A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes

Francesca Ronchini, Romain Serizel

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Overall goal

- Benchmark of a representative sampling of the state-of-the-art systems in Sound Event Detection task.
- Systems evaluated on synthetic soundscapes according to the polyphonic sound detection score.
- Analysis of robustness of the systems to varying level of target to non-target signal-to-noise ratio and to temporal localization of target sound events.

Problem definition

The task of Sound Event Detection (SED) consists in correctly detecting target sound events present in an audio clip.
SED systems are expected to produce strongly-labeled outputs.

In this study synthetic soundscapes are used as evaluation set in order to obtain a benchmark of Detection and Classification Acoustic Scene and Events 2021 Task 4 submissions (which represent a sample of the state-of-the-art in SED), analyzing:

- robustness of the systems to varying levels of target to non-target signal-to-noise ratio (TNTSNR);
- robustness of the systems to varying time localization of target sound events;
- impact of non-target sound events.

Problem setting

- **Scenario 1**: The system needs to react fast upon an event detection.
- **Scenario 2**: The system must avoid confusion between classes but the reaction time is less crucial than in the first scenario.

The systems selected for the study are reported in Table 1, together with the **PSDS_1** and **PSDS_2** metrics.

Dataset generation

The dataset considered for this study is the **DESED dataset**. It is composed of 10 seconds length audio clips either recorded in a domestic environment or synthesized to reproduce such an environment.

This study aims to investigate challenges related to real SED aspects. In order to so, the following different versions of the synthetic part of the DESED dataset have been generated:

- synthetic set with varying TNTSNR;
- synthetic set with varying onset time;
- sysnthetic set including only non-target events.

Submission code system 1	PSDS_1	PSDS_2	Submission code system 2	PSDS_1	PSDS_2
Zheng_USTC_task4_SED_1	0.45	0.67	Zheng_USTC_task4_SED_3	0.39	0.75
Iu_kwai_task4_SED_1	0.42	0.66	Iu_kwai_task4_SED_3	0.15	0.69
Kim_AiTeR_GIST_SED_4	0.44	0.67	Kim_AiTeR_GIST_SED_4	0.44	0.67
Nam_KAIST_task4_SED_2	0.40	0.61	Nam_KAIST_task4_SED_4	0.06	0.72
Tian_ICT_TOSHIBA_task4_SED_1	0.41	0.59	Tian_ICT_TOSHIBA_task4_SED_1	0.41	0.59
Gong_TAL_task4_SED_3	0.37	0.63	Gong_TAL_task4_SED_3	0.37	0.63
Baseline	0.31	0.55	Baseline	0.31	0.55

Table 1:PSDS_1 and PSDS_2 of six systems selected for the analysis, plus the baseline.

Impact of TNTSNR on Scenario 1

- **Focus:** investigate the impact of the TNTSNR.

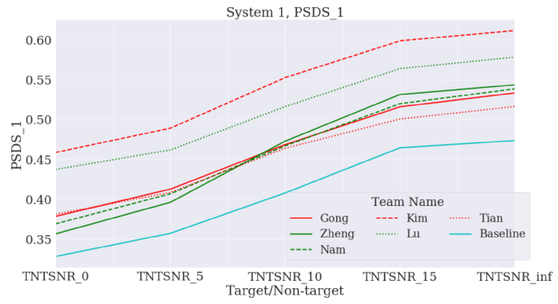

Figure 1:PSDS_1 results for systems selected for scenario 1.

- All the submissions perform better when only target events are present in the evaluation. The performance decreases with the TNTSNR getting lower.
- Probably, the TNTSNR has little effect on the segmentation performance of the systems.

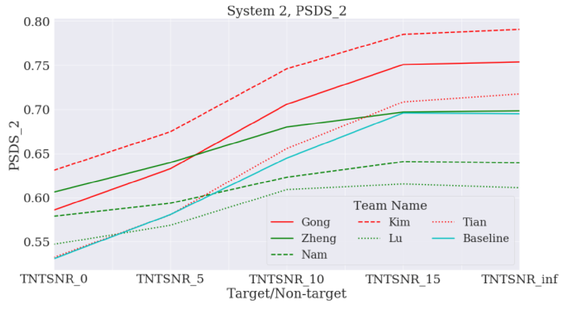

Figure 2:PSDS_2 results for systems selected for scenario 2.

- In Fig 2, the systems tailored to provide coarse segmentation show more robust results with respect to different TNTSNR.

Impact of time localization of the original event

- **Focus:** investigate sound event localization within the clip.

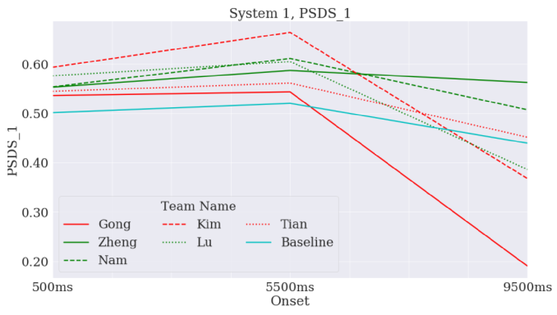

Figure 3:PSDS_1 results for system selected for scenario 1.

- Performance consistently drop with general systems.
- Systems adapted to the scenario show attenuated performance drop.

Impact of non-target sound events

- **Focus:** investigate the impact of non-target sound events.
- **Evaluation:** synthetic set with only non-target events.

Table 2 and 3 present the number of target events detected by the systems on clips that do not contain any target event.

Table 2 shows the systems tailored to have a fine segmentation.

Table 3 shows the systems tailored to have coarse segmentation.

The results on the tables are splitted depending on the average length of the target classes detected.

Submission code	All events	Short events	Long events
Zheng_SED_1	721	665	56
Lu_SED_1	781	719	62
Nam_SED_2	1098	1044	54
Baseline	831	697	134

Table 2:Non-target events detected by fine-segmentation systems evaluated on synth_ntg.

Submission code	All events	Short events	Long events
Zheng_SED_3	448	392	56
Lu_SED_3	282	225	57
Nam_SED_4	500	434	66
Baseline	831	697	134

Table 3:Non-target events detected by coarse-segmentation systems evaluated on synth_ntg.

- Systems tend to predict short events more than long events, especially systems with fine segmentation.
- This sensitivity probably has to be taken into account when designing systems with fine segmentation.

Conclusions

- Systems that are tailored for a fine time segmentation are generally more robust to the event localization within the clips.
- Fine time segmentation systems can also be more sensitive to false alarm triggered by non-target events.
- Systems that are tailored for coarse time segmentation generally provide an event classification that is more robust to low TNTSNR.

Acknowledgments

We would like to thank all the other organizers of DCASE 2022 Challenge Task 4.

References

- Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon, “Sound event detection in synthetic domestic environments,” in ICASSP, 2020.
- Nicolas Turpault, Romain Serizel, Scott Wisdom, Hakan Erdogan, John R Hershey, Eduardo Fonseca, Prem Seetharaman, and Justin Salamon, “Sound event detection and separation: a benchmark on desed synthetic soundscapes,” in ICASSP, 2021