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

Francesca Ronchini, Romain Serizel

A BENCHMARK OF STATE-OF-THE-ART SOUND EVENT DETECTION SYSTEMS EVALUATED ON SYNTHETIC SOUNDSCAPES

Francesca Ronchini[1]*, Romain Serizel*[1]

[1]Université de Lorraine, CNRS, Inria, Loria, Nancy, France

ABSTRACT

This paper proposes a benchmark of submissions to Detection and Classification Acoustic Scene and Events 2021 Challenge (DCASE) Task 4 representing a sampling of the state-of-the-art in Sound Event Detection task. The submissions are evaluated according to the two polyphonic sound detection score scenarios proposed for the DCASE 2021 Challenge Task 4, which allow to make an analysis on whether submissions are designed to perform fine-grained temporal segmentation, coarse-grained temporal segmentation, or have been designed to be polyvalent on the scenarios proposed.

We study the solutions proposed by participants to analyze their robustness to varying level target to non-target signal-to-noise ratio and to temporal localization of target sound events. A last experiment is proposed in order to study the impact of non-target events on systems outputs. Results show that systems adapted to provide coarse segmentation outputs are more robust to different target to non-target signal-to-noise ratio and, with the help of specific data augmentation methods, they are more robust to time localization of the original event. Results of the last experiment display that systems tend to spuriously predict short events when non-target events are present. This is particularly true for systems that are tailored to have a fine segmentation.

Index Terms— Sound event detection, synthetic soundscapes, open-source datasets, deep learning

1. INTRODUCTION

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 (i.e. detect sound events with a start time, end time, and sound class label) [1]. Multiple events can be present in each audio recording, including overlapping target sound events and potentially non-target sound events.

Since 2018, DCASE Challenge Task 4 proposes to address the SED problem in a context where systems are provided with unlabeled and weakly labeled recorded clips (without any timing information) for training [2]. In 2019, we proposed to use an additional training set composed of strongly labeled synthetic soundscapes [3]. These are cheap to obtain but they can introduce a domain mismatch when SED systems have to operate on recorded clips.

This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS Learning to understand audio scenes (ANR-18-CE23-0020), the project CPS4EU Cyber Physical Systems for Europe (Grant Agreement number: 826276) and the French region Grand-Est. Experiments presented in this paper were carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000).

The authors would like to thank all the other organizers of DCASE 2021 Challenge Task 4.

Task 4 of the DCASE Challenge has been motivated by the fact that manually obtaining strong annotations is time consuming and therefore expensive. Additionally, strong annotations are known to be error-prone, mainly because of the subjective judgement of each annotator regarding onsets and offsets localization. A large set of annotations have been recently released on a portion of Audioset [4]. Doing this for each specific application is not feasible and strong labels data remain subject to human annotators interpretations. Therefore, it is still relevant to explore to what extent it is possible to cheaply train a SED system from an heterogeneous dataset [5].

An additional advantage of using synthetically generated soundscapes is the possibility to have full control over the properties of the soundscapes [6]. This allows to generate custom synthetic clips to untangle some of the many problems faced by a SED system operating under real conditions. Since 2019, within DCASE Task 4, synthetic soundscapes have been designed to target some specific SED open problems and they have been proposed in the evaluation set to the challenge participants in order to obtain a benchmark of state-of-the-art SED systems on these specific problems [7, 8].

During previous iterations of the challenge, we have observed that obtaining an accurate time segmentation was one of the most prominent challenges of the SED systems [8, 9, 10]. In 2021, we proposed to change the metric from an event-based F-score [11] to a polyphonic sound detection score (PSDS) [12]. This metric has been proven to be more robust in labeling subjectivity than the collar-based match metric (such as event-based F-Score) and it allows focusing on specific application scenarios [13]. In particular depending on how finely sound events need to be localized in time, a specific set of parameters can be used.

This paper benchmarks DCASE 2021 Challenge Task 4 submissions which represent a sample of the state-of-the-art in SED. We investigate the solutions proposed by participants depending on time segmentation constraints, analyse the robustness of the different submitted systems to varying levels of target to non-target signal-to-noise ratio (TNTSNR) and a varying time localization of target sound events. The SED submissions are evaluated according to the two PSDS [24] scenarios defined for the DCASE 2021 Challenge Task 4. [1].

2. DATASETS AND TASK SETUP

2.1. Task setup and evaluation metrics

In order to better understand the expected behaviour of each submission and with the aim to emphasize different systems properties for the two scenarios considered in the DCASE 2021 Challenge Task 4, the following definitions are given.

[1]To promote reproducibility, the mapping files, the ground-truth of the proposed subsets, and the submissions of the teams are made available under an open-source license `https://zenodo.org/record/5949149`

Ref	Submission code system 1	PSDS_1	PSDS_2	Submission code system 2	PSDS1	PSDS2
[14]	Zheng_USTC_task4_SED_1	0.45	0.67	Zheng_USTC_task4_SED_3	0.39	0.75
[15]	Kim_AiTeR_GIST_SED_4	0.44	0.67	Kim_AiTeR_GIST_SED_4	0.44	0.67
[16]	lu_kwai_task4_SED_1	0.42	0.66	lu_kwai_task4_SED_3	0.15	0.69
[17]	Nam_KAIST_task4_SED_2	0.40	0.61	Nam_KAIST_task4_SED_4	0.06	0.72
[18]	Ebbers_UPB_task4_SED_3	0.42	0.64	Ebbers_UPB_task4_SED_4	0.36	0.64
[19]	Tian_ICT_TOSHIBA_task4_SED_1	0.41	0.59	Tian_ICT_TOSHIBA_task4_SED_1	0.41	0.59
[20]	Gong_TAL_task4_SED_3	0.37	0.63	Gong_TAL_task4_SED_3	0.37	0.63
[21]	Wang_NSYSU_task4_SED_3	0.34	0.65	Wang_NSYSU_task4_SED_4	0.30	0.66
[22]	Cai_SMALLRICE_task4_SED_2	0.37	0.58	Cai_SMALLRICE_task4_SED_3	0.37	0.60
[23]	Baseline	0.31	0.55	Baseline	0.31	0.55

Table 1. PSDS_1 and PSDS_2 of the nine highest-ranked teams based on rankings score plus the baseline.

Parameter	Scenario 1	Scenario 2
DTC	0.7	0.1
GTC	0.7	0.1
α_{CT}	0	0.5
cttc	-	0.3

Table 2. PSDS parameters of each scenario defined.

Scenario 1: The system needs to react fast upon an event detection (e.g. to trigger an alarm, adapt home automation system ...). The localization of the sound event is then really important.

Scenario 2: The system must avoid confusion between classes but the reaction time is less crucial than in the first scenario.

Different PSDS sets of parameters need to be defined in order to reflect the particular needs of each scenario. For the DCASE 2021 Challenge Task 4, four parameters have been customized according to the different needs: Detection Tolerance criterion (DTC), Ground Truth intersection criterion (GTC), Cross-Trigger Tolerance criterion (cttc) and Cost of Cross-Triggers (CT) on the user experience (α_{CT}). Table 2 summarizes the PSDS values given the scenario. PSDS values are computed using 50 operating points (linearly distributed from 0.01 to 0.99). More information regarding PSDS can be found in Bilen et al. [24].

The PSDS will be indicated throughout the paper as **PSDS_1** when evaluated on scenario 1 and **PSDS_2** when evaluated on scenario 2, regardless of the system.

The systems analyzed on this paper have been selected according to the official ranking. The ranking criterion is the aggregation of PSDS_1 and PSDS_2. Each separate metric considered in the final ranking criterion is the best separate metric among all teams submission (each team is allowed 4 different submissions and PSDS_1 and PSDS_2 can be obtained by two different systems of the same team).

$$\text{Ranking Score} = \overline{\text{PSDS\_1}} + \overline{\text{PSDS\_2}} \quad (1)$$

with $\overline{\text{PSDS\_1}}$ and $\overline{\text{PSDS\_2}}$ being the PSDS on scenario 1 and 2 normalized by the baseline PSDS on these scenarios, respectively.

In this paper, we analyze different experiments to understand the systems behavior. The two different scenarios allows us to understand more in detail the possible adaptation of each system. The setup has been chosen in order to favor experiments on the systems behavior and their adaptation to different target scenario.

2.2. Datasets

2.2.1. DESED dataset

The dataset considered on this paper is the DESED dataset[2] [10, 25], which is the same as provided for the DCASE 2021 Challenge Task 4. It is composed of 10 seconds length audio clips either recorded in a domestic environment or synthesized to reproduce such an environment[3]. The synthetic part of the dataset is generated with Scaper [6], a Python library for soundscape synthesis and augmentation. The foreground events (both target and non-target) are obtained from the Freesound Dataset (FSD50k) [26], while the background sounds are obtained from the SINS dataset (activity class "other") [27] and TUT scenes 2016 development dataset [28]. The event co-occurences are computed on a set of strong annotations from Audioset [4]. More information regarding the generation of the DESED dataset can be found in Ronchini et al. [23].

2.3. Synthetic evaluation datasets

The aim of this study is to investigate challenges related to real SED aspects such as target to non-target signal-to-noise ratio (TNTSNR), sound events localization in time and the impact of non-target sound events. Starting from the reference synthetic soundscapes evaluation set, eight different evaluation sets have been designed to specifically target these challenges. These additional evaluation datasets have been proposed to DCASE 2021 Task 4 participants together with the official evaluation set in order to be able to benchmark state-of-the-art systems on these particular aspects.

2.3.1. Reference synthetic soundscapes evaluation set

The synthetic 2021 evaluation set is composed of 1000 audio clips. In the context of the challenge, this subset is used for analysis purposes. We will refer to it as **synth**, being the reference evaluation synthetic set. For consistency, on the figures on the paper the set **synth** will be referred to as **TNTSNR_0**, being 0 the TNTSNR. It contains target and non-target events distributed between the different audio clips according to pre-calculated co-occurrences [23]. Two additional versions of the **synth** set have been generated, **TNTSNR_inf** (only target sound events) and **synth_ntg** (only non-target sound events).

[2] https://project.inria.fr/desed/

[3] For a detailed description of the DESED dataset and how it is generated the reader is referred to the original DESED article [25] and DCASE 2021 task 4 webpage: `http://dcase.community/challenge2021`

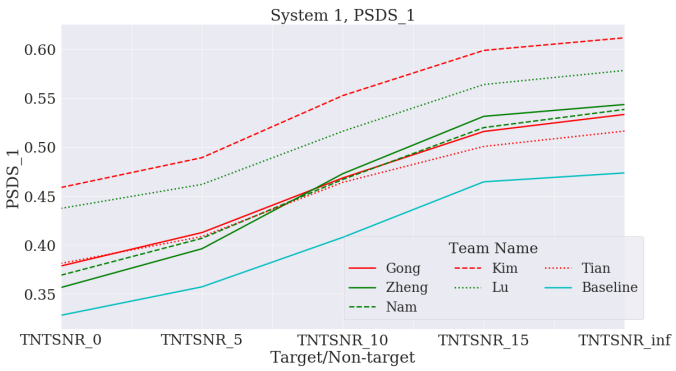

Fig. 1. PSDS_1 results for systems selected for scenario 1, evaluated with reference evaluation set and sets with varying TNTSNR.

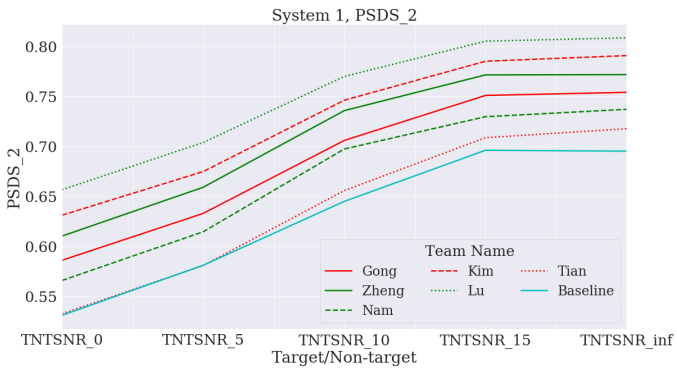

Fig. 2. PSDS_2 results for systems selected for scenario 1, evaluated with reference evaluation set and sets with varying TNTSNR.

2.3.2. Synthetic set with varying TNTSNR

With the aim of studying what would be the impact of varying the TNTSNR on the system performance, three different versions of **synth** have been generated. The SNR of non-target events have been decreased by 5 dB, 10 dB and 15 dB compared to their original value. The original SNR of the sound events is randomly selected between 6 dB and 30 dB, so the more we decrease the SNR, the less the sound will be audible, with some of the events that will not be audible at all. These subsets will be subsequently referred to as **TNTSNR_5**, **TNTSNR_10**, **TNTSNR_15**.

2.3.3. Synthetic set with varying onset time

A subset of 1000 soundscapes is generated with a uniform sound event onset distribution and only one event per soundscape. Three variants of this subset have been generated with the event onset shifted in time. The sound event onsets are located between 250 ms and 750 ms in the first version, between 5.25 s and 5.75 s in the second version and between 9.25 s and 9.75 s in the third version. These subsets will be hereafter referred to as **500ms**, **5500ms** and **9500ms**, respectively. These subsets are designed to study the sensibility of the SED segmentation to the sound event localization in time. A similar experiment has already been conducted for DCASE 2019 and DCASE 2020 Task 4 [7, 8]. The main purpose here is to analyze whether the systems have improved on this particular aspect.

3. IMPACT OF THE EVALUATION SCENARIO

One novelty of the DCASE Challenge 2021 Task 4 is the introduction of two evaluation scenarios. As described in Section 2.1, one scenario is strict in terms of temporal segmentation (scenario 1) while the other is more permissive in terms of temporal segmentation but systems must avoid confusion between classes (scenario 2). Since the participants were allowed to present up to 4 different systems, we first analyze whether or not the participants submitted specific systems for the scenarios.

Table 1 presents the PSDS scores obtained on the official evaluation set for both scenarios. In particular, the table reports the results of the 9 highest-ranked teams and the baseline. For each team, the PSDS for the two systems selected for the ranking score are reported. For three participants, the systems that perform best on scenario 1 and scenario 2 have different behaviors [14, 16, 17]. Zheng et al. adjust the sigmoid temperature parameter to obtain soft or sharp detection output depending on the scenario [14]. Lu et al. use a convolutional nerual network (CNN) model on scenario 1 and a conformer on scenario 2 [16]. Nam et al. propose to train a weak SED system for scenario 2, primarily focusing on event classification [17]. For some participants, the best system on scenario 1 is also the best system on scenario 2 [15, 19, 20]. In the remainder of the paper we focus on the six submissions listed above, comparing their performance to those obtained with the baseline. Table 1 also reports performance obtained by other participants for which the difference between the best system on scenario 1 and the best system on scenario 2 is marginal (the difference is mainly due to a different system adjustment) [18, 21, 22].

4. ANALYSIS ON THE SYNTHETIC DATASETS

This section compares the different submissions evaluated on the different synthetic datasets described in Section 2 in order to investigate the impact of the TNTSNR (Section 4.1), the sound event localization within the clip (Section 4.2) and the impact of non-target sound events when considered alone (Section 4.3).

The objective here is to highlight the different impact of using systems that are fine-tuned according to the different scenarios compared to using a general system. Therefore, all the plots in this paper use different colors to identify the teams. The performance for the teams using different systems for the two scenarios [14, 16, 17] are represented with green lines, while the performance for the teams using the same system for both scenarios [15, 19, 20] are presented in red. The cyan line represents the baseline performance.

4.1. Impact of TNTSNR

Figures 1, 2 and 3 report the results obtained with the submissions evaluated on the dataset with varying TNTSNR (see also Sections 2.3.1 and 2.3.2). Figures 1 and 2 show the performance of the systems selected on scenario 1, reporting both PSDS_1 and PSDS_2 whereas Figure 3 reports the PSDS_2 for the submissions selected on scenario 2.

As can be observed from Figures 1 and 2, all the submissions (selected on scenario 1 emphasizing the temporal segmentation) perform better when only target events are present in the evaluation set, with the performance that consistently decreases with the TNTSNR getting lower. This confirms the results observed in previous work [29]. The drop between the best performance (without

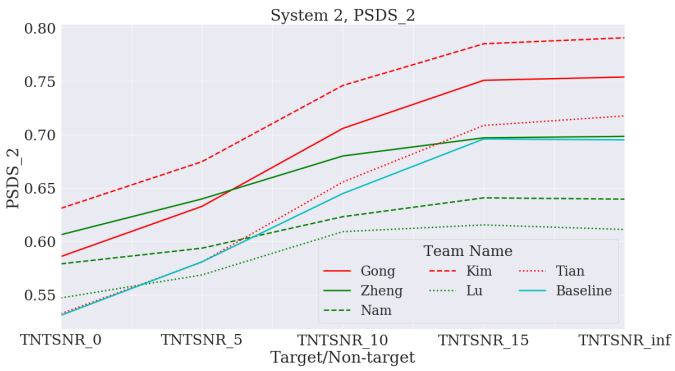

Fig. 3. PSDS_2 results for systems selected for scenario 2, evaluated with reference evaluation set and sets with varying TNTSNR.

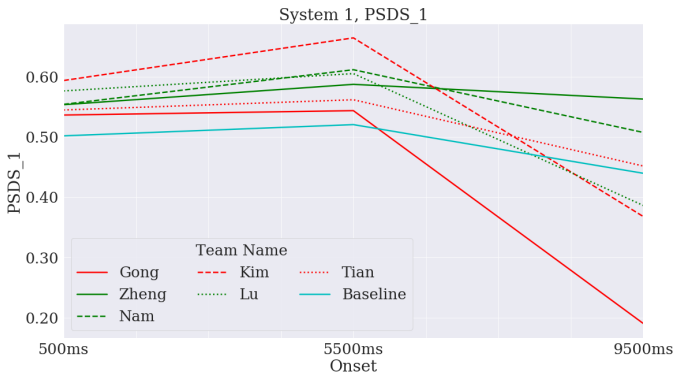

Fig. 4. PSDS_1 results for systems selected for scenario 1, evaluated with synthetic set with varying onset time.

non-target sound events) and the worst performance (TNTSNR_0) is similar with both metrics (Figure 1 and 2) regardless of the systems. This could indicate that the TNTSNR has little effect on the segmentation performance of the systems.

Figure 3 shows the PSDS_2 of the submissions selected for scenario 2 (relaxed time segmentation constraint). It is interesting to notice that the results obtained with systems that are tailored to provide coarse segmentation are more robust to different TNTSNR. In fact, with respect to this scenario, which emphasizes sound event class prediction over time segmentation, they evince a relatively smaller drop between the best performance (without non-target sound events) and the worst performance (TNTSNR_0) compared to the general systems. This tends to indicate that strong sound event classification capabilities are particularly important when considering robustness to varying TNTSNR.

4.2. Impact of time localization of the original event

Figure 4 shows the PSDS_1 for the submissions selected on scenario 1. In this experiment, we focus on systems selected for scenario 1 as the problem considered in this part of the analysis is related to time localization. We evaluate the systems using the evaluation sets described in Section 2.3.3. This experiment has already been proposed for DCASE 2019 and DCASE 2020 submissions [7, 8]. Previous analyses showed that performance consistently drop when the onsets of the sound events are located towards the end of the clips [8]. Similar performance trends are obtained with general systems [15, 19, 20], while systems that have been adapted for scenario 1 [14, 16, 17] generally show attenuated performance drop towards the end of the clips. In particular, the approaches proposed by Zheng et al. [14] and Nam et al. [17] have robust performance regardless of the onset timing. According to the systems description, this improvement seems to be related to the use of specific data augmentation methods.

4.3. Impact of non-target sound events

In this last experiment, we investigate the impact of non-target sound events on the systems outputs. Similarly to the study reported on Ronchini et al. [23], we are interested in studying whether false positive events can be triggered by non-target events and identify which ones. In order to do so, the systems have been evaluated with the **synth_ntg** evaluation set, described in Section 2.3.1.

Table 3 presents the number of target events detected by the systems on clips that do not contain any target event. The results are split depending on the average length of the target classes detected: *Alarm bell ringing, Cat, Dishes, Dog and Speech* are considered as short event classes while *Blender, Running water, Electric shaver toothbrush, Frying and Vacuum cleaner* are considered as long event classes. Systems tend to spuriously predict short events more than long events. This is particularly true for systems that are tailored to have a fine segmentation (Zheng_SED_1, Lu_SED_1, Nam_SED_2). This sensitivity probably has to be taken into account when designing systems with fine segmentation.

Submission code	All events	Short events	Long events
Zheng_SED_1	721	665	56
Zheng_SED_3	448	392	56
Lu_SED_1	781	719	62
Lu_SED_3	282	225	57
Nam_SED_2	1098	1044	54
Nam_SED_4	500	434	66
Baseline	831	697	134

Table 3. Number of non-target events detected by the systems when evaluated with **sytnh_ntg**.

5. CONCLUSIONS

This paper presents a benchmark of state-of-the-art Sound Event Detection systems (submissions to DCASE 2021 Challenge Task 4) on evaluation sets composed of synthetic soundscapes designed to target specific challenges of the Sound Event Detection task. The main challenges addressed in this paper are the impact of varying target to non-target signal-to-noise ratio, the impact of time localization of the sound event and the impact of non-target sound events. We observe that systems that are tailored for a fine time segmentation are generally more robust to the event localization within the clips but can also be more sensitive to false alarm triggered by non-target events. On the other end, systems that are tailored for coarse time segmentation generally provide an event classification that is more robust to low TNTSNR.

6. REFERENCES

[1] Tuomas Virtanen, Mark D Plumbley, and Dan Ellis, *Computational analysis of sound scenes and events*, Springer, 2018.

[2] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," in *Proc. DCASE Workshop*, United Kingdom, 2018.

[3] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *DCASE Workshop*, United States, 2019.

[4] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP*, 2021.

[5] Nicolas Turpault and Romain Serizel, "Training Sound Event Detection On A Heterogeneous Dataset," in *DCASE Workshop*, Japan, 2020.

[6] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *WASPAA*, United States, 2017.

[7] Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP*, Spain, 2020.

[8] 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.

[9] Romain Serizel and Nicolas Turpault, "Sound Event Detection from Partially Annotated Data: Trends and Challenges," in *IcETRAN conference*, Serbia, 2019.

[10] Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP*, Spain, 2020.

[11] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, 2016.

[12] Çağdaş Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP*, Spain, 2020.

[13] Giacomo Ferroni, Nicolas Turpault, Juan Azcarreta, Francesco Tuveri, Romain Serizel, Çagdaş Bilen, and Sacha Krstulović, "Improving Sound Event Detection Metrics: Insights from DCASE 2020," in *DCASE Workshop*, Japan, 2020.

[14] Xu Zheng, Han Chen, and Yan Song, "Zheng ustc team's submission for dcase2021 task4 – semi-supervised sound event detection," Tech. Rep., DCASE2021 Challenge, 2021.

[15] Nam Kyun Kim and Hong Kook Kim, "Self-training with noisy student model and semi-supervised loss function for dcase 2021 challenge task 4," Tech. Rep., DCASE2021 Challenge, 2021.

[16] Rui Lu, Wenzheng Hu, Duan Zhiyao, and Ji Liu, "Integrating advantages of recurrent and transformer structures for sound event detection in multiple scenarios," Tech. Rep., DCASE2021 Challenge, 2021.

[17] Hyeonuk Nam, Byeong-Yun Ko, Gyeong-Tae Lee, Seong-Hu Kim, Won-Ho Jung, Sang-Min Choi, and Yong-Hwa Park, "Heavily augmented sound event detection utilizing weak predictions," Tech. Rep., DCASE2021 Challenge, 2021.

[18] Reinhold Ebbers, Janek Haeb-Umbach, "Self-trained audio tagging and sound event detection in domestic environments," Tech. Rep., DCASE2021 Challenge, 2021.

[19] Gangyi Tian, Yuxin Huang, Zhirong Ye, Shuo Ma, Xiangdong Wang, Hong Liu, Yueliang Qian, Rui Tao, Long Yan, Kazushige Ouchi, and Reinhold Ebbers, Janek Haeb-Umbach, "Sound event detection using metric learning and focal loss for dcase 2021 task 4," Tech. Rep., DCASE2021 Challenge, 2021.

[20] Yaguang Gong, Changlong Li, Xintian Wang, Lu Ma, Song Yang, and Zhongqin Wu Wu, "Improved pseudo-labeling method for semi-supervised sound event detection," Tech. Rep., DCASE2021 Challenge, 2021.

[21] Yih-Wen Wang, Chia-Ping Chen, Chung-Li Lu, and Bo-Cheng Chan, "Cht+nsysu sound event detection system with multi-scale channel attention and multiple consistency training for dcase 2021 task 4," Tech. Rep., DCASE2021 Challenge, 2021.

[22] Dongchi Yu, Xichang Cai, Duxin Liu, and Zihan Liu, "Semi-supervised sound event detection using multi-scale convolutional recurrent neural network and weighted pooling," Tech. Rep., DCASE2021 Challenge, 2021.

[23] Francesca Ronchini, Romain Serizel, Nicolas Turpault, and Samuele Cornell, "The impact of non-target events in synthetic soundscapes for sound event detection," *DCASE Workshop*, 2021.

[24] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP*, Spain, 2020.

[25] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *DCASE Workshop*, United States, 2019.

[26] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "Fsd50k: an open dataset of human-labeled sound events," *arXiv preprint arXiv:2010.00475*, 2020.

[27] Gert Dekkers, Steven Lauwereins, Bart Thoen, Mulu Weldegebreal Adhana, Henk Brouckxon, Bertold Van den Bergh, Toon Van Waterschoot, Bart Vanrumste, Marian Verhelst, and Peter Karsmakers, "The sins database for detection of daily activities in a home environment using an acoustic sensor network," in *DCASE Workshop*, Germany, 2017.

[28] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Tut database for acoustic scene classification and sound event detection," in *EUSIPCO*, Hungary, 2016.

[29] Nicolas Turpault, Scott Wisdom, Hakan Erdogan, John R Hershey, Romain Serizel, Eduardo Fonseca, Prem Seetharaman, and Justin Salamon, "Improving Sound Event Detection In Domestic Environments Using Sound Separation," in *DCASE Workshop*, Japan, 2020.