A DATASET OF REVERBERANT SPATIAL SOUND SCENES WITH MOVING SOURCES FOR SOUND EVENT LOCALIZATION AND DETECTION

Archontis Politis, Sharath Adavanne, and Tuomas Virtanen

Audio and Speech Processing Research Group, Tampere University, Finland

ABSTRACT

This report details the dataset and the evaluation setup of the Sound Event Localization & Detection (SELD) task for the DCASE 2020 Challenge. Training and testing SELD systems requires datasets of diverse sound events occurring under realistic acoustic conditions. A significantly more complex dataset is created for DCASE 2020 compared to the previous challenge. The two key differences are a more diverse range of acoustical conditions, and dynamic conditions, i.e. moving sources. The spatial sound scene recordings for all conditions are generated using real room impulse responses, while ambient noise recorded on location is added to the spatialized sound events. Additionally, an improved version of the SELD baseline used in the previous challenge is included, providing benchmark scores for the task.

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

1. INTRODUCTION

Sound event localization and detection (SELD) takes the currently active research topic of temporal sound event detection (SED) [1] and connects it with the spatial dimension of event location or direction-of-arrival (DoA). Hence SELD aims to a more complete spatiotemporal characterization of the acoustic sound scene, with predictions on the type of sounds of interest in the scene, their temporal activations, and their spatial trajectories when they are active. This spatiotemporal scene description has a wide range of applications in machine listening, ranging from acoustic monitoring and robot navigation to intelligent human-machine interaction and deployment of immersive services.

Until the DCASE2019 Challenge[1], only a handful of approaches in literature were aiming some form of SELD [2–8]. Apart from [7, 8] which are fully deep-neural network (DNN) based approaches, these earlier works employed more traditional source localization methods such as time-difference-of-arrival (TDoA) [2,6], steered-response power [3], or acoustic intensity vector analysis [5], and Gaussian mixture models [2], hidden Markov models [3], support vector machines [5], or a simple artificial neural network [6] for classification. Additionally, most of them treated detection and localization independently, with only [4, 6] joining beamforming outputs after localization with the event classifiers.

Recently, DNNs have dominated SED approaches [1], and they have been applied successfully to pure source localization [9, 10], showing potential for joint modeling of the SELD task. The first works we are aware of this approach are [7, 8]. Hirvonen in [7] used a convolutional neural network (CNN) with localization targets at discrete DoAs, setting the SELD task as a multilabel-multiclass classification problem. In [8] we proposed the SELDnet, a convolutional recurrent neural network (CRNN) with two output branches, a classification one for SED and a regression one for DoA estimation. Both proposals were using simple generic features, such as multichannel power [7], or phase and magnitude [8], spectrograms.

Due to its relevance in all the aforementioned applications, SELD was introduced as a new task in DCASE2019 Challenge, and as such, it required a new dataset for training and evaluation of the submitted methods. This dataset, the **TAU Spatial Sound Events 2019**[2], comprised scenes with events from 11 classes, spatialized through captured room impulse responses (RIRs) as static sources at 504 possible locations for each of 5 different spaces [11]. Along with the dataset, a SELDnet implementation was provided by the authors as a baseline for the challenge participants[3]. The challenge attracted more than 20 original methods, with most methods surpassing significantly the baseline[4]. Many innovative solutions were presented for the task, such as more refined SED and localization features [12,13], a multi-stage modeling and training approach [12], data augmentation [14, 15], exploitation of domain-specific knowledge [13, 16], state-of-the-art network architectures [17], ensembles, or combinations of model-based localization and DNN-based event classification [18].

In this work we present the new dataset **TAU-NIGENS Spatial Sound Events 2020**[5] aimed for the next iteration of the SELD task in DCASE2020 challenge[6]. The dataset preserves all the realistic properties of the previous one while surpassing its major limitations: more sound examples per class, a greater number of rooms, much more diverse acoustic conditions, and non-quantized source positions in a predefined grid of directions. More importantly, the dataset introduces moving sources, which makes it significantly more challenging and closer to real-life conditions.

Along with the dataset, we introduce improvements to the SELDnet baseline of [8] that reflect some effective proposals by DCASE2019 participants, to make it more competitive. Moreover, instead of measuring performance independently for SED and localization, as in DCASE2019, we adopt recently proposed metrics for joint SELD measurement [19] which can distinguish between systems that localize the correct events at their correct position, and systems that detect and/or localize well independently.

This work has received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND.

[1] http://dcase.community/challenge2019/task-sound-event-localization-and-detection

[2] https://zenodo.org/record/2580091

[3] https://github.com/sharathadavanne/seld-dcase2019

[4] http://dcase.community/challenge2019/task-sound-event-localization-and-detection-results

[5] https://zenodo.org/record/3740236

[6] http://dcase.community/challenge2020/task-sound-event-localization-and-detection

2. REVERBERANT DYNAMIC DATASET

2.1. Sound events

Sound event samples were sourced from the recently published NIGENS General Sound Events Database[7]. This database provides a higher number of samples and classes than the one used in the previous challenge. 714 monophonic sound examples, recorded at 44.1 kHz, are distributed between 14 classes of *alarm, crying baby, crash, barking dog, running engine, burning fire, footsteps, knocking on door, female & male speech, female & male scream, ringing phone, piano*. For more details on the recordings and the database in general, the reader is referred to [20].

2.2. Recording of multichannel RIRs

The overall recording procedure was similar to the one used in the previous dataset [11], with differences highlighted below. In both datasets, instead of multiple RIR measurements at discrete source-receiver points, a very large range of source positions is covered by recording pseudo-random maximum-length sequences (MLS) emitted by a slowly moving source along predefined tracks [21]. The source is a Genelec G Three[8] loudspeaker mounted on a wheeled platform. The recording is done with a 32-channel compact spherical microphone array (SMA), the em32 Eigenmike[9]. An SMA with high channel count is chosen due to its uniform spatial resolution up to high frequencies, and to its flexibility in allowing us to extract a variety of smaller spatial formats from the same recording.

For DCASE2019, real recorded RIRs were captured from 5 rooms of similar type, with high direct-to-reverberant ratios. For DCASE2020, to add more variability in acoustical conditions and more challenging reverberation, we recorded 10 additional rooms of diverse shapes and types, such as lecture halls, large classrooms, small classrooms and meeting rooms, a modern sports hall, and a sports hall in an underground nuclear shelter with rock walls. Moreover, the recording trajectories in the new rooms were different for each one of them. In half of the rooms the trajectories were circular, but at differing distances and elevations, while in the rest linear trajectories at various heights were used. The RIRs were extracted from the moving source recordings through a simple linear regression on the filter coefficients between the clean MLS sequence and the recorded output, similar to [22]. RIRs extracted along circular trajectories have a more or less constant elevation, distance, and DRR, while ones extracted along linear trajectories have varying elevation, distance, and DRR, with respect to the recording position.

Finally, similarly to DCASE2019, apart from the MLS noise sequences an additional 30 mins of spatial ambient noise were captured in each room with the recording setup unchanged. Contrary to the 5 earlier rooms which were accessible by passing crowds at any time, the new room recordings contained mostly ventilation noise.

2.3. Reference RIRs and positional labels

During the synthesis of the spatial mixtures, sound events are intended to be spatialized at consistent DoAs across different environments, meaning that the direct path for the same DoA as encoded in the array channels should be similar between rooms, and the methods can rely on it for localization while being robust to the dissimilar reverberation patterns that follow. In the DCASE2019 dataset, the recorded circular trajectories were assumed to have the exact same geometry with respect to the microphones, and the final grid of reference positions was intended to be the same for all rooms. The temporal locations at each trajectory for those predetermined reference DoAs were located through a continuous DoA analysis of the recorded MLS, followed by the reference RIR extraction.

For the new more challenging dataset, we decided to estimate the reference DoAs acoustically, directly from the extracted RIRs, as these would reflect consistently the ones encoded into the multichannel mixture during the synthesis stage using the same RIRs. To that purpose, for each source trajectory we: a) extracted the multichannel RIRs at 200-millisecond intervals, b) estimated direct path delays from geometry and measurements, c) windowed the RIRs around their direct path, and d) applied a broadband version of the MUSIC algorithm [23] for estimation of the DoA corresponding to that early part of the RIR. From that list of RIR-DoA pairs, the final reference ones were determined by selecting the ones closest to the geometric reference trajectory, at approximately 1° intervals. Note that the same process was applied also to the 5 earlier rooms recorded for the DCASE2019 dataset.

2.4. Dataset Synthesis

All extracted multichannel RIRs and sound event samples were resampled to 24 kHz. From the 8 provided splits of the NIGENS dataset, 6 were used for the creation of the development, and the remaining 2 for the evaluation datasets. One or two rooms were assigned to each split, and 100 one-minute-long recordings of spatialized sound events were generated for each such combination of event samples and rooms. The onsets of sound events in each recording were randomly distributed but constrained by the allowed level of polyphony (number of simultaneous events, either 1 or 2).

An event was randomly chosen to be either *static* or *moving*. Static events were assigned randomly a DoA from the list of reference ones available for the specific room. Moving sound events were assigned randomly one of the RIR recording trajectories, limiting their motion along that path. The direction of movement was randomized (*forward* or *backward*), while the speed of motion was randomly chosen from three levels of *slow* ($\sim 10^\circ$/sec), *medium* ($\sim 20^\circ$/sec), and *fast* ($\sim 40^\circ$/sec). Additionally, since each trajectory was recorded at different heights, moving events reaching the end of a path had the possibility to jump to a higher or lower elevation and continue their motion on the respective path of that height.

Static events were spatialized by convolution with the respective RIRs for their intended DoA, and added to the mixture. Moving events were spatialized by a time-variant convolution scheme, performed between the STFT of the event sample and the STFTs of all the RIRs encountered along the path of motion [24]. Since the reference DoAs were extracted at about 1° intervals along a trajectory, the speed of motion was controlled by using 10 (slow), 20 (medium), or 40 (fast) consecutive RIRs per 1 second of output. Very short events of up to 2 seconds were excluded from being dynamic, and were assigned static DoAs instead.

After the reverberant spatialized events were layered with the intended polyphony, ambient noise from the same room was additionally mixed. The original ambient noise recordings were split into 1-minute segments and added to the mixtures at varying signal-to-noise ratios (SNR) from 30 dB to 6 dB. Since the duration of the recorded ambience was less than the total duration of the generated mixtures, additional noise segments were artificially generated by mixing two randomly chosen segments of the recording.

[7] https://zenodo.org/record/2535878
[8] https://www.genelec.com/g-three
[9] https://mhacoustics.com/products#eigenmike1

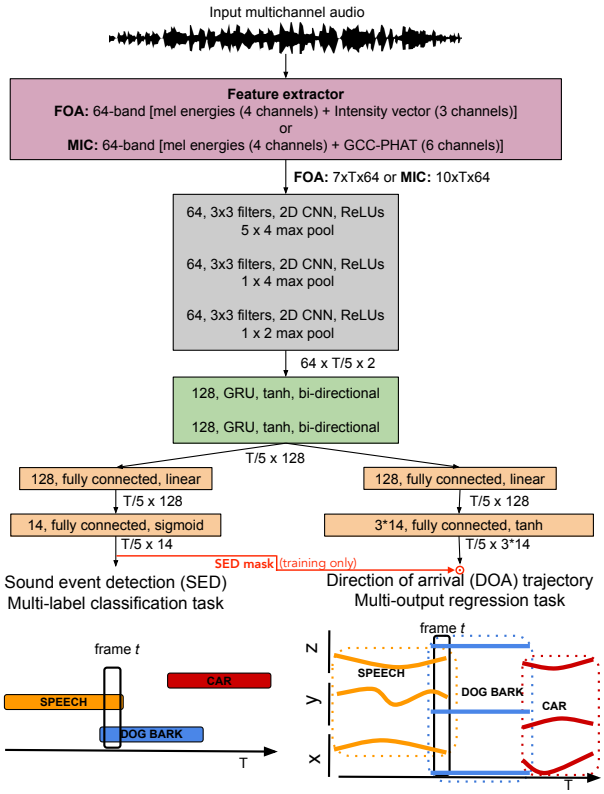

Figure 1: Baseline CRNN architecture for SELD.

2.5. Dataset Formats

As in the previous dataset for DCASE2019, we opted for delivering the synthesized sound recordings in two different 4-channel spatial sound formats, extracted from the 32-channel Eigenmike format. The first one is a 4-channel microphone array corresponding to a tetrahedral capsule arrangement (MIC), obtained directly by selecting channels 6,10,26, and 22 [11]. The second one is the widespread first-order Ambisonics (FOA), extracted through a matrix of 4×32 conversion filters, as detailed in [25]. The rationale behind offering the dataset in both the MIC and FOA formats is that each one encodes spatial information differently. Hence, different spatial features are better suited to each format, and participants could exploit one of the two or both. Analytical expressions of the spatial encoding of each formats can be found on the previous dataset report [11]. Note that contrary to DCASE2019, the FOA format omits the $\sqrt{3}$ factors on channels 2–4, to be compliant with the *SN3D* normalization scheme of Ambisonics.

3. BASELINE METHOD

As the benchmark method, we employ an updated version of SELDnet [8][10]. Specifically, we adopt changes to SELDnet that helped to improve its performance consistently across different submissions of the DCASE 2019 SELD task. Among those improvements, we include array-dependent acoustic feature extraction, and train a single model to jointly estimate SED and DOA as shown in Figure 1. Additionally, during the training, the DOA estimation branch uses the SED output as the mask, and the mean squared error loss is only computed for the sound events that are active. This strategy was first published by [12], with significant improvements on the results, and adopted by other participants in the challenge. Cross-entropy loss is used for the SED branch. Similar to the original SELDnet, we do not perform any post-processing on its output.

The updated SELDnet receives multichannel audio at 24 kHz sampling rate. For each of the MIC and FOA datasets two input features are extracted. The first feature, the multichannel mel-band power spectrogram, is common to both datasets, and, apart from being a widespread SED feature, it additionally captures inter-channel level differences (ILDs). For each channel 64 log mel-band energies are computed, with a 40 ms window and 20 ms hop length using a 1024-point FFT. The second, format-specific, spatial feature for the FOA dataset is the acoustic intensity vector, which expresses net acoustic energy flux, and is computed at each of the 64 mel-bands similar to [12, 17]. For the MIC dataset, we employ the generalized-cross-correlation with phase-transform (GCC-PHAT) feature computed in each of the 64 mel-bands similar to [12, 14, 15].

Based on the chosen dataset, the SELDnet is trained using the corresponding features. For the FOA dataset, the input is of $7 \times T \times 64$ dimension, where T is the number of time frames in the input sequence, and the number 7 arises from 4 channels of 64 dimension log mel-band energies computed for each of the 4 audio-channels, and 3 channels of FOA intensity vectors. Similarly for the MIC dataset, the input is of $10 \times T \times 64$, where 10 arises from 6 channels of GCC-PHAT computed between all pairs of audio-channels of the MIC dataset and 4 channels of log mel-band energies.

Details on the SELDnet architecture and training can be found on its original publication [8] and on DCASE2019 task report [11]. Some further differences with the DCASE2019 SELDnet baseline is temporal max-pooling at $T/5$, reducing the output temporal resolution to 100 ms, as specified by the challenge submission format. The SED branch outputs $T/5 \times C$ classification probabilities, and the DOA branch outputs $T/5 \times 3C$ Cartesian vector components, which are converted to azimuth & elevation during inference.

Table 1: Challenge setup

	Splits		
Dataset	**Training**	**Validation**	**Testing**
Development	3, 4, 5, 6	2	1
Evaluation	2, 3, 4, 5, 6	1	7, 8 (unlabeled)

4. EVALUATION

4.1. Evaluation Setup

The evaluation setup for the development dataset is shown in Table 1. Among the six splits in the dataset, the first is used as the unseen test split, the second as the validation split during training, and the remaining ones are used for training. The best parameters for a SELD method are chosen based on the validation split, without using the testing split. The performance of the best validation model on the unseen testing split is then reported as the development dataset score for the SELD method.

[10] https://github.com/sharathadavanne/seld-dcase2020

In order to have a fair comparison of the SELD performance across different submitted systems during development, participants are required to employ the proposed development dataset setup and report the performance of their method on the unseen test split 1. However, for the evaluation dataset, the participants are required to produce outputs on the unlabeled testing splits 7,8, with no restrictions on how they use the 1–6 development splits. The evaluation results presented here for the baseline are based on the evaluation dataset setup shown in Table 1.

4.2. Metrics

The 2019 version of the SELD task employed individual metrics for SED and DOA estimation. The SED performance was evaluated using the F-score (F) and error rate (ER) calculated in non-overlapping one-second segments [26]. The DOA estimation was evaluated using frame-wise metrics [9] of DOA error (DE) and frame recall (FR). The DOA error represents the average angular error in degrees between the predicted and reference DOAs. The frame recall represents the percentage of frames in which the estimated number of DOAs were identical to the reference.

Recently, in [19] we discussed the drawbacks of the above metrics for the SELD task and proposed metrics to evaluate the performance of joint detection and localization. The first two metrics, on location-aware detection, consider a prediction to be correct if the sound class of the prediction and reference are the same, and the distance between them is less than an application-specific threshold. We propose a threshold of 20° for the challenge, as an acceptable localization tolerance for a practical SELD system, and compute the corresponding metrics, error rate (ER_{20°) and F-score (F_{20°), in one-second non-overlapping segments. An ideal SELD method will have $ER_{20^\circ} = 0$ and $F_{20^\circ} = 100\%$.

The next two metrics, on class-aware localization, do not use any distance threshold, like above, but consider the error only between same-class predictions and references. The respective localization error (LE_{CD}) and its corresponding localization recall (LR_{CD}) are computed in one-second non-overlapping segments, where the subscript refers to classification-dependent. An ideal SELD method will have $LE_{CD} = 0^\circ$ and LR_{CD} of 100%.

Although information on joint localization/detection performance is gained by either location-aware detection, or class-aware localization, a thorough picture is given by all four. Hence, we evaluate the submissions in the DCASE2020 task using all four metrics. The submitted methods are ranked individually for each one of them, and the final positions are obtained using the cumulative minimum of the ranks.

5. RESULTS

The performance of the SELDnet method for the proposed evaluation setup of the DCASE 2020 SELD task is tabulated in Table 2. The results for both the DCASE2019 metrics and the official DCASE2020 metrics are reported. The 2020 metrics evaluate jointly detection and localization performance and hence provide deeper insights on the SELD performance. For instance, the 2019 detection metrics of DE and FR suggest that the SELDnet estimated the correct number of DOAs in 66.6% of the frames for the FOA test data with an average DOA error of 20.4°. But, this localization metric does not use the knowledge of detection and computes DOA error for all the detected sound classes, irrespective of them being correct or wrong. Although we have the corresponding detection scores of ER and F scores, there is no straightforward approach to assess a joint detection and localization performance. In contrast, the 2020 metrics of class-aware localization (LE_{CD} and LR_{CD}) and location-aware detection (ER_{20°) and F-score (F_{20°) can both independently provide insights on the joint performance. For instance, on the FOA test data, 60.7% (LR_{CD}) of the sound class instances were recalled successfully by the SELDnet with an average location error (LE_{CD}) of 22.8°. Similarly, if we consider that the predicted sound class is correct if it is within a margin of 20° from the reference sound class location, then we obtain an F-score (F_{20°) of 37.4% and error rate (ER_{20°) of 0.72.

Table 2: SELD performance of the baseline method evaluated using independent (2019) and joint (2020) localization/detection metrics.

	FOA				MIC			
2019	DE	FR	ER	F	DE	FR	ER	F
Development results								
Val	20.2°	62.9	0.54	62	21.9°	63.8	0.53	62.8
Test	20.4°	66.6	0.54	60.9	22.6°	66.8	0.56	59.2
2020	LE_{CD}	LR_{CD}	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	ER_{20°	F_{20°
Development results								
Val	23.5°	62	0.72	37.7	27°	62.6	0.74	34.2
Test	22.8°	60.7	0.72	37.4	27.3°	59	0.78	31.4
Detailed development dataset test-split results								
Overlap 1	18.1°	69.7	0.63	49.2	20.8°	66.6	0.70	40.8
Overlap 2	26.3°	55.4	0.77	30.4	32.0°	54.6	0.82	25.8
Evaluation results								
Val	22.8°	60.7	0.7	39.6	24.5°	58.7	0.72	36,9
Test	23.2°	62.1	0.7	39.5	23.1°	62.4	0.69	41.3
Detailed evaluation dataset test-split results								
Overlap 1	18.3°	69.9	0.58	51.3	16.0°	69.4	0.75	33.7
Overlap 2	26.7°	57.4	0.75	32.5	28.1°	58.3	0.75	33.7
Split 7	20.5°	65.0	0.66	43.3	21.8°	65.9	0.66	44.0
Split 8	26.2°	59.1	0.74	35.5	24.7°	58.9	0.72	38.6

In Table 2, although both the FOA and MIC datasets are synthesized from the same microphone array, the SELDnet is observed to perform better for FOA than the MIC dataset. This suggests that the spectral and spatial information in the two formats are not identical and methods trained with both the datasets can potentially benefit from mutual information. Finally, we observe that the performance of SELDnet on recordings without polyphony (overlap 1) is significantly better than with polyphony (overlap2). Additionally we can see, at the evaluation set results, that the model does not generalize equally well for different unseen spaces, as it performs better for one of the two rooms (split 7).

6. CONCLUSION

Herein, we outlined the sound event localization and detection task for the DCASE 2020 challenge, and its new complex dataset with dynamic sound scenes in reverberant rooms. This dataset is synthesized using impulse response trajectories measured at 15 indoor environments, with an elaborate measurement and RIR extraction procedure able to cover a large range of source positions for both static and moving events. Due to its larger scale and the dynamic conditions, it is a definite step towards realistic testing and training of SELD systems, compared to the previous dataset. Based on the same processing and synthesis framework, further realistic conditions can be integrated effectively on future datasets, such as moving receivers or out-of-class directional interferes. Absolute positional localization becomes also possible if recordings for more than one array position in a room are synthesized and mixed. Such advances are expected to be addressed in future work.

7. REFERENCES

[1] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

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

[3] 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 *19th European Signal Processing Conference (EUSIPCO)*, 2011.

[4] 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.

[5] K. Lopatka, J. Kotus, and A. Czyzewsk, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications Journal*, vol. 75, no. 17, 2016.

[6] C. Grobler, C. P. Kruger, B. J. Silva, and G. P. Hancke, "Sound based localization and identification in industrial environments," in *43rd Conference of the IEEE Industrial Electronics Society (IECON)*, 2017.

[7] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*, 2015.

[8] 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, March 2019.

[9] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *26th European Signal Processing Conference (EUSIPCO)*, 2018.

[10] 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.

[11] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[12] 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 *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[13] F. Grondin, I. Sobieraj, M. Plumbley, and J. Glass, "Sound event localization and detection using CRNN on pairs of microphones," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[14] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order Ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[15] P. Pratik, W. J. Jee, S. Nagisetty, R. Mars, and C. Lim, "Sound event localization and detection using CRNN architecture with Mixup for model generalization," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[16] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of CRNN models," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[17] S. Park, "Trellisnet-based architecture for sound event localization and detection with reassembly learning," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[18] A. Perez-Lopez, E. Fonseca, and X. Serra, "A hybrid parametric-deep learning approach for sound event localization and detection," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[19] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[20] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS general sound events database," *arXiv preprint arXiv:1902.08314*, 2019.

[21] N. Hahn and S. Spors, "Comparison of continuous measurement techniques for spatial room impulse responses," in *24th European Signal Processing Conference (EUSIPCO)*, 2016.

[22] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[23] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, 1986.

[24] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019.

[25] A. Politis and H. Gamper, "Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," in *Applied Sciences*, vol. 6, no. 6, 2016.