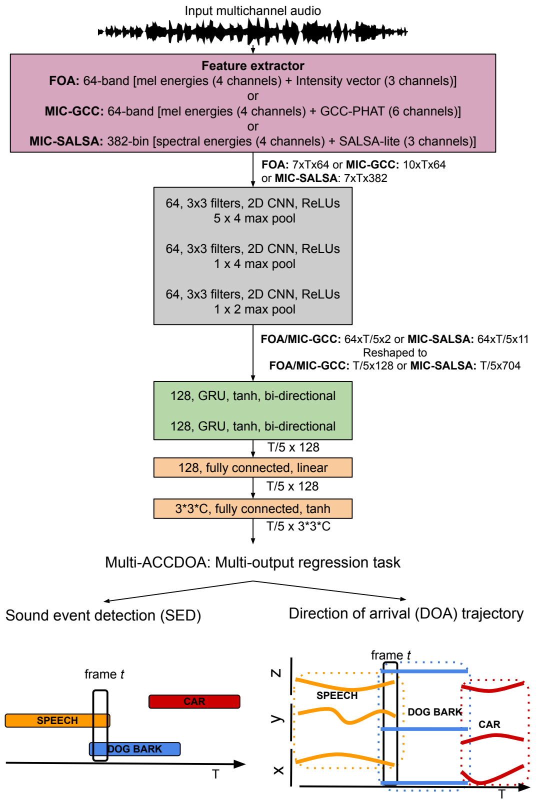
Input multichannel audio
Feature extractor
FOA: 64-band [mel energies (4 channels) + Intensity vector (3 channels)]
or
MIC-GCC: 64-band [mel energies (4 channels) + GCC-PHAT (6 channels)]
or
MIC-SALSA: 382-bin [spectral energies (4 channels) + SALSA-lite (3 channels)]
FOA: 7xTx64 or MIC-GCC: 10xTx64
or MIC-SALSA: 7xTx382
64, 3x3 filters, 2D CNN, ReLUs
5 x 4 max pool
64, 3x3 filters, 2D CNN, ReLUs
1 x 4 max pool
64, 3x3 filters, 2D CNN, ReLUs
1 x 2 max pool
FOA/MIC-GCC: 64xT/5x2 or MIC-SALSA: 64xT/5x11
Reshaped to
FOA/MIC-GCC: T/5x128 or MIC-SALSA: T/5x704
128, GRU, tanh, bi-directional
128, GRU, tanh, bi-directional
T/5 x 128
128, fully connected, linear
T/5 x 128
3*3*C, fully connected, tanh
T/5 x 3*3*C
Multi-ACCDOA: Multi-output regression task
Sound event detection (SED)
Direction of arrival (DOA) trajectory
frame t
CAR
SPEECH
DOG BARK
T
frame t
Z
Y
X
SPEECH
DOG BARK
CAR
T

Figure 1: Baseline CRNN model with mACCDOA output.

4. EVALUATION

Intermediate development set results are reported by the participants in the development set, while evaluation set results are computed by the organizers based on the submitted system outputs on the unseen evaluation set. Contrary to the previous challenges, participants are allowed to use external data during training, such as sample banks of sound events, room simulators, SRIR databases, spatial background noise recordings, pre-trained networks, and others. Generating the *synth-set* dataset to improve the baseline performance constitutes just one such example of external data usage.

The submissions are evaluated with the joint localization-detection metrics studied in [16, 1] and introduced first-time in DCASE2020. These are the location-dependent error rate (ER_{20°) and F1-score (F_{20°) for a spatial threshold 20° and the class-dependent localizastion error (LE_{CD}) and localization recall (LR_{CD}). Contrary to the previous challenges, in which F_{20° was micro-averaged, in this challenge evaluation is based on macro-averaging of F1-score to account better for the imbalanced presence of the target classes in the dataset.

Phase	**Training**	**Testing**
Development	synth-set + dev-set-train	dev-set-test
Evaluation	synth-set + dev-set-train + dev-set-test	eval-set

Table 3: Datasets & splits used for baseline training and evaluation.

	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$ (macro)	$F_{20^\circ} \uparrow$ (micro)	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
Development set					
FOA-real	0.78	0.11	-	64.1°	0.24
FOA-mixed	0.71	0.21	0.36	29.3°	0.46
MIC-mixed	0.71	0.18	0.36	32.2°	0.47
Evaluation set					
FOA-mixed	0.61	0.24	0.39	22.9°	0.51
MIC-mixed	0.61	0.22	0.41	25.9°	0.48

Table 4: Baseline results on development and evaluation set. *FOA-real* refers to training only on the development set of STARSS22, *FOA/MIC-mixed* refers to training using additionally synthetic data.

4.1. Results

Results of the baseline on the development and evaluation set are presented on Table 4, for both FOA and MIC formats. The baseline was trained as indicated in Sec. 3.2 using the additional synthetic spatial mixtures of *synth-set*. For comparison purposes, an example of the model with FOA input trained only with real recordings is also reported (FOA-real), with the training and testing splits of Table 3 excluding the synthetic data (*synth-set*). It can be seen that the performance is very low in this case, at least without using data augmentation strategies. Two training strategies were tested with regards to incorporating the synthetic data. The first was based on initial training of the model on the synthetic data, followed by fine-tuning with the development dataset. The second simply mixed both the synthetic and the development recordings and trained with the combined dataset. Better results were obtained with the mixed strategy and these are the ones presented here (FOA/MIC-mixed). It is noted that the SRIRs used for the generation of *synth-set* were captured in TAU spaces that were different than the ones were the scene recordings of the STARS22 dataset occurred. Regarding the MIC format, both the GCC features and the SALSA-lite features were tested. Slightly better results were obtained with the GCC features and reported here. That may be attributed to the fact that even though the SALSA-lite features show a clear advantage for densely populated multi-source scenes such as the ones in DCASE2021 dataset [14], for the more sparse scenes of STARSS22 that advantage may be diminished. Finally, both the micro and macro versions of the F1-score are presented here, with a clear drop in performance in the macro version, as expected with a dataset of such unbalanced presence of target classes (evident in Table 2).

5. CONCLUSIONS

This report presents the specifications of the STARS22 dataset, intended for evaluation of SELD systems in challenging real conditions with a natural composition of sound events. The dataset serves as the development and evaluation dataset of the SELD challenge of DCASE2022 and it is accompanied by a baseline model which, with use of external data and a suitable training strategy, can achieve a reasonable performance on the evaluation dataset.

6. ACKNOWLEDGMENT

The dataset collection and annotation at Tampere University has been funded by Google. This work was carried out with the support of Centre for Immersive Visual Technologies (CIVIT) research infrastructure, Tampere University, Finland.

7. REFERENCES

[1] 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 Trans. Audio, Speech, and Language Proc.*, 2020.

[2] S. Adavanne, A. Politis, and T. Virtanen, “A multi-room reverberant dataset for sound event localization and detection,” in *Work. on Detection and Classification of Acoustic Scenes and Events (DCASE)*, October 2019, pp. 10–14.

[3] A. Politis, S. Adavanne, and T. Virtanen, “A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection,” in *Work. on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 165–169.

[4] 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 *Work. on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021, pp. 125–129.

[5] M. Brousmiche, J. Rouat, and S. Dupont, “Secl-umons database for sound event classification and localization,” in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2020, pp. 756–760.

[6] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, “Wearable seld dataset: Dataset for sound event localization and detection using wearable devices around head,” in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2022, pp. 156–160.

[7] E. Guizzo, R. F. Gramaccioni, S. Jamili, C. Marinoni, E. Massaro, C. Medaglia, G. Nachira, L. Nucciarelli, L. Paglialunga, M. Pennese, *et al.*, “L3DAS21 challenge: Machine learning for 3D audio signal processing,” in *IEEE Int. Work. on Machine Learning for Sig. Proc. (MLSP)*, 2021, pp. 1–6.

[8] A. Politis and H. Gamper, “Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays,” in *IEEE Work. on Applications of Sig. Proc. to Audio and Acoustics (WASPAA)*, 2017, pp. 224–228.

[9] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2017, pp. 776–780.

[10] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE J. Selected Topics in Sig. Proc.*, vol. 13, no. 1, pp. 34–48, 2018.

[11] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, “ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection,” in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2021, pp. 915–919.

[12] 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 Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2022, pp. 316–320.

[13] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, “Polyphonic sound event detection and localization using a two-stage strategy,” in *Work. on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.

[14] 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 *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2022, pp. 716–720.

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

[16] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, “Joint measurement of localization and detection of sound events,” in *IEEE Work. on Applications of Sig. Proc. to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337.