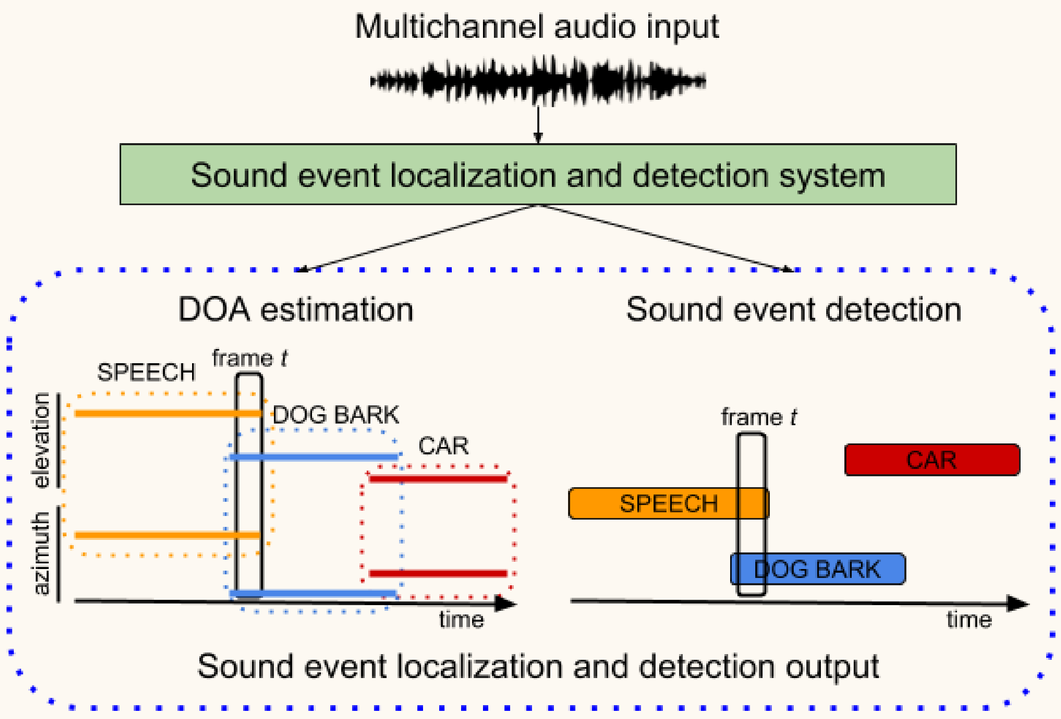
DCASE2021 Challenge

Summary & Results, Task 3

T3 Sound Event Localization and Detection

Coordinators
Archontis Politis, Antoine Deleforge, Sharath Adavanne,
Prerak Srivastava, Daniel Krause and Tuomas Virtanen

Task description



- **Localize and recognize individual sound events and their respective temporal onset and offset times, in the presence of interfering directional events not belonging to the target classes and spatial ambient noise.**
- *Motivation*: Enables an automated description of human activities with a spatial dimension, and help machines to interact with the world more seamlessly.
- *Examples*: Robots can independently recognize and spatially track the sound source of interest.

Dataset

	DCASE2019	DCASE2020	DCASE2021
# rooms	5 rooms	13 rooms	13 rooms
# spatial RIRs/positions	504 discrete positions	~200 spatial trajectories (continuously captured SRIRs)	~200 spatial trajectories (continuously captured SRIRs)
Source-to-receiver distances	1m-2m	1m-5m	1m-5m
Spatial ambient noise	30dB SNR	6-30dB SNR	0-30dB SNR
Moving sources	No	Yes	Yes
Non-target interfering events	No	No	**Yes**
# polyphony/overlapping events	≤2	≤2	**≤3 (+ ≤1 interf. event)**
% same-class overlapping events	low	low	**high**
# target classes	11	14	12
# event samples	220	~700	~500 (target events) ~400 (interferer events)

Figure 1: Comparison of SELD datasets created for DCASE Challenges.

- measured spatial room impulse responses from multiple rooms
- recorded spatial ambient noise from the same rooms
- increased occurrence of same-class overlapping events
- contains non-target interfering sound events
- two spatial formats derived from a spherical microphone array

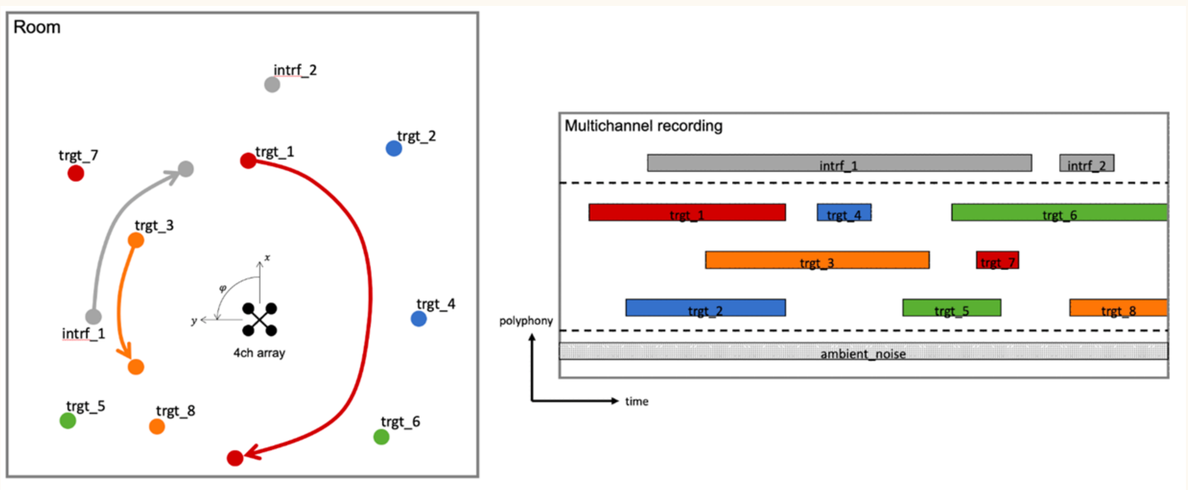


Figure 2: Exemplary depiction of an emulated recording in the dataset.

Submissions

- **Baseline method**: Modified SELDnet.
  - SELDnet is a convolutional and recurrent neural network (CRNN) that jointly performs sound event detection (SED) and direction-of-arrival (DOA) estimation as a multi-output regression task.
  - This year, SELDnet was modified to have activity-coupled Cartesian direction-of-arrival (ACCDOA) output representation [Shimada et.al.]
    - ACCDOA unifies the SED and DOA estimation into a single homogeneous regression output, simplifying the overall architecture, while simultaneously improving its performance.
- **Evaluation metric**: Joint detection and localization metrics computed over one-second non-overlapping segments, similar to the DCASE2020.
  - Location-dependent SED metrics: error rate ER_{20° and F_{20°. Predicted true positive are considered only if they are less than 20° from reference.
  - Classification-dependent localization metrics: Localization error (LE) and Localization recall (LR).
- **Challenge rank**: The submissions are ranked for each of the four metrics individually, and the final rank of each system is determined by the sum of the four individual ranks.

Results

Systems	Format	Method	Features	ER_{20°	F_{20°	LE	LR
Shimada	FOA	RD3Net, EINV2	IPD, cosIPD, sinIPD, magnitude, PCEN spectra	0.32	79.1	8.5	82.8
Nguyen	FOA	CRNN, MHSA	mel spectra, direct-to-reverberant ratio, eigenvectors of spatial-covariance matrix	0.32	78.3	10.0	78.3
Parrish	FOA	CNN, MHSA	mel & constant-Q spectra, intensity vector	0.39	73.8	12.8	76.8
Lee	FOA	Transformer	mel spectra, intensity vector	0.40	72.9	13.2	76.5
Park	BOTH	Transformer	mel spectra, intensity vector	0.46	67.8	12.8	72.3
Zhang	BOTH	Conformer	mel spectra, intensity vector, GCC	0.46	64.7	12.8	61.9
Ko	FOA	Transformer	mel spectra, intensity vector	0.58	60.3	15.1	70.7
Huang	BOTH	Transformer	Waveform	0.57	52.3	18.5	58.5
Yalta	FOA	Transformer	mel spectra, intensity vector	0.72	52.5	20.1	71.1
Naranjo	FOA	CRNN	mel spectra, intensity vector	0.68	37.7	25.3	53.9
Baseline	FOA	CRNN	mel spectra, intensity vector	0.67	37.2	23.9	45.8
Bai	MIC	CRNN	mel spectra, GCC	0.79	16.4	66.5	35.5
Sun	FOA	CRNN	mel spectra, intensity vector	0.95	2.7	84.5	17.4

Discussion

- 31 submissions: 12 Teams, 48 Authors, 17 Affiliations (3 Industry).
- 3 of the 12 teams supported same-class overlapping events detection.
- Popular choices: ACCDOA representation (8 of 12 teams), data augmentation (10/12), self-attention in the form of transformer blocks, conformer blocks, and cross-modal attention (10/12).
- Unlike previous years, none of the teams employed a) model-based or parametric localization estimators or b) separate modelling of SED and DOA estimation tasks.

TAU-NIGENS Spatial Sound Events 2021 dataset
https://doi.org/10.5281/zenodo.5476980

DCASE2021 baseline system
https://github.com/sharathadavanne/seld-dcase2021