DCASE2021 CHALLENGE

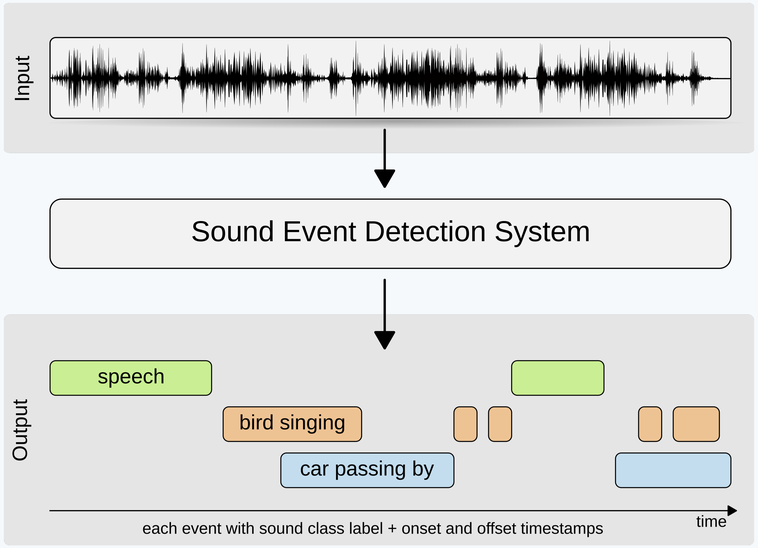
Summary & Results, Task 4

T4 Sound Event Detection and Separation in Domestic Environments

Coordinators
Romain Serizel, Francesca Ronchini, Nicolas Turpault, Scott Wisdom, Hakan Erdogan, John Hershey, Justin Salamon, Prem Seetharaman, Eduardo Fonseca, Samuele Cornell, Daniel P. W. Ellis

Results
https://tinyurl.com/DCASE2021-T4

Task description



- **Detecting and classifying sound events within 10-second audio clips from youtube and vimeo**
- *Motivation*: Smart home applications, assisted living
- *Challenges*: Partly and weakly labeled real training data + synthetic soundscapes (strongly labeled)

DESED Dataset

Novelties since 2020:

- **Non-target events:**
  - Clips from FUSS containing the non-target classes
  - Selection based on FSD50K annotations
- **Event distribution:** computed on annotations obtained by humans for ≈90k clips from Audioset.
- **Additionnal datasets:**
  - Sound events: FSD50K (both target and non-target)
  - Sound sources: YFCC100M (annotations not necessarily consistent with DESED)

Submissions

- 78 Systems
  - 22 Teams
  - 98 Authors

Ranking metric

Polyphonic sound detection score for two different scenarios

- **Scenario 1:** localization of the sound event is really important ($\mathrm{PSDS\_1}$)
- **Scenario 2:** relaxed localization constraint but strong constraint on class confusion ($\mathrm{PSDS\_2}$)

Ranking score:

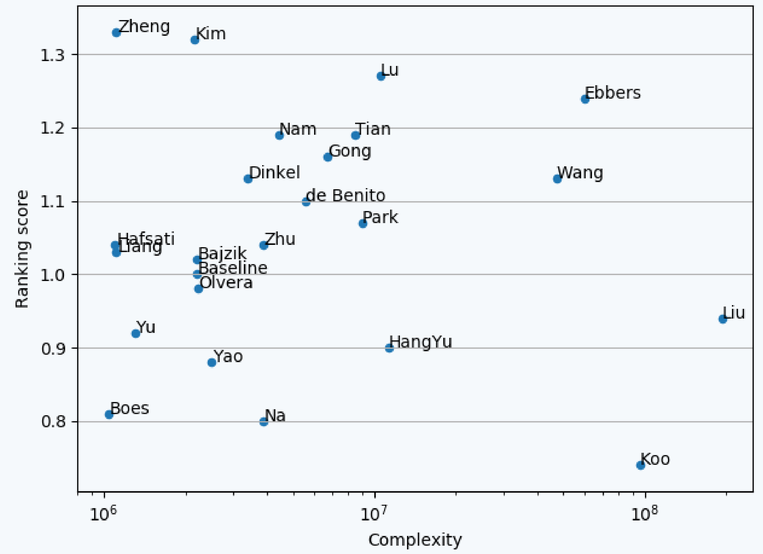
$$\mathrm{PSDS\_1} + \mathrm{PSDS\_2}$$

with $\overline{\mathrm{PSDS\_1,2}}$: the PSDS on scenario 1 and 2 normalized by the baseline PSDS.

Results and systems description, Top 10

	System Id		Scores		
Submission	PSDS1	PSDS2	Ranking	PSDS1	PSDS2
Zheng_USTC	SED_1	SED_3	1.4	0.452	0.746
Kim_AiTeR_GIST	SED_4	SED_4	1.32	0.442	0.674
Nam_KAIST	SED_2	SED_4	1.29	0.399	0.715
lu_kwai_task4	SED_1	SED_3	1.29	0.419	0.686
Ebbers_UPB_task4	SED_3	SED_4	1.24	0.416	0.637
Tian_ICT-TOSHIBA	SED_1	SED_1	1.19	0.413	0.586
Gong_TAL	SED_3	SED_3	1.16	0.37	0.626
Cai_SMALLRICE	SED_2	SED_3	1.14	0.373	0.596
Wang_NSYSU	SED_3	SED_4	1.14	0.339	0.662
Baseline_SSep_SED			1.11	0.364	0.58
deBenito_AUDIAS	SED_2	SED_4	1.1	0.363	0.577
Park_JHU	SED_2	SED_2	1.07	0.327	0.603
Liang_SHNU	SED_2	Ssep_SED_1	1.05	0.325	0.588
Hafsati_TUITO	SED_2	SED_2	1.04	0.336	0.55
Zhu_AIAL-XJU	SED_1	SED_1	1.04	0.318	0.583
Bajzik_UNIZA	SED_2	SED_2	1.02	0.33	0.544
Baseline_SED			1	0.315	0.547

Complexity



Take-away message

- Most of the systems used:
  - C(R)NN
  - Log-mel energies
  - Data augmentation
  - Teacher teacher-student
  - Median filtering
- Self-training is used by a few submissions
- Top performing systems are using ensembles
  - Best performing single system is ranked 11^{th}
- A few systems were specialized to scenarion 1/2
- **Complexity**:
  - Many systems are more complex than the baseline
  - The top performing system is simpler than the baseline
  - Overall complexity did not increase since last year

DETECTION AND CLASSIFICATION OF ACOUSTIC SCENES AND EVENTS