Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019

Archontis Politis, *Member, IEEE,* Annamaria Mesaros, *Member, IEEE,* Sharath Adavanne, Toni Heittola, Tuomas Virtanen, *Fellow, IEEE*

***Abstract*—Sound event localization and detection is a novel area of research that emerged from the combined interest of analyzing the acoustic scene in terms of the spatial and temporal activity of sounds of interest. This paper presents an overview of the first international evaluation on sound event localization and detection, organized as a task of the DCASE 2019 Challenge. A large-scale realistic dataset of spatialized sound events was generated for the challenge, to be used for training of learning-based approaches, and for evaluation of the submissions in an unlabeled subset. The overview presents in detail how the systems were evaluated and ranked and the characteristics of the best-performing systems. Common strategies in terms of input features, model architectures, training approaches, exploitation of prior knowledge, and data augmentation are discussed. Since ranking in the challenge was based on individually evaluating localization and event classification performance, part of the overview focuses on presenting metrics for the joint measurement of the two, together with a reevaluation of submissions using these new metrics. The new analysis reveals submissions that performed better on the joint task of detecting the correct type of event close to its original location than some of the submissions that were ranked higher in the challenge. Consequently, ranking of submissions which performed strongly when evaluated separately on detection or localization, but not jointly on both, was affected negatively.**

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

I. Introduction

Recognition of classes of sound events in an audio recording and identification of their occurrences in time is a currently active topic of research, popularized as sound event detection (SED), with a wide range of applications [1]. While SED can reveal a lot about the recording environment, the spatial locations of events can bring valuable information for many applications. On the other hand, sound source localization is a classic multichannel signal processing task, based on sound propagation properties and signal relationships between channels, without considering the type of sound characterizing the sound source. A sound event localization and detection (SELD) system aims to a more complete spatiotemporal characterization of the acoustic scene by bringing SED and source localization together. The spatial dimension makes SELD suitable for a wide range of machine listening tasks, such as inference on the type of environment [2], robotic simultaneous localization and mapping [3], navigation without visual input or with occluded targets, tracking of sound sources of interest [4], and audio surveillance [5]. Additionally, it can aid human-machine interaction, scene-information visualization systems, scene-based deployment of services, and assisted-hearing devices, among others.

The SELD task was included for the first time in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge of 2019[1]. In addition to the related studies that aim at detecting and localizing multiple speakers (see e.g. [6]), only a handful of approaches could be found in the literature up to that point [5], [7]–[12]. Earlier studies were treating the two problems of detection and localization separately, without trying to associate source positions and events. In those works, Gaussian mixture models (GMMs) [5], hidden Markov models (HMMs) [7], or support vector machines [9] were used for detection, while localization relied on classic array processing approaches such as time difference of arrival (TDOA) [5], steered response power [7], or acoustic intensity vector analysis [9]. An early attempt in joining estimates from the two problems was presented in [8], where beamforming outputs from distributed arrays along with an HMM-GMM classifier are used to build a maximum-a-posteriori criterion on the most probable position in a room of a certain class.

During the last decade, deep neural networks (DNNs) have become the most established method on SED, offering ample modeling flexibility and surpassing traditional machine learning methods when trained with adequate data [13]. Recently, DNNs have also been explored for machine learning-based source localization [14]–[16] with promising results. Hence, DNNs seem like a good candidate for joint modeling of localization and detection in the SELD task. The first works we are aware of that use this approach are [11] and [12]. Hirvonen [11] proposed to set joint modeling as a multilabel-multiclass classification problem, mapping two event classes to eight discrete angles in azimuth. A convolutional neural network (CNN) was trained to infer probabilities of each sound class at each position, after which a predefined threshold was used to decide the final class presence and location. Adavanne et al. [12] proposed as an alternative a regression-based localization approach. Modeling was performed by a convolutional and recurrent neural network (CRNN) with two output branches, one performing SED and the other localization. In the localization branch, one regressor per class returned continuous

Manuscript received Month, Day, 2020; revised Month, Day, 2020. This work received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND.

A. Politis, A. Mesaros, S.Adavanne, T. Heittola and T. Virtanen are with the Faculty of Information Technology and Communication Sciences, Tampere University, Finland, e-mail: {archontis.politis, annamaria.mesaros, tuomas.virtanen}@tuni.fi

[1]http://dcase.community/challenge2019/

azimuth and elevation angles. Binary thresholding was used in the detection branch to indicate the temporal activity of each class and that output was used to gate the respective direction-of-arrival (DoA) output, joining them together during inference. The proposed system, named SELDnet, was extensively compared against other architectures, for a variety of simulated and real data and for different array configurations. Note that both DNN-based proposals were using simple generic input features, such as multichannel power spectrograms in [11] or magnitude and phase spectrograms in [12].

Due to its relevance in the aforementioned applications, the SELD task was introduced for the first time in the DCASE 2019 Challenge and received a remarkable number of submissions for a novel topic. A new dataset of spatialized sound events was generated for the task [17] and a SELDnet implementation was provided by the authors as a baseline for the challenge participants[2]. Beyond the works associated with the challenge [18]–[39], multiple works have followed aiming to address the SELD task in a new way or improve on the limitations of the challenge submissions [40]–[43].

This paper serves three major aims. Firstly, it presents an overview of the first SELD-related challenge. Secondly, it presents common considerations of SELD systems and discusses how these were addressed by the participants, highlighting novel solutions and common elements of the challenge submissions. Thirdly, the performance of the systems is analyzed by addressing the issue of evaluating joint detection and localization. Following the ranking of the systems in the challenge, we calculate confidence intervals for the challenge evaluation metrics and analyze submissions with respect to their performance in detection and localization separately. Additionally, we reevaluate the systems using novel metrics proposed for joint evaluation of localization and detection [44] and investigate correlations between the different metrics and the ranking of the systems.

The paper is organized as follows: Section II presents the task description, dataset, baseline system, and evaluation, as defined in the challenge. Section III introduces and formulates the joint metrics for evaluation of localization and detection. Section IV presents the analysis of submitted systems, including the challenge results and detailed systems characteristics. In Section V we reevaluate the submissions with the new joint metrics, and analyze the results with a rank correlation analysis of the different metrics. Finally, Section VI presents the concluding remarks on the challenge task organization.

II. Sound event detection and localization in DCASE 2019 Challenge

The goal of the SELD task, given a multichannel recording, can be summarized as identifying individual sound events from a set of given classes, their temporal onset and offset times in the recording, and their spatial trajectories while they are active. In the 2019 challenge, the spatial parameter was the DoA in azimuth and elevation, and only static scenes were considered, meaning that each individual sound event instance in the provided recordings was spatially stationary with a fixed location during its entire duration. Some common approaches to SELD systems found in the challenge are depicted in Fig. 1, including a single DNN modeling jointly the class and location of events, separate DNNs for classification and localization, or systems combining DNN-based classification with parametric localization.

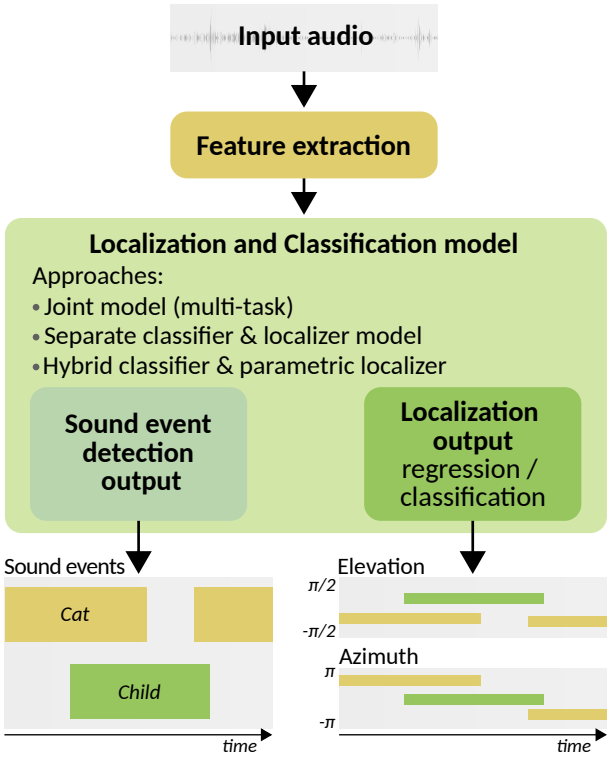


Fig. 1. General SELD system approaches common in the challenge.

A. Dataset

Creating a dataset for a SELD task presents some challenges, reflecting the high complexity of the problem. Ideally, a large range of sound events representative of each sound class should be reproduced at different times and temporal overlaps, at an enormous range of different positions in azimuth, elevation, and possibly distance from the microphones, covering the localization domain of interest. Furthermore, if the system is to be robust to varying acoustic conditions and different spaces, all the previous dimensions should be varied across different rooms. Staging real recordings with this degree of variability is not practical. Acoustic simulations of spatial room impulse responses (RIRs) for various room shapes and positions, and then subsequent convolution of the sound event samples with them is a viable alternative, explored for example in [12]. However, such simulators, with simplifications on room geometry and acoustic scattering behavior, can deviate significantly from real spatial RIRs. Additionally, the non-directional ambient noise characteristic of the function of each space is present in reality, adding another component the SELD system should be robust to.

For DCASE2019, we opted for a hybrid recording-simulation strategy that allowed us to control the detection, localization, and acoustical variability we needed. Real-life impulse responses were recorded at 5 indoor locations in the

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

Hervanta campus of Tampere University, at 504 unique combinations of azimuth-elevation-distance around the recording position. The measurements were covering a domain of 360° in azimuth, -40° ∼ 40° in elevation, and 1∼2m in distance. Additionally, realistic ambient noise was recorded on-site with the recording setup unchanged.

Each spatial sound recording was synthesized as a one-minute multichannel mixture of spatialized sound events convolved with RIRs from the same space, with randomized onsets and source positions, and with up to two simultaneous events allowed. The RIRs were convolved with the isolated sound events dataset[3] provided with DCASE 2016 Task 2 Sound event detection in synthetic audio[4], containing 20 event samples for each of the 11 event classes. Finally, the recorded natural ambient noise from the same space was added to the synthesized mixture, at a 30 dB signal to noise ratio relative to the average power of the sound-event mixture at the array channels. Each mixture was provided in two different 4-channel recording formats, extracted from the same 32-channel recording equipment. The first was a tetrahedral microphone array of capsules mounted on a hard spherical body, while the second was the first-order Ambisonics (FOA) spatial audio format. The two recording formats offer different possibilities in exploiting the spatial information captured between the channels. A development set was available during the challenge[5], and for the evaluation set only the audio without labels was released[6]. The development and evaluation sets consist of 400 and 100 one-minute recordings, respectively. Half of the material has no overlapping events, while the other half has two overlapping events active for most of its duration. Note that two simultaneous events of the same class can occur in the overlapping case. A detailed description of the generation of the dataset is given in [17].

B. Baseline system

The SELDnet architecture of [12] was provided as the baseline architecture of the challenge. The rationale behind this choice was its conceptual and implementation simplicity, and its generality with respect to input features. Furthermore, even though SELDnet was very recent and had the best results between the tested methods in its publication, it still left a significant margin for improvements with realistic data, both at localization and detection accuracy. The architecture of the system is depicted in Fig. 2. It consists of three convolutional layers modeling spatial interchannel and sound event intrachannel time-frequency representations, followed by two bi-directional recurrent layers with gated recurrent units (GRU) capturing longer temporal dependencies in the data. The following two output branches of fully connected layers correspond to the individual tasks of SED and DoA estimation. The SED output is optimized with a cross-entropy loss, while the DoA output is optimized using the mean squared error

[3]https://archive.org/details/dcase2016_task2_train_dev

[4]http://dcase.community/challenge2016/task-sound-event-detection-in-synthetic-audio

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

[6]https://zenodo.org/record/3066124

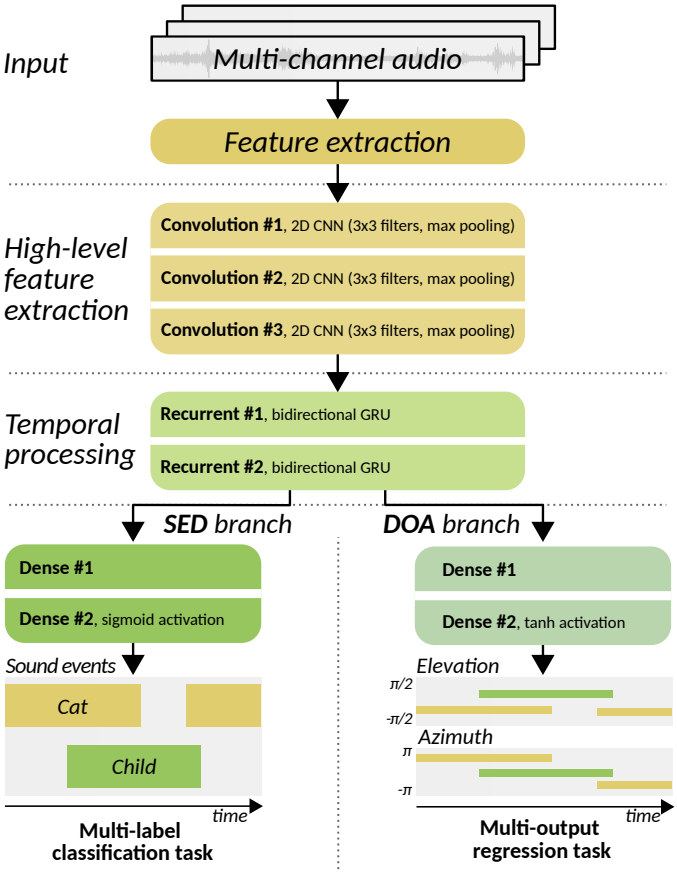


Fig. 2. Detailed SELDnet network architecture of the baseline.

of angular distances between reference and predicted DoAs. Contrary to the original SELDnet in [12] which was outputting Cartesian vector DoAs, the implementation for the challenge is returning directly azimuth and elevation angles. The network takes as input multichannel magnitude and phase spectrograms, stacked along the channel dimension. Reference SED outputs are expressed with one-hot encoding and reference DoAs with azimuth and elevation angles in radians. The network is trained using the Adam optimizer with a weighted combination of the two output losses, with more weight given to the localization loss. More details on the SELDnet challenge implementation can be found in [17].

C. Evaluation and ranking

In this first implementation of the challenge the submitted systems were evaluated with respect to their detection and localization performance individually. For SED, the detection metrics were the $F1$-score and error rate (ER) computed in non-overlapping one-second segments [45]. For DoA estimation, two additional frame-wise metrics were used. The first is a conventional *directional error* (DE) expressing the angular distance between reference and predicted DoAs. Since multiple simultaneous estimates are possible, references and predictions need to be associated before errors can be computed. The Hungarian algorithm [46] was used for that purpose, and the final DE was computed as the minimum cost association, divided with the number of associated DoAs. Since DE does not reflect on how successfully a system

detects localizable events, a second recall-type metric was introduced, termed *frame recall* (FR). Due to a more general introduction and reformulation of the metrics, DE is renamed in this work as *localization error* (LE), while FR is renamed as *event count recall* (ECR).

For a detailed picture of the overall performance, the submissions were ranked individually for each of the four $(F1, ER, LE, ECR)$ metrics. Hence, the jth submission had ranks $I_{F1}(j), I_{ECR}(j)$ based on its position after sorting the $F1, ECR$ results in descending order, and ranks $I_{ER}(j), I_{LE}(j)$ after sorting the ER, LE results in ascending order. A total ranking $I_{\text{tot}}(j)$ aiming to indicate systems achieving good performance in all metrics or exceptional performance in most of them, was obtained by summing the individual ranks $I_{F1}(j) + I_{ECR}(j) + I_{ER}(j) + I_{LE}(j)$ and sorting the results in increasing order.

III. Joint measurement of localization and detection performance

Sound localization and sound event detection are traditionally two different areas of research, but the recent research addresses joint modeling and prediction of the two, motivating a joint evaluation. An example case to illustrate the main drawback of employing separate evaluations for detection and localization (similar to Subsection II-C) is visualized in Fig. 3. Both the participating systems have detected the two sound events correctly, however, their spatial positions are swapped. Using a standalone detection metric will evaluate if the system has correctly predicted the presence of the sound events (without regard to their position), and similarly, a standalone localization metric will evaluate the spatial errors between the closest sound pairs (ignoring the underlying sound classes). Hence, those metrics individually give the exact same score for both systems A and B, even though it is obvious that system B is inferior to A.

A. Metrics formulation

Since a spatial event is not distinguished only by its class, but also by its location, measurement ideally happens at the event level. Let us consider a SELD system that at a given temporal step predicts a set of M events $P = \{p_1, ..., p_i, ..., p_M\}$, where each event prediction is associated with a class label index $\tilde{b}_i$ and a positional vector $\tilde{\mathbf{x}}_i$, such that $p_i = \{\tilde{b}_i, \tilde{\mathbf{x}}_i\}$. At the same time, N reference events exist as $R = \{r_1, ..., r_j, ..., r_N\}$, with each reference event being of class index b_j at position $\mathbf{x}_j$, denoted as $r_j = \{b_j, \mathbf{x}_j\}$. We assume a total of C possible class labels that are ordered, such that $b \in [1, .., C]$. Note that contrary to traditional SED, where predictions and references are class based, it is possible that more than one of the events in P or R are of the same class.

We begin by considering localization-only metrics, neglecting classification. Every combination of prediction $\tilde{\mathbf{x}}_i$ and reference $\mathbf{x}_j$ is associated spatially with an appropriate distance metric $d(\tilde{\mathbf{x}}_i, \mathbf{x}_j)$. Two most common examples are

$$d(\tilde{\mathbf{x}}_i, \mathbf{x}_j) = \arccos\left(\frac{\tilde{\mathbf{x}}_i \cdot \mathbf{x}_j}{||\tilde{\mathbf{x}}_i||||\mathbf{x}_j||}\right) \quad \text{angular distance} \quad (1)$$

$$d(\tilde{\mathbf{x}}_i, \mathbf{x}_j) = ||\tilde{\mathbf{x}}_i - \mathbf{x}_j|| \quad \text{Cartesian distance} \quad (2)$$

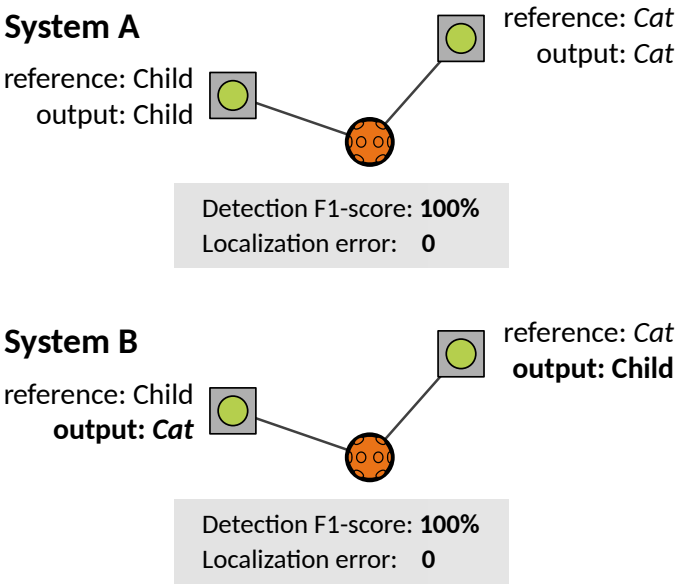


Fig. 3. Example reference and predicted sound events and locations. Circles denote reference sounds, rectangles system output. Two systems evaluated separately for detection and localization performance. Based on the measured performance, they both have perfect score.

In this work evaluation is based on angular distances since only directions of events are measured, instead of absolute positions. All such distances can be expressed with an $M \times N$ distance matrix $\mathbf{D}$, where each element is given by $[\mathbf{D}]_{ij} = d(\tilde{\mathbf{x}}_i, \mathbf{x}_j)$. Before measuring a mean LE across events, references and predictions should be associated using, for example, a minimum cost assignment algorithm such as the Hungarian algorithm, $\mathbf{A} = \mathcal{H}(\mathbf{D})$. The association should ensure that if predictions are more than the references $M > N$, only N predictions are associated with N references, while if predictions are less than the references $M < N$, only M references are associated with M predictions. The non-associated $M - N$ predictions in the first case or the non-associated $N - M$ references in the second case, would constitute an equal number of false alarms (false positives FP) or misses (false negatives FN) respectively, in a detection-like localization metric. Consequently, if there are no predictions, all references are counted as N misses, and similarly if there are no references, all predictions are counted as M false alarms, while in such cases the LE is undefined. Based on the above, the $M \times N$ binary association matrix $\mathbf{A}$ can have maximum one unity entry at each column and row, meaning that only $K = \min(M, N) = ||\mathbf{A}||_1$ predictions and references are associated and contribute to the LE

$$LE = \frac{1}{K}\sum_{i,j} a_{ij} d_{ij} = \frac{||\mathbf{A} \odot \mathbf{D}||_1}{||\mathbf{A}||_1}, \quad (3)$$

where $d_{ij} = [\mathbf{D}]_{ij}$, $a_{ij} = [\mathbf{A}]_{ij}$, $||\cdot||_1$ is the $L_{1,1}$ entrywise matrix norm, and $\odot$ the entrywise matrix product.

The above localization precision gives a partial performance picture because it does not take into account misses or false alarms of localized sounds. To that purpose, we introduce a

simple metric termed *localization recall* (LR), expressed as

$$LR = \frac{\sum_l \min(M^{(l)}, N^{(l)})}{\sum_l N^{(l)}} = \frac{\sum_l ||\mathbf{A}^{(l)}||_1}{\sum_l N^{(l)}}, \tag{4}$$

where the summation $\sum_l$ happens across $l = 1, ..., L$ temporal frame outputs or some other preferred averaged segmental representation, and $M^{(l)}, N^{(l)}$ are the number of predictions and references at the lth frame or segment. Finally, a related but more concentrated metric of interest may be the ratio of frames or segments for which the system detects the correct number of references $M = N$. We name this metric *event count recall* (ECR). ECR corresponds to

$$ECR = \frac{\sum_l \mathbb{1}\left(M^{(l)} = N^{(l)}\right)}{L}, \tag{5}$$

and $\mathbb{1}(\cdot)$ is the indicator function, returning 1 if its argument is true, and 0 otherwise. Note that ECR was termed *frame recall* in the challenge evaluation, and in [12], [14], but we opted here for a more descriptive name of its counting objective.

Often, a localization method needs to be evaluated only under a certain level of spatial precision, usually expressed through an application-dependent threshold Θ. Such a threshold on the above metrics can be applied by constructing an $M \times N$ binary matrix $\mathbf{T}$ with unity entries only on the reference-prediction pairs that are closer than the threshold, $[\mathbf{T}]_{ij} = \mathbb{1}([\mathbf{D}]_{ij} \leq \Theta)$. The number of associated predictions that pass the threshold are then given by $K_{\leq\Theta} = ||\mathbf{T} \odot \mathbf{A}||_1$. The thresholded metrics are

$$LE_{\leq\Theta} = \frac{1}{K_{\leq\Theta}} \sum_{i,j} t_{ij} a_{ij} d_{ij} = \frac{||\mathbf{T} \odot \mathbf{A} \odot \mathbf{D}||_1}{||\mathbf{T} \odot \mathbf{A}||_1} \tag{6}$$

$$LR_{\leq\Theta} = \frac{\sum_l K^{(l)}_{\leq\Theta}}{\sum_l N^{(l)}} = \frac{\sum_l ||\mathbf{T}^{(l)} \odot \mathbf{A}^{(l)}||_1}{\sum_l N^{(l)}} \tag{7}$$

$$ECR_{\leq\Theta} = \frac{\sum_l \mathbb{1}\left(K^{(l)}_{\leq\Theta} = N^{(l)}\right)}{L}, \tag{8}$$

with $t_{ij} = [\mathbf{T}]_{ij}$. Note that the thresholded $LE_{\leq\Theta}$ in Eq. (6) is undefined when there are no associations passing the threshold $K_{\leq\Theta} = 0$.

Considering the fact that events have a class label in SELD, it is more informative to measure localization performance only between events that are correctly classified (class-aware localization). Similarly, we may want to impose a spatial constraint on correct classifications, such that events classified correctly, but very far from their spatial reference are considered invalid (location-aware detection). For both modes, we:

1) Find subsets $P_c = \{p_i | \tilde{b}_i = c\}$ of predictions and $R_c = \{r_j | b_j = c\}$ of reference events classified on class $c \in [1, ..., C]$. The resulting class-specific number of predictions is M_c and of references N_c.
2) Compute a class-dependent $M_c \times N_c$ distance matrix $\mathbf{D}_c$ between predictions P_c and references R_c, and compute the respective association matrix $\mathbf{A}_c = \mathcal{H}(\mathbf{D}_c)$.
3) Determine a suitable application-specific spatial threshold Θ, for location-aware detection. Construct the thresholding binary matrix $\mathbf{T}_c$ from $\mathbf{D}_c$, and determine the number of associated predictions $K_c = ||\mathbf{A}_c||_1 = \min(M_c, N_c)$, and the number of associated predictions which pass the threshold $K_{c,\leq\Theta} = ||\mathbf{T}_c \odot \mathbf{A}_c||_1$.
4) After association, count true positives TP, false negatives FN, and false positives FP as follows:

$$TP_{c,\leq\Theta} = K_{c,\leq\Theta} \tag{9}$$

$$FP_{c,\leq\Theta} = \max(0, M_c - N_c) + \min(M_c, N_c) - K_{c,\leq\Theta} \tag{10}$$

$$FN_c = \max(0, N_c - M_c). \tag{11}$$

A simple example is illustrated in Fig. 4, where the reference annotation contains four sound events: *dog*, *dog*, *car horn*, and *child*, while the system output contains three: *dog*, *car horn*, and *cat*, at their respective positions. The joint evaluation will compare for correctness of both the labels and the locations, therefore it will characterize the localization error in the *dog-dog* pair and the *car horn-car horn* pair, and consider the other events as errors (false positives and false negatives). Note that with the above setup false negatives do not depend on the threshold, while false positives include both the extraneous predictions and associated predictions that did not pass the threshold (the *car horn* example in Fig. 4). Based on the above, we are able to measure location-aware detection metrics such as precision, recall, F1-score, or error rates.

Regarding class-aware localization, we compute the localization error (LE_c) and localization recall (LR_c) of Eq. (3)–(4) only between predictions and references of class c

$$LE_c = \frac{||\mathbf{A}_c \odot \mathbf{D}_c||_1}{||\mathbf{A}_c||_1} \tag{12}$$

$$LR_c = \frac{\sum_l ||\mathbf{A}^{(l)}_c||_1}{\sum_l N^{(l)}_c}. \tag{13}$$

The overall *class-dependent* LE_{CD}, LR_{CD}, are computed as the class means of Eq. (12)–(13)

$$LE_{CD} = \frac{1}{C \cdot L} \sum_c \sum_l LE^{(l)}_c \tag{14}$$

$$LR_{CD} = \frac{1}{C} \sum_c LR_c. \tag{15}$$

In some applications it may be of interest to have both class-dependent, and thresholded localization metrics, similar to Eq. (6)–(8). In the joint measurement results of this study we use the non-thresholded versions of Eq. (12)–(13). It is also worth noting that different thresholds per class Θ_c may be accommodated in the above framework, to reflect different spatial tolerances for certain classes depending on the application. In our evaluation we opted for non-thresholded localization metrics since we deemed it more beneficial to have a localization measurement of all detected estimates in a class, providing complementary information to the spatially-thresholded detection metrics.

It is worth noting here the relation between the proposed metrics and dedicated tracking metrics such as the OSPA [47] or the CLEAR MOT [48] metrics, which evaluate the performance of systems in identifying distinct contiguous spatial trajectories from instantaneous spatial estimates. A form of

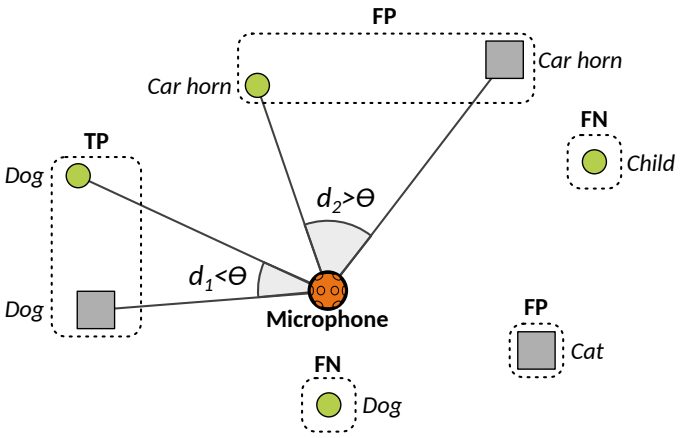


Fig. 4. Example reference and predicted sound events and locations. Circles denote reference sounds, rectangles system output. The dashed rectangles indicate associated predictions with references, and d_1, d_2 are the respective angular distances expressing localization error. The TP, FP, FN indicate how the respective predictions and references contribute to the true positive, false positive, and false negative count.

tracking occurs in SELD systems through location-aware classification; the positional estimates of a source emitting a signal of a certain type are joined to a consistent spatiotemporal trajectory when that class is detected active. However, tracking metrics evaluate trajectory consistency without having to resort to classes and they penalize identity switches, something that the proposed metrics do not do e.g. in the case of two simultaneous events of the same class.

B. Segment-based measurement

Segment-based detection metrics generalize the frame-based binary activity of sound events to its corresponding activity at segment-level and are common in SED. In [45], this generalization is done by considering an event to be active at segment-level if it is active in at least one frame within the segment. A similar generalization of the localization metrics to a different time-scale can be formulated through a spherical mean DoA vector or Cartesian mean positional vector $\hat{\mathbf{x}}$ of all predictions $\tilde{\mathbf{x}}^{(l)}$ of the corresponding event within the segment, before localization errors are measured. Alternatively, the average localization error within a segment can be computed based on the frame-based pairs of reference and predicted events. Both approaches are introduced and compared in [44] with comparable results. Herein, we present results evaluated in one-second segments, apart from the reevaluation in Sec. V where additional frame-level localization results are included in the analysis.

IV. Challenge results

Even though the SELD task was introduced in DCASE2019 for the first time, it attracted a lot of interest and received the second highest number of submissions among other tasks. In total 58 systems were submitted, from a total of 22 teams consisting of in total 65 members. The participants were affiliated with 16 universities and 8 companies.

A. Overall challenge results

The overall results of the challenge are presented in Table I. Only the best system of each team is presented and the systems are ordered by their official challenge rank as described in Section II-C. In addition to the results displayed on the challenge webpage, this table includes the 95% confidence intervals for each separate metric, estimated using the jackknife procedure presented in [1]. The method is a resampling technique that estimates a parameter from a random sample of data for a population using partial estimates [49]. Confidence intervals by jackknifing are coarse approximations, but applicable in cases where the underlying distribution of the parameter to be estimated is unknown. In our case the parameters are metrics that depend on individual combinations of active sounds at each time and the jackknife method allows estimating the confidence intervals without making any assumption on their distribution. The partial estimates for all metrics were calculated in a leave-one-out manner, excluding, in turns, one audio file from the evaluation set.

Considering the best-performing system of each team, 17 out of the 22 submitted systems ranked higher than the baseline system using the official ranking method. In terms of the individual metrics, 17 systems had better ER and $F1$-scores than the baseline, with the best ER and $F1$-scores of 0.06 [20], [21] and 96.7% [21] respectively. Similarly, 18 systems had better LE and 14 systems had higher ECR, with the best LE of 2.7° [25] and ECR of 96.8% [18].

The top-10 systems of Table I are illustrated with respect to detection metrics in Fig. 5a and localization metrics in Fig. 5b. The best system in both these plots is in their corresponding top left corner. We observe that the ranking order of the submitted systems is different for detection and localization metrics. For instance, the best system according to detection metrics - *He THU* [21] (Fig. 5a top-left corner), has a high LE compared to the other top-10 systems and hence achieves an overall rank of four. Similarly, although *Chang HYU* [25] achieved the best LE among the top-10 systems, its detection performance was among the poorest of top-10 systems and hence achieved a rank of eight. In general, ER and $F1$-scores of event detection are correlated and hence all the submitted systems are observed along the diagonal. This diagonal behavior is not observed with the localization metrics as LE and ECR are only weakly correlated.

The system characteristics of all the submissions are summarized in Table II. All systems had at least one deep learning component in their approach. Specifically, apart from [36] and [38] that employed a CNN architecture with no recurrent layers the remaining 20 systems employed different versions of the baseline CRNN architecture as one of their components. Four of the submitted systems employed model-based parametric DoA estimation [20], [23], [32], [35] along with CRNN-based classification. The best purely parametric DoA approach [23] achieved the 6th position. Among the DNN-based SELD methods, nine of them employed multi-task learning [50] for joint SED and DoA estimation. The remaining systems, including the top ranked system [18], employed separate networks for SED and DoA estimation

TABLE I
CHALLENGE RESULTS OF SUBMITTED SYSTEMS. THE RANK IS BASED ON THE CUMULATIVE RANK BASED ON THE FOUR CALCULATED METRICS. BEST SYSTEM PER TEAM ACCORDING TO THE OFFICIAL CHALLENGE RANKING. BEST SCORE INDICATED FOR THE SEPARATE METRICS.

Rank	System	ER	F1	LE	ECR
1	Kapka_SRPOL_2 [18]	0.08 ± 0.01	94.7 ± 0.8	3.7 ± 0.6	**96.8 ± 0.6**
2	Cao_Surrey_4 [19]	0.08 ± 0.01	95.5 ± 0.4	5.5 ± 0.7	92.2 ± 1.0
3	Xue_JDAI_1 [20]	**0.06 ± 0.01**	96.3 ± 0.5	9.7 ± 1.3	92.3 ± 1.3
4	He_THU_2 [21]	**0.06 ± 0.01**	**96.7 ± 0.4**	22.4 ± 1.7	94.1 ± 1.0
5	Jee_NTU_1 [22]	0.12 ± 0.01	93.7 ± 0.5	4.2 ± 0.5	91.8 ± 1.0
6	Nguyen_NTU_3 [23]	0.11 ± 0.01	93.4 ± 0.7	5.4 ± 0.4	88.8 ± 1.6
7	MazzonYasuda_NTT_3 [24]	0.10 ± 0.01	94.2 ± 0.5	6.4 ± 0.9	88.8 ± 1.3
8	Chang_HYU_3 [25]	0.14 ± 0.01	91.9 ± 0.5	**2.7 ± 0.3**	90.8 ± 1.3
9	Ranjan_NTU_3 [26]	0.16 ± 0.01	90.9 ± 0.8	5.7 ± 0.5	91.8 ± 1.0
10	Park_ETRI_1 [27]	0.15 ± 0.01	91.9 ± 0.6	5.1 ± 0.7	87.4 ± 1.3
11	Leung_DBS_2 [28]	0.12 ± 0.01	93.3 ± 0.6	25.9 ± 1.3	91.1 ± 1.3
12	Grondin_MIT_1 [29]	0.14 ± 0.01	92.2 ± 0.7	7.4 ± 0.6	87.5 ± 1.7
13	ZhaoLu_UESTC_1 [30]	0.18 ± 0.01	89.3 ± 0.8	6.8 ± 0.9	84.3 ± 1.4
14	Rough_EMED_2 [31]	0.18 ± 0.01	89.7 ± 0.7	9.4 ± 0.9	85.5 ± 1.5
15	Tan_NTU_1 [32]	0.17 ± 0.02	89.8 ± 0.9	15.4 ± 1.4	84.4 ± 2.1
16	Cordourier_IL_2 [33]	0.22 ± 0.01	86.5 ± 0.8	20.8 ± 1.2	85.7 ± 1.5
17	Krause_AGH_4 [34]	0.22 ± 0.02	87.4 ± 0.9	31.0 ± 1.0	87.0 ± 1.8
18	Adavanne_TAU_FOA [17]	0.28 ± 0.02	85.4 ± 0.9	24.6 ± 1.1	85.7 ± 1.9
19	Perezlopez_UPF_1 [35]	0.29 ± 0.03	82.1 ± 1.5	9.3 ± 0.4	75.8 ± 2.5
20	Chytas_UTH_1 [36]	0.29 ± 0.01	82.4 ± 0.8	18.6 ± 1.3	75.6 ± 2.4
21	Anemueller_UOL_3 [37]	0.28 ± 0.02	83.8 ± 1.2	29.2 ± 1.1	84.1 ± 2.3
22	Kong_SURREY_1 [38]	0.29 ± 0.01	83.4 ± 0.9	37.6 ± 1.7	81.3 ± 1.9
23	Lin_YYZN_1 [39]	1.03 ± 0.01	2.6 ± 0.7	21.9 ± 8.2	31.6 ± 2.5

TABLE II
SUMMARY OF SUBMITTED SYSTEMS. THE RANK IS BASED ON THE CUMULATIVE RANK BASED ON THE FOUR CALCULATED METRICS. BEST SYSTEM PER TEAM ACCORDING TO THE OFFICIAL CHALLENGE RANKING.

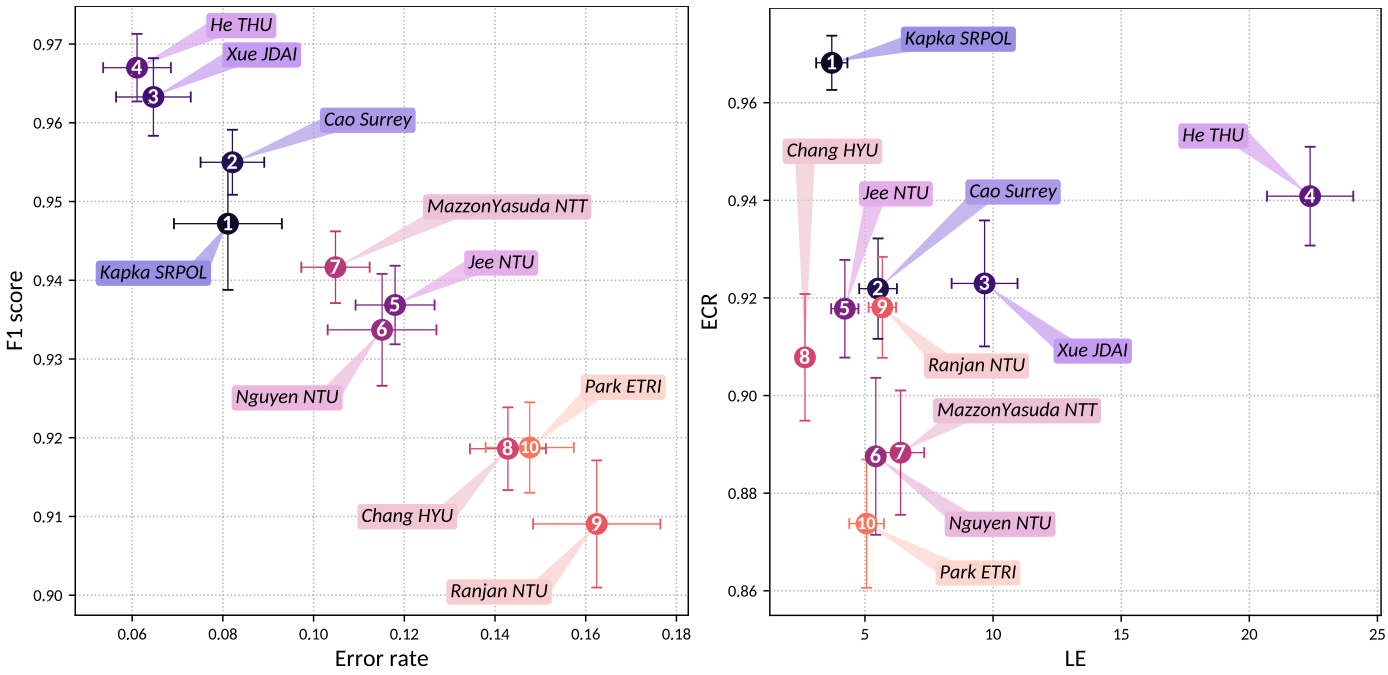
	System	Audio	Features	Classifier	Multi-task
1	Kapka_SRPOL_2 [18]	AMB	Phase and magnitude spectra	CRNN	×
2	Cao_Surrey_4 [19]	Both	Log-mel, GCC, and intensity vectors	CRNN ensemble	×
3	Xue_JDAI_1 [20]	MIC	Log-mel, Q-transform, multiple spectra	CRNN ensemble, parametric DoA	✓
4	He_THU_2 [21]	AMB	Log-mel, phase, and magnitude spectra	CRNN	×
5	Jee_NTU_1 [22]	MIC	Log-mel spectra and GCC	CRNN	×
6	Nguyen_NTU_3 [23]	AMB	Log-mel, phase, and magnitude spectra	CRNN, parametric DoA	×
7	MazzonYasuda_NTT_3 [24]	Both	Log-mel spectra and GCC	CRNN, ResNet ensemble	×
8	Chang_HYU_3 [25]	MIC	Log-mel spectra, cochleagram, and GCC	CRNN, CNN	×
9	Ranjan_NTU_3 [26]	MIC	Log-mel and phase spectra	ResNet RNN	×
10	Park_ETRI_1 [27]	Both	Log-mel and intensity vectors	CRNN, TrellisNet	✓
11	Leung_DBS_2 [28]	AMB	Log-magnitude, phase, and cross spectra	CRNN ensemble	✓
12	Grondin_MIT_1 [29]	MIC	Phase and magnitude spectra, GCC and TDOA	CRNN ensemble	×
13	ZhaoLu_UESTC_1 [30]	MIC	Log-mel spectra	CRNN	✓
14	Rough_EMED_2 [31]	MIC	Phase and magnitude spectra	CRNN	×
15	Tan_NTU_1 [32]	MIC	Log-mel spectra and GCC	ResNet RNN, parametric DoA	×
16	Cordourier_IL_2 [33]	MIC	Phase and magnitude spectra, and GCC	CRNN ensemble	✓
17	Krause_AGH_4 [34]	AMB	Phase and magnitude spectra	CRNN ensemble	✓
18	Adavanne_TAU_FOA [17]	AMB	Phase and magnitude spectra	CRNN	✓
19	Perezlopez_UPF_1 [35]	AMB	Log-mel spectra	CRNN, parametric DoA	×
20	Chytas_UTH_1 [36]	MIC	Raw audio and power spectra	CNN ensemble	×
21	Anemueller_UOL_3 [37]	AMB	Group-delay and magnitude spectra	CRNN	✓
22	Kong_SURREY_1 [38]	AMB	Magnitude spectra	CNN	✓
23	Lin_YYZN_1 [39]	AMB	Phase and magnitude spectra	CRNN	✓

and performed engineered data-association of their respective outputs. Finally, there was no significant improvement in SELD performance with the choice of either of the two audio formats in the dataset. Among the top 10 ranked systems, four of them used the microphone array format, three used the Ambisonic format, and the rest used both formats as input.

B. Analysis of individual systems

A detailed analysis of some of the systems follows, along with a summary of the most prominent architectural, input feature, or training characteristics.

Kapka & Lewandowski (*Kapka SRPOL*) [18] was the top performing system of the challenge, with very high performance in both localization and detection. There was minimal feature engineering and the pure magnitude and phase spectrograms of the FOA format were used as input. However, the approach was highly coupled to the task, by splitting it into four well defined subtasks and then dedicating one CRNN model to infer each one of them. The subtasks were: a) estimation of the number of sources, b) estimation of DoA



(a) Detection results

(b) Localization results

Fig. 5. Separately calculated detection and localization performance of top 10 systems (best system per team). The official rank of the systems is indicated in the center of the marker for each scatter plot.

for an active source, c) estimation of a second DoA in the case that two simultaneous events are detected, d) classification of events whose number equals the number of detected sources. Well-engineered post-processing of outputs, from source count to localization to event durations to classification, coupled the method to prior knowledge of the dataset and ensured consistent association and information flow between modules. It is worth noting that their architecture seems able to resolve two simultaneous instances of the same class at different directions. Since the architecture relied on prior knowledge, such as a maximum of two simultaneous sources and discrete DoAs at 10° intervals, it was not as general as most of the other approaches.

Cao et al. (*Cao Surrey*) [19], had the second best performing system, following the first one closely. However, the authors kept the general SELDnet architecture and advanced it with a number of informed domain-specific choices. The most important ones seem to be improved input features and disassociation of the detection and localization losses. In particular, the losses were separated by duplicating the SELDnet and training each clone for SED and localization separately. Here, the ground truth SED activations were used as masks on the localization loss. Additionally, they used both FOA and MIC input and ensemble averaging. According to ablation studies in [19], the better input features and the two-stage training architecture have a drastic effect in performance.

The system of Xue et al. (*Xue JDAI*) [20] outperformed the first two in detection results, but had lower localization performance resulting in the third best average rank. Its success seems to be a combination of multiple spectral and spatial features and elaborate post-processing. DoA estimation from the CRNN model was also abandoned in favour of a traditional SRP estimation, refined by the former only in the case of simultaneous events. Additionally, separate CNN branches were used for SED and localization features, before being merged at the recursive layers.

The fourth best system of Zhang et al. (*He THU*) [21] follows the same architecture as [19]. It had the best SED performance overall, but its localization accuracy was only marginally better than the baseline. The large difference compared to the second system may be due to the basic spectrogram feature for localization, instead of the more effective directional features used in [19]. On the other hand, the higher detection performance may be attributed to the SpecAugment [51] data augmentation strategy used. The same architecture was also employed by the fifth best system of Jee et al. [22], aiming to improve its performance. They introduced a number of incremental modifications to the SED features, CRNN layers, pooling, and activation functions, along with a mixup [52] data augmentation strategy, without, however, achieving better results at the challenge evaluation.

Nguyen et al. (*Nguyen NTU*) [23] took the concept of independent localization and detection to its extreme, performing them separately and then associating DoAs to overlapping detected events randomly. Good overall performance brought them to the sixth place. Note that their approach exploits the fact that detection and estimation performance are evaluated independently and correct associations between the two are not measured, as discussed in the next sections.

The next best system of Mazzon et al. (*MazzonYasuda NTT*) [24] was also based on the architecture of the second best system [19], trying to improve on it with a Resnet network

replacing CNNs, an elaborate ensemble strategy, and, most importantly, an original spatial data augmentation approach exploiting the rotation and reflection properties of the spherical harmonic bases encoding the sound field in Ambisonics [53]. The authors limited the input features to only GCC-PHAT for both FOA and microphone array signals, potentially limiting their effectiveness for the FOA set which encodes DoA information by amplitude differences.

Noh et al. (*Chang HYU*) [25] added an overall sound activity detection model on top of the SED one. Two additional independent CRNN models were trained to detect presence of one or two events respectively, using cochleagram features as input. Their binary outputs were used to select whether none, one, or two event classes with the highest probabilities of the dedicated CRNN SED model were outputted. The authors employed just a CNN network for DoA estimation, performed as a classification task on 324 classes, inferred from the grid of potential DoAs in the dataset. Interestingly, their model achieved the lowest localization error in the challenge. That may be attributed to their DoA classification matching the DoA discretized grid in the dataset, along with their spatial data augmentation technique, mixing recordings from non-overlapping events to generate additional overlapping segments for training. No information was provided on how or if DoAs were associated with events and from further analysis on the following sections we assume that such association was done randomly, as in [23]. The same approach of independent SED and localization networks, a classification-based DoA estimation, and random association between the two was followed by the next best performing system of Ranjan et al. (*Ranjan NTU*) [26]. Additionally, the authors replaced CNN layers with Resnets in the typical CRNN networks followed by most participants.

The tenth-best performing system of Park et al. (*Park ETRI*) [27] attempted to combine the success of the two-stage training approach [19] with the assumed consistency of joint-modeling. They performed two stages of weight transfer from separately trained SED and DoA estimation networks, into a new network with a SED and DoA branch trained with a combined detection and localization loss, as in the baseline SELDnet. Additionally they experimented with TrellisNet layers instead of RNNs and alternative activation functions.

We note some interesting investigations in the rest of the submitted systems. Grondin et al. (*Grondin MIT*) [29] used one CRNN for each microphone pair in the array format, performing joint event detection and localization. The network was trained to output intermediate TDOA values, mapped afterwards to DoAs. Tan et al. (*Tan NTU*) [32] was one of the four systems that did not use machine learning for DoA estimation, computing time-domain cross-correlations between microphone pairs and their respective TDOA and converting it to a DoA by a least-squares geometric fit. Krause and Kowalczyk (*Krause AGH*) [34] explored various combinations of layers processing localization and SED features before fusion, as well as early branching for the two tasks. *Grondin MIT* [29] showed similar considerations on the fusion of input features, since the approach of the baseline stacking phase and magnitude spectrograms into a single tensor could be suboptimal. Chytas and Potamianos (*Chytas UTH*) [36] proposed to perform SELD directly from downsampled audio waveforms, with some additional help for SED using power spectrograms. Even though their CNN-only approach underperformed on SED, it showed that competitive localization can be achieved using DNNs directly on time-domain multichannel audio.

Finally, a special mention should go to the system by Perez et al. (*PerezLopez UPF*) [35] since, along with the best performing system of [18], it was the only other system following a localize-before-detect paradigm. Their approach was based on model-based DoA estimation on the FOA format, determination of the number of sources based on the DoA estimates, determination of the event onset/offset, and beamforming towards the prominent DoAs. The beamformed signals, being essentially estimates of separated event signals, were fed to a CRNN classifier for SED. Contrary to the majority of submissions in the challenge, such an architecture is capable of detecting simultaneous instances of the same class localized at different directions.

C. Discussion on submitted systems

One obvious observation on the results is that the SELDnet baseline, as implemented for the challenge, had a suboptimal performance compared to the majority of the submissions. An initial weakness seems to be the input features. A number of submissions indicated that by switching to features with more concentrated information on each of the two tasks, detection (log-mel spectra) or localization (GCC-PHAT arrays, active intensity vectors), improved performance significantly. These three sets of features were the most popular overall in the top submissions, with only the third best system relying on multiple other types of multichannel spectra. It has to be noted though, that the top system [18] used the raw multichannel phase and magnitude spectrograms, indicating that it is possible to perform SELD succesfully with such lower level features, but with model architectures exploiting prior knowledge and coupled tightly to the task.

The most popular network architecture and training choices seem to be the ones introduced by Cao et al. [19]. Essentially, their work disassociates the joint cost function combining SED losses and localization losses as realized in the baseline and trains individual models for each task. The SED and DoA estimates are then associated through a training strategy or assigned randomly between them [23], [25], [26]. It has to be noted that such random association takes advantage of the fact that detection and localization were evaluated independently in the challenge and would not be a good strategy in practice. Ranjan et al. [26] compared the two-stage architecture versus joint-modeling, with clearly improved results with the former. However, it is worth noting that two systems in the top ten places had a single network performing joint-modeling [20], [27], one of them being third best [20].

The SELD paradigm proposed by the SELDnet baseline, where one DoA output is tied to each class, was followed by most submissions including multi-stage approaches [19], [25]. This paradigm is forcing a convenient detect-before-localize approach. However, it limits the output of the system

to only one localized event per class, even in the presence of two same-class instances. Systems that were training an independent localization network as a DoA classification task, were not addressing that problem since association of DoAs to detected classes was ambiguous. The only two submissions that followed a localize-before-detect paradigm used localization information to determine the number and DoAs of events independently of their class [18], [35]. They were then passing that information to classifiers, turning the class-based outputs into event-based outputs and circumventing the same-class multi-instance problem of the detect-before-localize approach.

Certain architectural or training choices were specific to the localization task. Some of the submissions treated DoA estimation as a classification task [25], [26], e.g. similar to other DNN-based localization works [14]–[16], instead of the regression format of the baseline. Xue et al. [20] trained both DoA output formats simultaneously. However, it has to be noted that the systems that relied only on DoA classification were taking advantage of the the small set of 324 fixed DoAs embedded in the dataset. A dataset with a much more dense spatial resolution of possible DoAs, a continuous range of DoAs, or moving sources, may have needed a much larger number of classes to be modeled effectively (e.g. 2522 discrete angles for a resolution of 5° in azimuth and elevation covering the sphere). Moreover, classification-based DoA estimation was found successful in two-stage systems, training independently a DoA network. Joint-modeling of SELD based completely on classification, as pioneered by Hirvonen [11], seems feasible for a small number of classes and directions. Otherwise, such a classifier would require *no. of DoA classes* $\times$ *no. of event classes* outputs. With only a small number of them being positive at each frame, its training would face the issue of an imbalanced dataset. Additionally, training such a large number of classes requires an impractically huge dataset with enough examples for each class. On the other hand, the format of one DoA-regression-output per sound event class does not suffer from those limitations, but it is unable to detect multiple instances of the same class being active at different directions.

Finally, some of the submissions aimed for a parametric DoA estimation instead of a trainable DNN model [20], [23], [32], [35], including the third best system of Xue et al. [20]. Parametric DoA estimation has the advantage that it does not require training and that it is possible to generalize to completely unseen environments, since it requires only knowledge of the directional array response. Moreover, Nguyen et al. [23] had one of the smallest DoA errors in the challenge. However, it can be more susceptible to reverberation than DNN approaches if not accompanied with additional processing, such as detection of single-source dominated time-frequency blocks [23]. Interestingly, Xue et al. [20] did not utilize the provided theoretical steering vectors of the spatial format, but estimated them directly from the data.

V. Reevaluation of challenge entries using joint metrics

We evaluate all the systems submitted to DCASE 2019 Challenge Task 2 using the proposed joint measures in order to determine the most suitable single metric that encompasses all aspects when representing system performance in a single number. We compute all metrics in one-second segments and evaluate the location-aware detection metrics with an angular error threshold of 10 and 30 degrees. The results are presented in Table III, in order of the official challenge rank. Confidence intervals for all metrics were calculated according to the jackknife procedure by leaving out one file at a time for the partial evaluation. New cumulative ranks are estimated similar to the official ranks based on the proposed joint measures for the purpose of system comparison. The top 10 systems from Table III are also presented in Fig. 6.

A. Analysis of systems

The independent localization and evaluation metrics ($ECR, LE, F1, ER$) are more permissive than the joint ones ($LR_{CD}, LE_{CD}, F_{10^\circ}, ER_{10^\circ}$). We chose a threshold of 10° for a relatively strict localization criteria with respect to the average localization error of the systems presented in Table II. A ranking based on the new metrics is expected to be different at least for some of the submissions. Table III presents new ranks computed between class-dependent localization (LR_{CD}, LE_{CD}) and location-dependent classification ($F_{10^\circ}, ER_{10^\circ}$). Systems with equal ranks indicate that the sum of the individual ranks for each pair of metrics was the same. The greatest changes on the top ten systems seem to be induced by the location-dependent classification ($F_{10^\circ}, ER_{10^\circ}$), which is to be expected since it penalizes inadequately localized detections with a strict threshold of 10°.

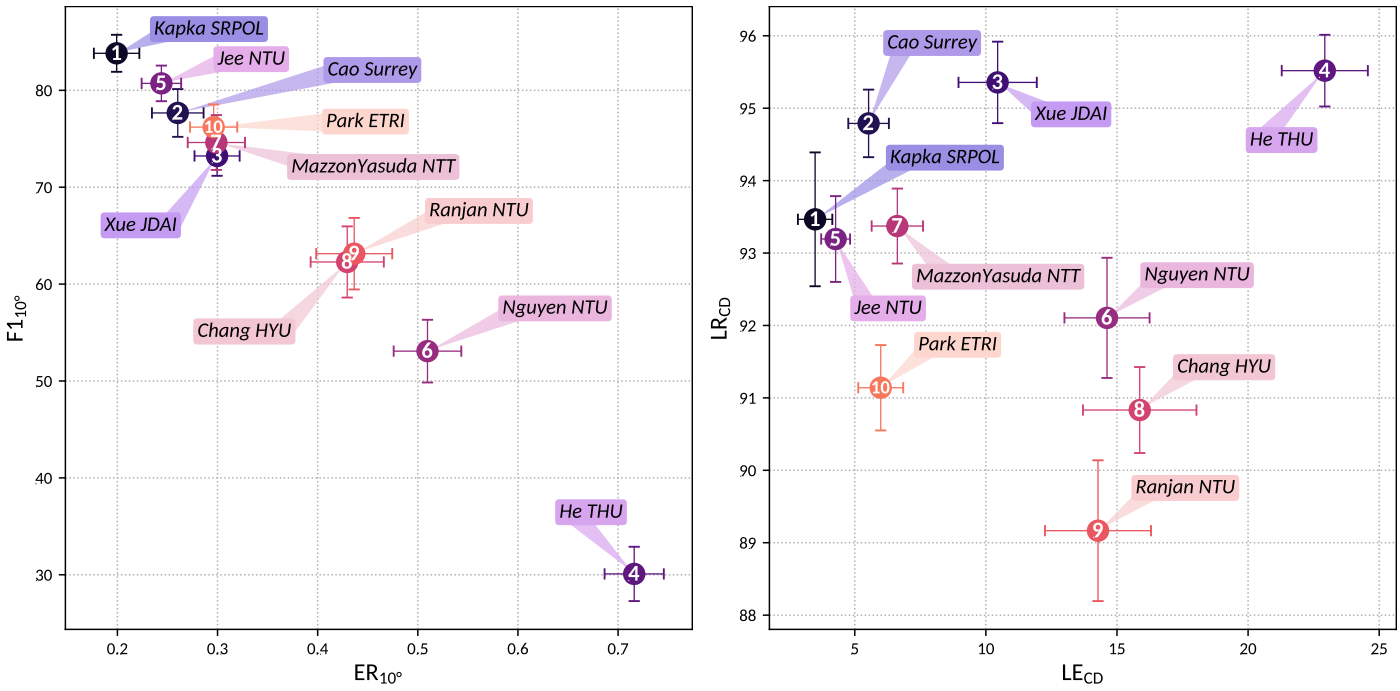
In general, it can be observed that submissions which employed separate localization and detection systems and did not handle association of the two properly were likely to slip in their ranks. This is especially evident on the systems that assigned randomly DoAs to detections, such as *Nguyen NTU* [23] and *Ranjan NTU* [26], including the best localization method of *Chang HYU* [25]. Their association problems are revealed both by their large drop in detection scores ($F_{10^\circ}, ER_{10^\circ}$) and by the large error increase between their original LE and the class-dependent one LE_{CD}.

Methods that performed significantly better detection than localization, such as *Xue JDAI* [20], *He THU* [21], and *Leung DBS* [28] also slipped in their ranks. This is mostly due to three of the original metrics ($F1, ER, ECR$) being directly associated to detection performance, boosting their overall rank. This imbalance is diminished with the new metrics, resulting in the drop of the aforementioned systems.

Among the methods that performed proper data association, the ones who had better localization scores [24], [27], [29], [30], [35], [36] and not the best detection scores improved in their ranks, due to the detection bias of official rankings mentioned above. Two examples worth mentioning are [27], [35]. The multi-task training strategy of *Park ETRI* [27] showed its benefits when evaluated jointly, taking them to 4th place. *PerezLopez UPF* [35] leaped from 19th place below the baseline to 7th place. Both systems achieved such rank advances when evaluated with the strict location-dependent detection ($F_{10^\circ}, ER_{10^\circ}$).

TABLE III
EVALUATION OF DCASE 2019 SUBMISSIONS USING THE JOINT METRICS CALCULATED IN ONE SECOND SEGMENTS. BEST SYSTEM PER TEAM, IN ORDER OF THE OFFICIAL CHALLENGE RANKING.

Official rank	System	LE_{CD}	LR_{CD}	Rank	ER_{10°	F_{10°	Rank	ER_{30°	F_{30°
1	Kapka_SRPOL_2 [18]	**3.5** ± 0.7	93.5 ± 0.9	1	**0.20** ± 0.02	**83.8** ± 1.9	1	0.13 ± 0.02	91.0 ± 1.4
2	Cao_Surrey_4 [19]	5.5 ± 0.8	94.8 ± 0.5	2	0.26 ± 0.03	77.7 ± 2.5	3	**0.13** ± 0.01	**91.0** ± 1.0
3	Xue_JDAI_1 [20]	10.5 ± 1.5	95.4 ± 0.6	5	0.30 ± 0.02	73.2 ± 2.1	6	0.16 ± 0.02	87.2 ± 1.8
4	He_THU_2 [21]	22.9 ± 1.6	**95.5** ± 0.5	8	0.72 ± 0.03	30.1 ± 2.8	16	0.28 ± 0.03	74.6 ± 2.8
5	Jee_NTU_1 [22]	4.3 ± 0.6	93.2 ± 0.6	3	0.24 ± 0.02	80.7 ± 1.8	2	0.15 ± 0.01	90.9 ± 0.8
6	Nguyen_NTU_3 [23]	14.6 ± 1.6	92.1 ± 0.8	9	0.51 ± 0.03	53.1 ± 3.2	13	0.25 ± 0.03	80.4 ± 2.4
7	MazzonYasuda_NTT_3 [24]	6.6 ± 1.0	93.4 ± 0.5	4	0.30 ± 0.03	74.6 ± 2.8	5	0.17 ± 0.01	88.2 ± 1.2
8	Chang_HYU_3 [25]	15.9 ± 2.2	90.8 ± 0.6	13	0.43 ± 0.04	62.3 ± 3.7	10	0.32 ± 0.03	74.4 ± 2.6
9	Ranjan_NTU_3 [26]	14.3 ± 2.0	89.2 ± 1.0	11	0.44 ± 0.04	63.1 ± 3.7	10	0.31 ± 0.03	76.7 ± 2.7
10	Park_ETRI_1 [27]	6.0 ± 0.9	91.1 ± 0.6	6	0.30 ± 0.02	76.2 ± 2.3	4	0.20 ± 0.01	86.9 ± 1.1
11	Leung_DBS_2 [28]	31.4 ± 1.6	92.3 ± 0.7	15	0.84 ± 0.02	17.7 ± 1.7	18	0.43 ± 0.03	59.9 ± 2.6
12	Grondin_MIT_1 [29]	8.0 ± 0.8	91.6 ± 0.8	7	0.40 ± 0.03	65.4 ± 3.1	9	0.19 ± 0.02	88.0 ± 1.3
13	ZhaoLu_UESTC_1 [30]	7.3 ± 1.0	88.3 ± 0.9	10	0.39 ± 0.03	67.5 ± 3.1	8	0.24 ± 0.02	83.8 ± 1.5
14	Rough_EMED_2 [31]	9.7 ± 1.0	88.7 ± 0.8	11	0.50 ± 0.03	55.3 ± 2.8	12	0.24 ± 0.02	83.4 ± 1.5
15	Tan_NTU_1 [32]	19.0 ± 1.8	88.8 ± 1.0	16	0.63 ± 0.02	41.4 ± 2.3	14	0.31 ± 0.03	76.0 ± 2.4
16	Cordourier_IL_2 [33]	22.6 ± 1.4	85.8 ± 0.9	17	0.78 ± 0.03	25.7 ± 2.6	17	0.39 ± 0.02	67.3 ± 2.3
17	Krause_AGH_4 [34]	36.9 ± 1.4	86.1 ± 1.0	19	0.95 ± 0.01	8.3 ± 0.8	21	0.56 ± 0.02	49.5 ± 2.3
18	Adavanne_TAU_FOA [17]	29.7 ± 1.3	83.8 ± 0.9	18	0.95 ± 0.01	10.5 ± 1.1	20	0.53 ± 0.02	56.5 ± 2.2
19	Perezlopez_UPF_1 [35]	5.9 ± 0.4	81.2 ± 1.6	14	0.38 ± 0.03	73.8 ± 1.8	7	0.32 ± 0.03	80.2 ± 1.8
20	Chytas_UTH_1 [36]	19.2 ± 1.5	81.0 ± 1.0	19	0.70 ± 0.02	37.0 ± 2.5	15	0.43 ± 0.02	67.7 ± 2.8
21	Anemueller_UOL_3 [37]	34.5 ± 1.4	82.6 ± 1.2	21	0.97 ± 0.01	7.9 ± 0.9	22	0.60 ± 0.03	46.8 ± 2.3
22	Kong_SURREY_1 [38]	42.7 ± 2.1	82.2 ± 1.0	22	0.92 ± 0.01	11.0 ± 1.4	19	0.65 ± 0.02	41.3 ± 2.5
23	Lin_YYZN_1 [39]	92.7 ± 20.9	1.1 ± 0.4	23	1.04 ± 0.01	0.0 ± 0.0	23	1.04 ± 0.01	0.2 ± 0.2



(a) Location-aware detection results

(b) Class-aware localization results

Fig. 6. Joint detection and localization performance of top 10 systems (best system per team). The official rank of the systems is indicated in the center of the marker for each scatter plot.

Even though the rank for the more permissive 30° location-dependent detection metrics ($F_{30^\circ}, ER_{30^\circ}$) is not displayed in Table III, it is closer to the original challenge ranking. This is explained a) by the more relaxed threshold, which as it increases it causes the metrics to approach their independent detection counterparts. B) by the fact that the threshold is larger than the average LE_{CD} of about 20° between systems.

B. Metrics analysis

The analysis of the metrics was performed using Spearman's rank correlation coefficient [54] to calculate how they correlate to each other. The correlation was calculated between all pairs of considered metrics, using the evaluated performance of all submissions to the task. Our purpose is to determine which single metric is capable of representing the desired properties

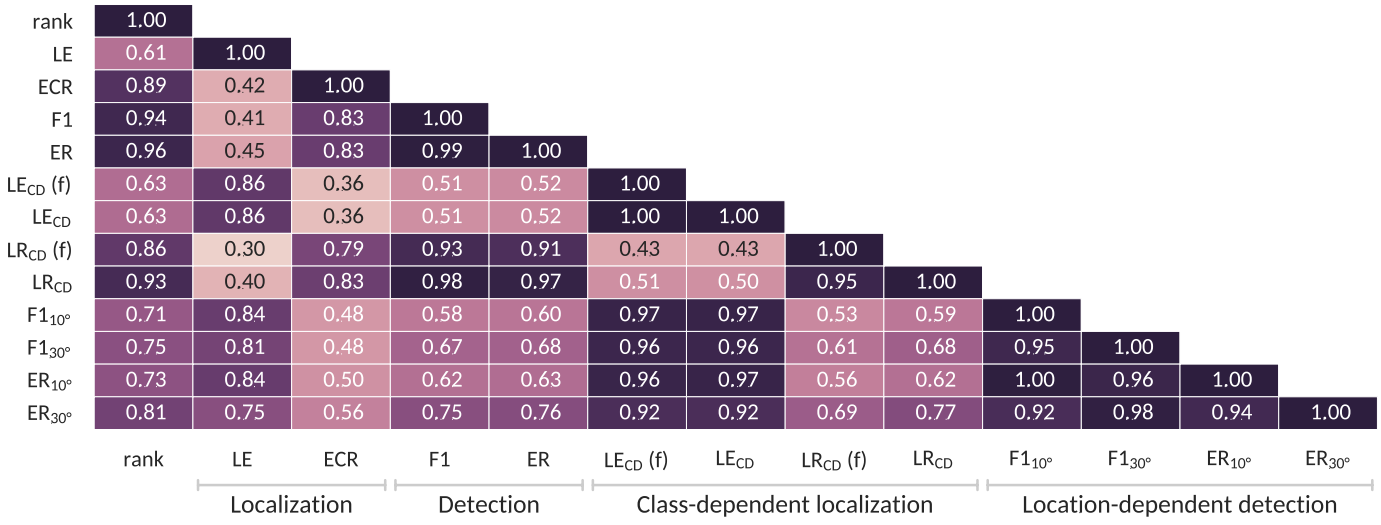


Fig. 7. Correlation between ranking order of submissions according to the different metrics and the official ranking in the challenge.

of the system in terms of localization and detection, instead of using the compound of four separate metrics as done in the challenge ranking. We rank all submissions using each metric separately and evaluate how correlated the different rankings are. Correlation values are presented in Fig. 7. The metrics marked with (f) are calculated frame-wise (in this case 20 ms). Among the four individual metrics (LE, ECR, $F1$, and ER), the detection scores ($F1$ and ER) are highly correlated with the ranking, indicating that good detection performance was important for obtaining a top rank. The localization error is less correlated with the overall rank.

Among the joint metrics, the class-dependent LR_{CD} score is highly correlated with the official ranking, more so for the segment-based than the frame-based measurement. This behavior is noticed in all metrics, with the more permissive metric being more correlated to the overall rank: a) segment-based LR_{CD} is more correlated to the rank than frame-based $LR_{CD}(f)$ and b) metrics with 30° threshold are more correlated to the rank than metrics with 10° threshold. This can be explained by the fact that joint metrics first perform the data association between detected and localized sound sources. The more permissive metrics allow a higher proportion of matches, which in turn is closer to the matching done by the detection-only and separation-only metrics.

We observe similar behavior between metric pairs with and without data association: a) correlation between localization-only metrics LE and ECR is moderate and similar to the one between LE_{CD} and LR_{CD}. b) High correlation is observed between detection-only ER and $F1$; similarly for the corresponding data associated versions. On the other hand, the correlation between detection-only ER and its counterparts ER_{10° or ER_{30° is moderate. Similar behaviour is observed between $F1$ and its counterparts $F1_{10^\circ}$ or $F1_{30^\circ}$. Basically, the data association makes the metrics less permissive (in a similar manner as the higher correlation for the more permissive threshold of 30° than for 10°).

Among the proposed joint-metrics, LR_{CD} has the best correlation (0.93) with the official DCASE2019 rankings, that is presumed to be a good approximation of the overall system performance. However, LE_{CD} is only moderately correlated (0.50) with LR_{CD}, hence, selecting an SELD model based on just LR_{CD} might not always guarantee the best LE_{CD}. On the other hand, the location-aware detection metrics are highly correlated with each other (ER_{10° vs. $F1_{10^\circ}$ or ER_{30° vs. $F1_{30^\circ}$) and have moderately high (0.71-0.81) correlation with the official rank. Furthermore, for a given distance threshold, the error rate metrics are more correlated to the official rank than the F1-scores and they are also highly correlated with LE_{CD}. If choosing a SELD model has to be limited to a single metric only, it seems that the error rate (ER_{10°/ER_{30°) is a suitable choice, since it combines high correlation with the original ranking, with the ranking based on ($F1_{10^\circ}$/$F1_{30^\circ}$), as well as the localization ranking of LE_{CD}. Hence, it is expected to guarantee an overall good SELD performance, a good counterpart F1-score and a low localization error.

C. Discussion

The very high performance of the top ranked systems, of a few degrees of mean localization error and more than 83% F1 score in the stricter setting, reveals additionally that the state of the art can potentially handle more challenging conditions than those reflected on the current dataset. The simulated spatial recordings, even though acoustically realistic, contained only static events well separated between them by at least 10°. Furthermore, the room IRs were captured in large open spaces and at fairly close distances from the microphone resulting in high direct-to-reverberant ratios, while the ambient noise was added at a very high SNR. As a consequence, the spatial and spectral characteristics of the events were not significantly corrupted by them and the methods had to learn mostly a model of the directional array response to infer location. Such conditions of up to two simultaneous foreground sound events of interest at differing directions and at 1–2 m away from the listener, in the presence of reverberation and low background noise can still occur frequently in real-life, but they are, of course, only a subset of real spatial sound scenes and of the associated

challenges for SELD systems. Most of these considerations were addressed in the recent dataset for the new DCASE2020 challenge [55]. A significant advance is the introduction of reverberant moving sources, still based on captured RIRs from real spaces [55], [56]. Moreover, ambient noise occurs at varying levels, reverberant conditions are stronger and more varied, and event locations do not occur in a sparse regular grid but can vary more or less continuously. Hence, after DCASE2019 confirmed that informed engineering can solve the SELD task successfully under the restricted conditions of its dataset, the DCASE2020 challenge focuses on presenting more challenging evaluation conditions closer to reality.

It should be pointed out that a rigorous analysis of the metrics and of their accuracy, consistency, and behaviour in general is still open and remains a topic for future work. It should be also emphasized that until such work has been done, the proposed results should be seen as empirical; they are based on observation and on correlating the proposed performance measurements with the expected behaviour of the measured systems. Similar empirical results were obtained in a second setting in [44], where a learning-based SELD system was measured while it was trained, with all scores increasing in accordance with its optimization objective.

VI. Conclusions and future work

This work presented and analyzed the submissions of the DCASE2019 SELD challenge, with a discussion on general and individual characteristics of the systems, how those reflected on their performance, and a comprehensive evaluation. This first challenge revealed a strong community focused on the joint localization and detection, coming both from the audio machine learning and the array signal processing fields. Compared to the few related studies before the challenge, the participants demonstrated strong advances in terms of SELD modeling, engineering, and in terms of raw performance. The majority of submissions surpassed the baseline with a large margin and the best ones reached almost perfect localization and detection scores.

Taking into account the advances in the recent DCASE2020 SELD challenge, we can envision some of the challenges in a SELD task that have not been addressed yet. In terms of the spatial properties of the scene, two points not addressed yet are moving receivers (together with moving sources) and directional interferences which represent clearly localized sounds of unknown types. Both of these properties are expected to be introduced in the upcoming challenges, after DCASE2020. Beyond spatial characteristics, an evolution of the challenge and its datasets would consider the overall spatiotemporal scene consistency. At the moment events are randomly chosen and spatialized. A realistic scene generator should spatialize events that fit a given space at their most probable locations, while respecting real-life co-occurence probabilities. Such consistency between space, sound source locations, respective sound emitting actions, and the sound events associated with all the above remains a topic for future research.

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

[5] 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*, London, UK, 2007, pp. 21–26.

[6] T. May, S. Van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, 2012.

[7] 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)*, Barcelona, Spain, 2011, pp. 1317–1321.

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

[9] 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 and Applications*, vol. 75, no. 17, pp. 10 407–10 439, 2016.

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

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

[12] 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.

[13] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[14] 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)*, Rome, Italy, 2018, pp. 1462–1466.

[15] 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.

[16] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[17] 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, pp. 10–14.

[18] 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, pp. 119–123.

[19] 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, pp. 30–34.

[20] W. Xue, T. Ying, Z. Chao, and D. Guohong, "Multi-beam and multi-task learning for joint sound event detection and localization," DCASE2019 Challenge, Tech. Rep., 2019.
[21] J. Zhang, W. Ding, and L. He, "Data augmentation and prior knowledge-based regularization for sound event localization and detection," DCASE2019 Challenge, Tech. Rep., 2019.
[22] 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, pp. 199–203.
[23] T. N. T. Nguyen, D. L. Jones, R. Ranjan, S. Jayabalan, and W. S. Gan, "DCASE 2019 Task 3: A two-step system for sound event localization and detection," DCASE2019 Challenge, Tech. Rep., 2019.
[24] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using FOA domain spatial augmentation," DCASE2019 Challenge, Tech. Rep., 2019.
[25] K. Noh, C. Jeong-Hwan, J. Dongyeop, and C. Joon-Hyuk, "Three-stage approach for sound event localization and detection," DCASE2019 Challenge, Tech. Rep., 2019.
[26] R. Ranjan, S. Jayabalan, T. N. T. Nguyen, and W. S. Gan, "Sound event detection and direction of arrival estimation using residual net and recurrent neural networks," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019, pp. 214–218.
[27] S. Park, W. Lim, S. Suh, and Y. Jeong, "Trellisnet-based architecture for sound event localization and detection with reassembly learning," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019, pp. 179–183.
[28] S. Leung and Y. Ren, "Spectrum combination and convolutional recurrent neural networks for joint localization and detection of sound events," DCASE2019 Challenge, Tech. Rep., 2019.
[29] 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, pp. 84–88.
[30] Z. Lu, "Sound event detection and localization based on CNN and LSTM," DCASE2019 Challenge, Tech. Rep., 2019.
[31] P. LiHong, Z. Xue, C. Ping, W. Zhe, and Z. Chun, "Polyphonic sound event detection and localization using a two-stage strategy," DCASE2019 Challenge, Tech. Rep., 2019.
[32] E. L. Tan, R. Ranjan, and S. Jayabalan, "Sound event detection and localization using ResNet RNN and time-delay DOA," DCASE2019 Challenge, Tech. Rep., 2019.
[33] H. Cordourier-Maruri, P. Lopez Meyer, J. Huang, J. Del Hoyo Ontiveros, and H. Lu, "GCC-PHAT cross-correlation audio features for simultaneous sound event localization and detection (SELD) on multiple rooms," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019, pp. 55–58.
[34] D. Krause and K. Kowalczyk, "Arborescent neural network architectures for sound event detection and localization," DCASE2019 Challenge, Tech. Rep., 2019.
[35] 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, pp. 189–193.
[36] S. P. Chytas and G. Potamianos, "Hierarchical detection of sound events and their localization using convolutional neural networks with adaptive thresholds," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019, pp. 50–54.
[37] E. Nustede and J. Anemuller, "Group delay features for sound event detection and localization (Task 3) of the DCASE 2019 challenge," DCASE2019 Challenge, Tech. Rep., 2019.
[38] Q. Kong, Y. Cao, T. Iqbal, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems," DCASE2019 Challenge, Tech. Rep., 2019.
[39] Y. Lin and Z. Wang, "A report on sound event localization and detection," DCASE2019 Challenge, Tech. Rep., 2019.
[40] I. Trowitzsch, C. Schymura, D. Kolossa, and K. Obermayer, "Joining sound event detection and localization through spatial segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 487–502, 2019.
[41] L. Pi, X. Zheng, C. Zhang, P. Chen, Z. Wang, and X. Li, "U recurrent neural network for polyphonic sound event detection and localization," in *5th International Conference on Multimedia Systems and Signal Processing*, Chengdu, China, 2020, pp. 86–91.
[42] W. Wang, F. Seraj, N. Meratnia, and P. J. Havinga, "Localization and classification of overlapping sound events based on spectrogram-keypoint using acoustic-sensor-network data," in *IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, Bali, Indonesia, 2019, pp. 49–55.
[43] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 71–75.
[44] 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)*, New Paltz, NY, USA, 2019.
[45] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.
[46] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[47] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE transactions on signal processing*, vol. 56, no. 8, pp. 3447–3457, 2008.
[48] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.
[49] H. Abdi and L. Williams, "Jackknife," in *Encyclopedia of research design*, N. J. Salkind, Ed. Thousand Oaks, CA, USA: Sage, 2010, vol. 1, pp. 655–665.
[50] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
[51] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech*, Graz, Austria, 2019, pp. 2613–2617.
[52] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.
[53] 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, pp. 154–158.
[54] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.
[55] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," DCASE2020 Challenge, Tech. Rep., 2020.
[56] 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, pp. 20–24.

Archontis Politis is a post-doctoral researcher at Tampere University, Finland. He obtained his M.Eng. degree in civil engineering from Aristotle University, Thessaloniki, Greece, and his M.Sc. degree in Sound & Vibration studies from the Institute of Sound and Vibration Research (ISVR), Southampton University, UK, in 2006 and 2008, respectively. In 2016 he obtained a Doctor of Science degree on parametric spatial sound recording and reproduction from Aalto University, Finland. His research interests include spatial audio technologies, virtual acoustics, array signal processing, and acoustic scene analysis.

Annamaria Mesaros is assistant professor at Tampere University, Finland. She received the M.Sc. and Ph.D degrees in electronics and telecommunications in 2001 and 2007, respectively, from Technical University of Cluj Napoca, Romania, and Doctor of Science degree in signal processing from TUT in 2012. She has also been working as a postdoctoral researcher at Aalto University, Helsinki, Finland, within the Finnish Centre of Excellence in Computational Inference Research. Her research focuses on sound event detection in real-world multisource environments, including semantic aspects of human-generated sound annotation.

Sharath Adavanne received his M.Sc. and Ph.D. degrees in signal processing from Tampere University, Finland in 2011 and 2020 respectively. Between 2011 and 2016 he worked in the industry solving problems related to music information retrieval, speech recognition, audio fingerprinting and general audio content analysis. His current research interest is in the application of machine learning based methods for real-life auditory scene analysis.

Toni Heittola is a doctoral student at Tampere University, Finland. He received his M.Sc. degree in Information Technology from Tampere University of Technology (TUT), Finland, in 2004. He is currently pursuing the Ph.D. degree at Tampere University. His main research interests are sound event detection in real-life environments, sound scene classification and audio content analysis.

Tuomas Virtanen is a professor at Tampere University, Finland. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition, music content analysis and audio event detection. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored about 100 scientific publications on the above topics. He is a member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society.