THE IMPACT OF NON-TARGET EVENTS IN SYNTHETIC SOUNDSCAPES FOR SOUND EVENT DETECTION

Francesca Ronchini[1]*, Romain Serizel*[1]*, Nicolas Turpault*[1]*, Samuele Cornell*[2]

[1]Université de Lorraine, CNRS, Inria, Loria, Nancy, France
[2] Department of Information Engineering, Universita Politecnica delle Marche, Italy

ABSTRACT

Detection and Classification Acoustic Scene and Events Challenge 2021 Task 4 uses a heterogeneous dataset that includes both recorded and synthetic soundscapes. Until recently only target sound events were considered when synthesizing the soundscapes. However, recorded soundscapes often contain a substantial amount of non-target events that may affect the performance. In this paper, we focus on the impact of these non-target events in the synthetic soundscapes. Firstly, we investigate to what extent using non-target events alternatively during the training or validation phase (or none of them) helps the system to correctly detect target events. Secondly, we analyze to what extend adjusting the signal-to-noise ratio between target and non-target events at training improves the sound event detection performance. The results show that using both target and non-target events for only one of the phases (validation or training) helps the system to properly detect sound events, outperforming the baseline (which uses non-target events in both phases). The paper also reports the results of a preliminary study on evaluating the system on clips that contain only non-target events. This opens questions for future work on non-target subset and acoustic similarity between target and non-target events which might confuse the system.

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

1. INTRODUCTION

The main goal of ambient sound and scene analysis is to automatically extract information from sounds that surround us and analyze them for different purposes and applications. Between the different area of interest, ambient sound analysis have a considerable impact on applications such as noise monitoring in smart cities [1, 2], domestic applications such as smart homes and home security solutions [3, 4], health monitoring systems [5], multimedia information retrieval [6] and bioacoustics domain [7]. Sound Event Detection (SED) aims to identify the onset and offset of the sound events present in a soundscape and to correctly classify them, labeling the events according to the target sound classes that they belong to. Nowadays, deep learning is the main method used to approach the problem. However, one of the main limitations of deep learning models is the requirement of large amounts of labeled training data to reach good performance. The process of labeling data is time-consuming and bias-prone mainly due to human errors and disagreement given the subjectivity in the perception of some sound event onsets and offsets [8]. To overcome these limitations, recent works are investigating alternatives to train deep neural networks with a small amount of labeled data together with a bigger set of unlabeled data [3, 9, 10, 8, 11]. Among them, Detection and Classification Acoustic Scenes and Events Challenge (DCASE) 2021 Task 4 uses an heterogeneous dataset that includes both recorded and synthetic soundscapes [8]. This latter soundscapes provide a cheap way to obtain strongly labeled data. Until recently, synthesized soundscapes were generated considering only target sound events. However, recorded soundscapes also contain a considerable amount of non-target events that might influence the performance of the system.

The purpose of this paper is to focus on the impact on the system's performance when non-target events are included in the synthetic soundscapes of the training dataset. The study has been mainly divided into three stages. Firstly, we investigate to what extent using non-target events alternatively during training or validation helps the system to correctly detect the target sound events. Mainly motivated from the results of the first experiment, in the second part of the study, we focus on understanding to what extend adjusting the target to non-target signal-to-noise ratio (TNTSNR) at training improves the sound event detection performance. Results regarding a preliminary study on the evaluation of the system using clips containing only non-target events are also reported, opening questions for future studies on possible acoustic similarity between target and non-target sound events which might confuse the SED system. [1].

2. PROBLEM DEFINITION AND DATASET GENERATION

2.1. Problem definition

The primary goal of the DCASE 2021 Challenge Task 4 is the development of a semi-supervised system for SED, exploiting an heterogeneous and unbalanced training dataset. The goal of the system is to correctly classify the sound event classes and to localize the different target sound events present in an audio clip in terms of timing. Each audio recording can contain more than one event. Some of those could also be overlapped. The use of a larger

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).

[1]To promote reproducibility, the code, `https://github.com/DCASE-REPO/DESED_task`, and pre-trained models `https://zenodo.org/record/5529692`, are made available under an open-source license.

amount of unlabeled recorded clips is motivated by the limitations related to annotating a SED dataset (human-error-prone and time-consuming). Alternatively, synthesized soundscapes are an easy way to have strongly annotated data. In fact, the user can easily generate the soundscapes starting from isolated sound events. On the other hand, in most of the recorded soundscapes the target sound classes are almost never present alone. For this reason, one of the main novelties of the DCASE 2021 Challenge Task 4 is the introduction of non-target isolated events in the synthetic soundscapes[2]. This paper explores the impact of the non-target sound events on the baseline system performance, with the final goal of understanding and highlighting how to correctly exploit them to generate realistic soundscapes.

2.2. Dataset generation

The dataset used in this paper is the DESED dataset[3] [12, 13], which is the same 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[4]. The synthetic part of the dataset is generated with Scaper [14], a Python library for soundscape synthesis and augmentation, which allows to control audio parameters. The recorded soundscapes are taken from AudioSet [15]. The foreground events (both target and non-target) are obtained from the Freesound Dataset (FSD50k) [16], while the background sounds are obtained from the SINS dataset (activity class "other") [17] and TUT scenes 2016 development dataset [18]. In particular, non-target events are the intersection of FUSS dataset [19] and FSD50k dataset in order to have compatibilty with the source separation baseline system.

In this article, we modify only the synthetic subset of the dataset. Starting from the synthetic part of the DESED dataset, we generated different versions of it in order to investigate how non-target events impact the system performance and to what extent their relationship with the target events affects the training phase of the system. The following subsections describe the different subsets used for the experiments, which have been generated using Scaper.

2.2.1. Synthetic training set

The synthetic training set is the same set of data released for the DCASE 2021 Challenge Task 4. It includes 10000 audio clips where both target and non-target sound events could be present in each clip. The distribution of the sound events among the files have been determined considering the co-occurrences between the different sound events. The co-occurrences have been calculate considering the strong annotations released for the AudioSet dataset [20][5]. A second version of this dataset has been generated where only target events are present. The datasets will be hereafter referred as **synth_tg_ntg** (used by the official baseline system) and **synth_tg** for the synthetic subset including target and non-target events and the synthetic subset including only target events, respectively.

2.2.2. Synthetic validation set

The synthetic validation set is the same as the synthetic validation dataset supplied for the DCASE 2021 Challenge Task 4. It includes 3000 audio clips including target and non-target events, which distribution has been defined calculating the co-occurrences between sound events. We generated a second version of the dataset containing only target events. The datasets will be referred to as **synth_tg_ntg_val** (used by the baseline system) and **synth_tg_val** (only target sound events).

2.2.3. Synthetic evaluation set

The synthetic 2021 evaluation set is composed by 1000 audio clips. In the context of the challenge, this subset is used for analysis purposes. We will refer to it as **synth_tg_ntg_eval**. It contains target and non-target events distributed between the different audio clips according to the pre-calculated co-occurrences. Two different versions of the **synth_tg_ntg_eval** set have been generated, **synth_tg_eval** (only target sound events) and **synth_ntg_eval** (only non-target sound events).

2.2.4. Varying TNTSNR training and validation set

With the aim of studying what would be the impact of varying the TNTSNR on the system performance, different versions of **synth_tg_ntg** and **synth_tg_ntg_val** have been generated. In particular, for each of them, three versions have been created. The SNR of the 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 **synth_5dB**, **synth_10dB**, **synth_15dB** for the training subsets and **synth_5dB_val**, **synth_10dB_val**, **synth_15dB_val** for the validation subsets.

2.2.5. Public evaluation set

The public evaluation set is composed of recorded audio clips extracted from Youtube videos that are under creative common licenses. This is part of the evaluation dataset released for the evaluation phase of the DCASE 2021 Challenge Task 4 and considered for ranking. The set will be referred to as **public**.

3. EXPERIMENTS TASK SETUP

In order to compare the results with the official baseline, we used the same SED mean-teacher system released for this year challenge. More information regarding the system can be found at Turpault et al. [8] and on the official webpage of the DCASE Challenge Task 4. All the different models have been trained 5 times. This paper reports the average of the scores and the confidence intervals related to those. Only for the baseline model we do no report the confidence intervals because we have considered the results using the checkpoint made available for it [6]. The metrics considered for the study are the two polyphonic sound detection score (PSDS) [21] scenarios defined for the DCASE 2021 Challenge Task 4, since these are the official metrics used in the challenge.

[2] http://dcase.community/challenge2021

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

[4] For a detailed description of the DESED dataset and how it is generated the reader is referred to the original DESED article [13] and DCASE 2021 task 4 webpage: http://dcase.community/challenge2021

[5] The co-occurrences distribution and the code used to compute them will be distributed.

[6] https://zenodo.org/record/4639817

Non-target		PSDS1	PSDS2
Train	Val		
✓		33.81 (0.36)	52.62 (0.19)
	✓	35.92 (0.49)	54.85 (0.29)
		34.90 (0.82)	53.07 (1.22)
✓	✓	**36.40**	**58.00**

Table 1: Evaluation results for the **public** set, considering the different combinations of using target and non-target sound events at training and validation.

The scope of these experiments is twofold: understand the impact of non-target events on the system performance and investigate to what extend the TNTSNR helps the network to correctly predict the sound events in both matched and mismatched conditions. In order to do so, we divided the experiment into three stages. The first part of the study is focused on understanding the influence of training the system with non-target events. This experiment is described and discussed in Section 4. Section 5 reports the results and the relative discussion of the second part of the experiment where we investigate if a mismatch in terms of TNTSNR between datasets could have an impact on the output of the system. Section 6 reports preliminary results of the last stage of the experiment, regarding the evaluation of the system on the **synth_ntg_eval** dataset, formed by only non-target sound events, in order to investigate if some classes could get acoustically confused at training, having a negative impact on the performance. The last stage has been motivated by the results of the second part of the experiment.

4. USING TARGET/NON-TARGET AT TRAINING

In the first experiment we concentrate on training the system with different combinations of the training dataset. Table 1 reports the results of the experiment evaluating the system on the **public** set. We check-marked the columns NT Train or/and NT Val according to if the non-target sound events are present or not in the synthetic sounscapes. From the results it is possible to observe that using non-target sound events during training and validation improves the performance by a large margin with relaxed segmentation constraints (PSDS2) but only marginally with strict segmentation constraints (PSDS1). In this latter case what matters the most is the use of non-target sound events during the validation. A possible explanation is that synthetic soundscapes with non-target sound events are actually too difficult and confuse the systems when used during the training but they still help reducing the mismatch with recorded soundscapes during model selection (validation).

Table 2 reports the results considering the **synth_tg_ntg_eval** and **synth_tg_eval** evaluation sets. In all cases the best performance is obtained in matched training/evaluation conditions. The performance obtained on **synth_tg_ntg_eval** are lower than the performance obtained on **synth_tg_eval** even in matched conditions. Not surprisingly, this confirm that including non-target sound events makes the SED task more difficult. Interestingly, as opposed to the previous experiment, the most important here is to have matched conditions during training and to a lesser extent during validation. In order to verify the low impact of non-target sound events at training when evaluating on recorded soundscapes, in the next experiment we investigate a possible mismatch in terms in TNTSNR.

Non-target		Eval set	PSDS1	PSDS2
Train	Val			
✓		synth_tg_ntg_eval	23.22 (1.33)	36.44 (2.62)
	✓	synth_tg_ntg_eval	20.08 (0.39)	31.33 (1.29)
		synth_tg_ntg_eval	20.13 (0.35)	30.99 (1.07)
✓	✓	synth_tg_ntg_eval	**25.14**	**40.12**
✓		synth_tg_eval	42.82 (2.42)	58.26 (2.08)
	✓	synth_tg_eval	46.92 (1.02)	**62.79 (0.55)**
		synth_tg_eval	**47.73 (0.33)**	62.54 (1.00)
✓	✓	synth_tg_eval	43.22	61.09

Table 2: Evaluation results for the **synth_tg_ntg_eval** set and **synth_tg_eval** set, considering the different combination of using target and non-target sound events at training and validation.

Non-target		PSDS1	PSDS2
Train	Val		
Original	5 dB	35.57 (0.28)	56.68 (1.77)
5 dB	Original	36.25 (1.26)	57.53 (1.06)
5 dB	5 dB	35.46 (0.46)	**58.09** (0.74)
Original	Original	**36.40**	58.00

Table 3: Evaluation results for the second part of the experiment, varying TNTSNR by 5 dB (**synth_5dB** and **synth_5dB_val**). Evaluating with **public** set.

Non-target		PSDS1	PSDS2
Train	Val		
Original	10 db	36.23 (1.11)	57.82 (1.37)
10 db	Original	**36.42 (0.77)**	**58.94 (0.89)**
10 db	10 db	36.20 (1.14)	57.92 (1.04)
Original	Original	36.40	58.00

Table 4: Evaluation results for the second part of the experiment, varying TNTSNR by 10 dB (**synth_10dB** and **synth_10dB_val**). Evaluating with **public** set.

Non-target		PSDS1	PSDS2
Train	Val		
Original	15 dB	36.08 (1.13)	57.78 (1.33)
15 dB	Original	**37.37 (0.70)**	**58.64 (1.34)**
15 dB	15 dB	36.10 (0.50)	57.36 (0.89)
Original	Original	36.40	58.00

Table 5: Evaluation results for the second part of the experiment, varying TNTSNR by 15 dB (**synth_15dB** and **synth_15dB_val**). Evaluating with **public** set.

5. VARYING TNTSNR AT TRAINING

The second part of the study focuses on understanding the impact of varying the TNTSNR at training and validation aiming at finding a TNTSNR condition that could match better the recorded soundscapes. For each TNTSNR, we use similar combinations as the ones used in Section 4, replacing the set without non-target sound events by a set with adjusted TNTSNR. For example, considering the 5 dB

Validation set	PSDS1	PSDS2
synth_5dB_val	38.68 (1.07)	60.57 (0.78)
synth_10dB_val	**39.07 (0.75)**	**60.75 (0.80)**
synth_15dB_val	37.95 (0.53)	59.99 (1.14)

Table 6: Evaluation results of the SED system, training with **synth_tg**, validating with varying TNTNSNR set and evaluating with **public** set.

case, the combinations considered would be:

- training using the **synth_tg_ntg** set and validating with **synth_5dB_val**;
- training with **synth_5dB** and validating with **synth_tg_ntg_val**;
- training and validating with **synth_5dB** and **synth_5dB_val**.

The fourth combination is the official DCASE Task 4 baseline. Repeating the experiment with all the varying TNTSNR, allow us to analyse to what extend the loudness of the non-target events helps matching the evaluation conditions on recorded clips. Table 3, 4 and 5 report the performance on the **public** set when using a TNTSNR of 5 dB, 10 dB and 15 dB, respectively. When the TNTSNR is 5 dB or 10 dB, the performance changes only marginally between configurations. Increasing the TNTSNR to 15 dB leads to a behaviour more similar to the one obtained in Table 1. The best performance is obtained when training with TNTSNR is 15 dB and validating on **synth_tg_ntg_val**. This could be explained by the fact TNTSNR 15 dB is a condition closer to that of the recorded soundscapes and the fact that it allows for selecting models that will be more robust towards non-target events at test time.

In the last experiment, we investigate the impact of varying the TNTSNR during validation phase, while using the **synt_tg** for training. Results are reported on Table 6, where it is possible to observe that all of them overcome the baseline or are comparable with it, with the best performance obtained for 10 dB TNTSNR. These experiments could indicate that recorded soundscapes in **public** in general have a TNTSNR of about 10 – 15 dB which should be confirmed by complementary experiments.

6. EVALUATING ON NON-TARGET EVENTS ONLY

Based on the previous experiments, TNTSNR could be one reason of mismatch between the synthetic soundscapes and the recorded soundscapes. But this could not explain all the performance differences observed here. In particular why in general having lower TNTSNR during training is decreasing the performance regardless of the validation. One possibility is that the system gets acoustically confused by a possible similarity in sound between events when soundscapes tend to be less dominated by target events. So we evaluated the system using the **synth_ntg_eval**, where only non-target events are considered, to see for which classes the system would output false positives. We evaluated the system on the **public** set; considering the systems trained for the first experiment (see Table 1). Results show that some sound events are detected more than others. For some classes as Speech, this could be explained by the original event distribution (indicated in the first column) but for some other classes as Dishes there is a discrepancy between the original distribution and the amount of false alarms. Interestingly the amount of false alarms is decreased sensibly for most of the classes when including non-target sound events during training.

Classes	Nref	Nsys A	Nsys B	Nsys C	Nsys Base
Dog	197	135	126	146	79
Vacuum_cleaner	127	31	42	44	47
Alarm_bell	191	47	50	52	59
Running_water	116	34	41	61	30
Dishes	405	1478	395	1270	305
Blender	100	63	32	55	19
Frying	156	70	41	60	33
Speech	1686	206	181	180	201
Cat	141	99	103	98	73
Electric_shaver	103	21	18	18	7

Table 7: Preliminary evaluation results by classes, evaluating the system with **synth_ntg_eval**. Nsys (A): training with **synth_tg**, validating with **synth_tg_val**; Nsys (B): training with **synth_tg_ntg**, validating with **synth_tg_val**; Nsys (C): training with **synth_tg**, validating with **synth_tg_ntg_val**; Base: baseline using target and non-target events for training and validation.

7. CONCLUSIONS AND FUTURE WORK

This paper analyzes the impact of including non-target sound events in the synthetic soundscapes of the training dataset for SED systems trained on heterogeneous dataset. In particular, the experiments are divided into three stages: in the first part, we explore to what extend using non-target sound events at training has an impact on the system's performance, secondly we investigate the impact of varying TNTSNR and we conclude the study by analyzing a possible confusion of the SED model in case of false alarms triggered by non-target sound events.

From the results reported on this paper, we can conclude that using non-target sound events can help the SED system to better detect the target sound events, but it is not clear to what extend and what would be the best way to generate the soundscapes. Results show that the final SED performance could depend on mismatches between synthetic and recorded soundscapes, part of which could be due to the TNTSNR but not only. Results on the last experiment show that using non-target events at training decreases the amount of false alarms at test but from this experiment it is not possible to conclude on the impact of non-target sound events on the confusion between the target sound events. This is a first track for future investigation on the topic. Additionally, the impact of the non-target sound events at training on the ability of the system to better segment the target sound events in noisy soundscapes would have to be investigated. A final open question is the impact of the per class distribution of the sound events (both target and non-target) and their co-occurrence distribution on the SED performance.

8. ACKNOWLEDGEMENTS

We would like to thank all the other organizers of DCASE 2021 Challenge Task 4. In particular, we thank Eduardo Fonseca and Daniel P. W. Ellis for their help with the strong labels of the AudioSet dataset used to compute the events co-occurrences, and Justin Salamon and Prem Seetharaman for their help with Scaper.

9. REFERENCES

[1] J. P. Bello, C. Silva, O. Nov, R. L. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for the monitoring, analysis and mitigation of urban noise pollution," *arXiv preprint arXiv:1805.00889*, 2018.

[2] J. P. Bello, C. Mydlarz, and J. Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 373–397.

[3] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[4] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[5] Y. Zigel, D. Litvak, and I. Gannot, "A method for automatic fall detection of elderly people using floor vibrations and sound—proof of concept on human mimicking doll falls," *IEEE transactions on biomedical engineering*, vol. 56, no. 12, pp. 2858–2867, 2009.

[6] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[7] V. Morfi, R. F. Lachlan, and D. Stowell, "Deep perceptual embeddings for unlabelled animal sound events," *The Journal of the Acoustical Society of America*, vol. 150, no. 1, pp. 2–11, 2021.

[8] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[9] R. Serizel and N. Turpault, "Sound event detection from partially annotated data: Trends and challenges," in *IcETRAN conference*, 2019.

[10] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *arXiv preprint arXiv:1804.09288*, 2018.

[11] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[12] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020. [Online]. Available: https://hal.inria.fr/hal-02355573

[13] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, Oct. 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[14] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[15] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[16] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *arXiv preprint arXiv:2010.00475*, 2020.

[17] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. Van den Bergh, T. Van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The sins database for detection of daily activities in a home environment using an acoustic sensor network," *Detection and Classification of Acoustic Scenes and Events 2017*, pp. 1–5, 2017.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[19] S. Wisdom, H. Erdogan, D. P. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 186–190.

[20] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[21] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 61–65.