Sound event detection in synthetic domestic environments

Romain Serizel, Nicolas Turpault, Ankit Shah, Justin Salamon

SOUND EVENT DETECTION IN SYNTHETIC DOMESTIC ENVIRONMENTS

Romain Serizel[1]*, Nicolas Turpault*[1]*, Ankit Shah*[2]*, Justin Salamon*[3]

[1]Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
[2]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, United States
[3]Adobe Research, San Francisco, CA, United States

ABSTRACT

We present a comparative analysis of the performance of state-of-the-art sound event detection systems. In particular, we study the robustness of the systems to noise and signal degradation, which is known to impact model generalization. Our analysis is based on the results of task 4 of the DCASE 2019 challenge, where submitted systems were evaluated on, in addition to real-world recordings, a series of synthetic soundscapes that allow us to carefully control for different soundscape characteristics. Our results show that while overall systems exhibit significant improvements compared to previous work, they still suffer from biases that could prevent them from generalizing to real-world scenarios.

Index Terms— Sound event detection, synthetic data, weakly labeled data, semi-supervised learning

1. INTRODUCTION

We are constantly surrounded by sounds and we rely heavily on these sounds to obtain important information about what is happening around us [1]. Ambient sound analysis aims at automatically extracting information from these sounds. It encompasses disciplines such as sound scene classification (in which context does this happen?) or sound event detection and classification (SED) (what happens during this recording?). This area of research has been attracting a continuously growing attention during the past years as it can have a great impact in many applications in noise monitoring in smart cities [2, 3], surveillance [4], urban planning [2], multimedia information retrieval [5, 6]; and domestic applications such as smart homes, health monitoring systems and home security solutions [7, 8, 9].

In Task 4 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge [10], an extension of the same task from the previous year [7], we proposed to investigate the added value of synthetic soundscapes with strong labels when training a system to perform SED (with time boundaries) in domestic environments. That is, a system had to detect the presence of a sound event as well as predict the onset and offset times of each occurrence of the event. We generated strongly annotated synthetic soundscapes using the Scaper library [11].

The ranking of the task was performed on real audio clips extracted from YouTube and Vimeo. However, the performance analysis and the study of the systems behavior under scenarios when using real-world data collected from the internet is limited, because we have no control over the characteristics of the soundscapes. Therefore, we provided an additional evaluation set composed of synthetic soundscapes. This dataset was inspired by our analysis of the performance of systems submitted to DCASE 2018 task 4 [12]. Using synthetic soundscapes allowed us to design specific scenarios to test the robustness of the submitted systems to corrupted signals, in terms discrimination between long sound events and short sound events, or in terms on sound event segmentation (localizing a sound event in time regardless of its class). This latter point in particular is crucial for SED from weakly labeled data. However, it remains under-investigated.

The synthetic soundscapes evaluation dataset included multiple subsets comprised of the same set of synthetic soundscapes, each with a different type of data degradation applied via the audio degradation toolbox [13] or with for different foreground-to-background signal-to-noise ratio (FBSNR), that is the ratio between the loudness of the sound event (foreground) and the loudness of the background noise. We also considered other subsets where the same sound event would be localized at different time instants within a sound clip and subsets using the long sound event classes as background noise. All these scenarios are realistic and represent conditions where the submitted systems are likely to fail. However, gathering enough real data that would cover each of these aspects is hardly feasible. Not to mention that in some cases (e.g., varying FBSNR) it would require recording new data to ensure that only the tested parameter is changing between experiments. Synthetic soundscapes then offer a flexible and realistic alternative to performance preliminary tests before considering further investigation on real data if need be.

In this paper, we present the results obtained by the participants of DCASE 2019 task 4 on the evaluation composed of synthetic soundscapes. We propose an analysis of the robustness of the submitted approaches to degradation of the recording quality and to varying FBSNR and a study of the robustness of the segmentation process implemented in the submitted systems.

2. DESED DATASET AND TASK SETUP

The DESED dataset is composed of 10-sec audio clips recorded in a domestic environment [7, 10] or synthesized to simulate a domestic environment. In this paper, we focus on the synthetic subset of the DESED dataset.

2.1. DESED synthetic soundscapes evaluation set

The DESED synthetic soundscapes evaluation set is comprised of 10-second audio clips generated with Scaper [11], a Python library

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) 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).

for soundscape synthesis and augmentation. Scaper operates by taking a set of foreground sounds and a set of background sounds and automatically sequencing them into random soundscapes sampled from a user-specified distribution controlling the number and type of sound events, their duration, signal-to-noise ratio, and several other key characteristics. This set is used for analysis purposes and its design is motivated by the analysis of DCASE 2018 task 4 results [12]. In particular, most submissions from DCASE 2018 task 4 performed poorly in terms of event segmentation, that is they were not able to localize sound events properly in time (regardless of the sound event classes).

The foreground events are obtained from the Freesound Dataset (FSD) [14, 15]. Each sound event clip was verified by a human to ensure that the sound quality and the event-to-background ratio were sufficient to be used as an isolated sound event. We also controlled for whether the sound event onset and offset were present in the clip. Each selected clip was then segmented when needed to remove silences before and after the sound event and between sound events when the file contained multiple occurrences of the sound event class. The number of unique isolated sound events per class used to generate the subset of synthetic soundscapes is presented in Turpault et al. [10].

Background sounds are extracted from YouTube videos under a Creative Commons license and from the Freesound subset of the MUSAN dataset [16]. These recordings were selected because they contain a low amount of sound events from our 10 target foreground sound event classes. However, there is no guarantee that these sound event classes are completely absent from the background clips.

DESED synthetic soundscapes evaluation set is further divided into several subsets (described below) for a total of 12,139 audio clips synthesized from 314 isolated events. The synthetic soundscapes are annotated with strong labels that are automatically generated by Scaper [11][1].

2.1.1. Varying foreground-to-background SNR

A subset of 754 soundscapes is generated with Scaper scripts are designed such that the distribution of sound events per class, the number of sound events per clip (depending on the class) and the sound event class co-occurrence are similar to that of the validation set which is composed of real recordings. The foreground event signal-to-noise ratio (SNR) parameter was uniformly drawn between 6 dB and 30 dB. Four versions of this subset are generated varying the value of the background SNR parameter:

- 0 dB (the FBSNR is between 6 dB and 30 dB);
- 6 dB (the FBSNR is between 0 dB and 24 dB);
- 15 dB (the FBSNR is between -9 dB and 15 dB);
- 30 dB (the FBSNR is between -24 dB and 0 dB).

In the remainder of the paper, these subsets will be referred to as **synth_30dB**, **synth_24dB**, **synth_15dB** and **synth_0dB**, respectively. This subset is designed to study the impact of the SNR on the SED systems performance. Related results are discussed in Section 3.2.

2.1.2. Audio degradation

Six alternative versions of the subset **synth_30dB** are generated introducing artificial degradation with the Audio Degradation Toolbox [13]. The signal degradations are generated to simulate degradation faced in real environments. The following degradations are used (with default parameters) : "smartPhonePlayback", "smartPhoneRecording", "unit_applyClippingAlternative", "unit_applyDynamicRangeCompression", "unit_applyLowpassFilter" and "unit_applyHighpassFilter". In the remainder of the paper, these subsets will be referred to as **phone_play**, **phone_record**, **clipping**, **compression**, **lowpass** and **highpass**, respectively. This subset is designed to study the robustness of the SED to audio degradation. Related results are discussed in Section 3.1.

2.1.3. Varying onset time

A subset of 750 soundscapes is generated with uniform sound event onset distribution and only one event per soundscape. The parameters are set such the FBSNR is between 6 dB and 24 dB. Three variants of this subset are generated with the same isolated events, only shifted in time. In the first version, all sound events have an onset located between 250 ms and 750 ms, in the second version the sound event onsets are located between 4.75 s and 5.25 s and in the last version the sound event onsets are located between 9.25 s and 9.75 s. In the remainder of the paper, these subsets will be referred to as **500ms**, **5500ms** and **9500ms**, respectively. This subset is designed to study of the sensibility of the SED segmentation to the sound event location in time. In particular, we wanted to control if SED systems were learning a bias in term of time localization depending on the event length (e.g., long sound events would most often start at the beginning of the sound clip). Related results are discussed in Section 4.

2.1.4. Long sound events vs. short sound events

A subset with 522 soundscapes is generated where the background is selected from one of the five long sound event classes (Blender, Electric shaver/toothbrush, Frying, Running water and Vacuum cleaner). The foreground sound events are selected from the five short sound event classes (Alarm/bell/ringing, Cat, Dishes, Dog and Speech). Three variants of this subset are generated with the same sound event scripts and varying values of the background SNR parameter. In a first subset the resulting FBSNR is 0 dB, the FBSNR is 15 dB in the second and 30 dB in the last subset. In the remainder of the paper, these subsets will be referred to as **ls_0dB**, **ls_15dB** and **30_dB**, respectively. This subset is designed to study of the impact of a sound event being in the background or the foreground on SED performance [17]. Related results are discussed in Section 3.2.

2.2. Evaluation metrics

Submissions were evaluated with event-based measures for which the system output is compared to the reference labels event by event [18]. The correspondence between sound event boundaries are estimated with a 200 ms tolerance collar on onsets and a tolerance collar on offsets that is the maximum of 200 ms and 20 % of the duration of the sound event. When sound event classes are taken into account, the overall F1-score is the unweighted average of the class-wise F1-scores (macro-average). The metrics are computed using the sed_eval library [18].

3. ROBUSTNESS TO NOISE AND DEGRADATIONS

In this section, we focus on the impact of signal degradation on the SED and the FBSNR performance. Each participant was allowed to submit up to four different systems. Only the F1-score for the top

[1]There is a plan to release the dataset within the coming weeks

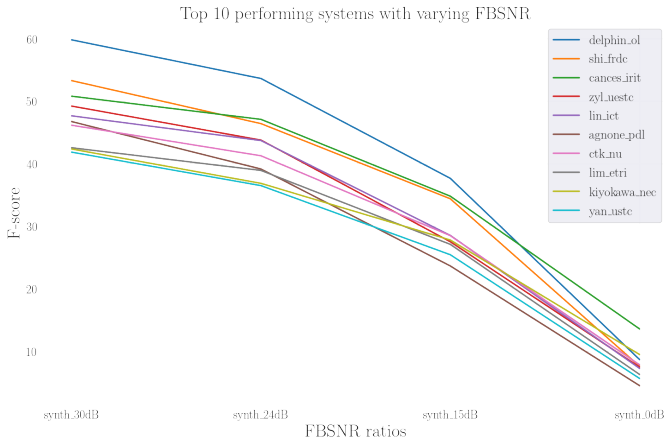


Fig. 1: SED performance depending on the FBSNR.

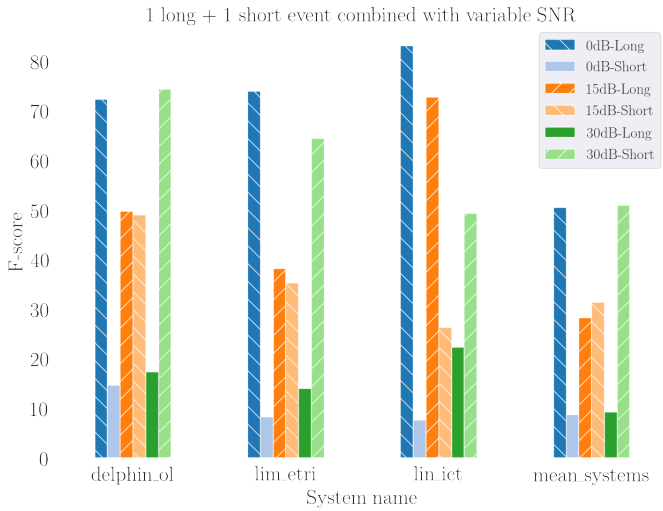


Fig. 2: SED performance depending on the FBSNR when the soundscape is composed of a long event and a short event.

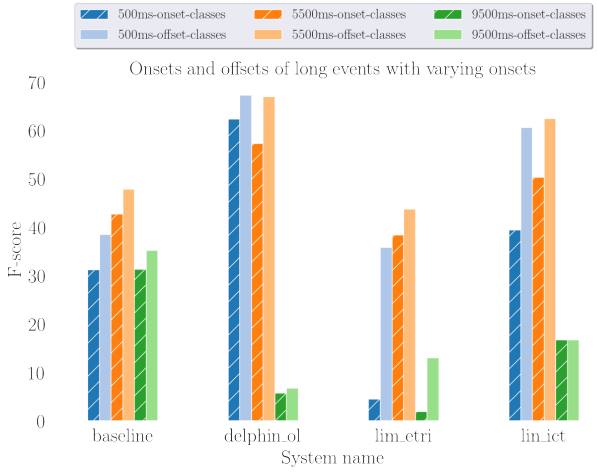


Fig. 3: Segmentation performance for the long sound event classes depending on the event localization in time.

performing system (on **synth_24dB**) for each participant is presented here. We limit the analysis to the 10 top-performing systems.

3.1. Simulated degradations

The F-1 score obtained on the degraded subsets is presented in Table 1. The performance on **synth_24dB** subset are presented here for comparison purpose. The system are ordered alphabetically.

Some systems seem to have over-fitted the synthetic soundscapes subset of the training set and as a result, their performance decreased for most of the degradations. Otherwise, the trend is similar for most of the systems. The submitted systems seem to be rather robust to smartphone related degradations and compression which can be related to the fact that they have been trained on audio data extracted for YouTube and that has most probably been recorded with smartphones. On the other hand, all systems seem to be very sensitive to low-pass and high-pass filtering which tends to indicate that systems are not robust to changes in the frequency range of the input representation between training and test.

3.2. Foreground-to-background Signal-to-noise ratio

In Figure 1, we present the F1-score performance for the 10 top-performing systems mentioned above under varying FBSNR (see Section 2.1.1). The trend for all systems is similar, so no submission really stands out in terms of robustness to noise. Interestingly, on **synth_15dB** where FBSNR should be distributed almost evenly around 0 dB, F1-score performance are still acceptable for most systems and remain in the range of what was obtained on real recording clips [10]. Unsurprisingly, on **synth_0dB**, the FBSNR is always negative and the performance for all systems collapses.

We then propose to analyze the systems' performance when the background is actually one event from the long sound event classes and the foreground sound events are selected in the short sound event classes (see Section 2.1.4). In Figure 2, we present the F1-score performance for the 3 top performing systems (on this particular task) together with the performance averaged over all systems.

In all cases, when the FBSNR is low, all systems consistently obtain better performance on long sound event classes. Whereas when the FBSNR is high, all systems obtain better performance on short sound event classes. When the FBSNR is 0 dB most of the systems perform similarly on short sound event classes and long sound event classes. This tends to show that the bias toward long event classes observed in DCASE 2018 [7, 12] is less important this year. The trend is confirmed by the performance on short or long sound event classes that are within the same range in the most favorable cases (0 dB FBSNR for the long sound event classes, 30 dB for the short sound event classes). However, the system submitted by Lin et al. [19] does not follow this trend and mostly performs better on long events. This could be due to the guided learning methods that biases the mean teacher model. The teacher labels are converted to 0/1 predictions instead of probabilities which may increase the number of with a positive labels and introduce a bias towards long events.

4. SEGMENTATION

In this section, we focus on the analysis of the systems' performance in term of segmentation, that is, the ability of the submitted system to localize a sound event in time (regardless of the sound event class). Sound event segmentation then relies on finding the time instant for

System	Event-based F1-score						
	synth_24dB	**phone_play**	**phone_record**	**clipping**	**compression**	**highpass**	**lowpass**
Agnone, PDL	39.1%	15.4%	9.2%	14.6%	29.6%	8.5%	0.9%
Cances, IRIT	47.1%	25.7%	35.8%	42.6%	44.3%	19.2%	1.2%
Chan, NU	41.2%	25.9%	17.5%	22.8%	33.4%	19.3%	1.2%
Delphin-Poulat, OL	**53.6%**	32.9%	23.7%	29.5%	48.2%	**23.3%**	**4.8%**
Kiyokawa, NEC	36.8%	33.9%	21.9%	35.6%	40.2%	22.1%	4.2%
Lim, ETRI	38.9%	26.9%	30.3%	39.7%	48.1%	15.4%	0.7%
Lin, ICT	43.7%	22.4%	9.3%	19.8%	35.3%	17.6%	0.5%
Shi, FRDC	46.4%	**35.0%**	**36.4%**	**48.3%**	**54.1%**	17.4%	4.0%
Yan, USTC	36.5%	22.1%	21.5%	18.3%	32.7%	16.6%	1.0%
Zhang, UESTC	43.7%	21.8%	15.3%	24.6%	41.4%	14.1%	1.7%
Average score (all participants)	33.9%	22.0 %	16.4%	21.6%	31.4%	15.8%	1.7%

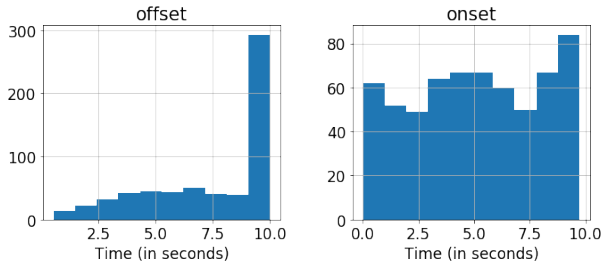
Table 1: F1-score performance on the degraded synthetic soundscapes

both the sound event onset and offset. In particular, we consider the scenario described in Section 2.1.3 in which three versions of a sound clip are generated with the same background and the same sound event starting either at the beginning, in the middle or at the end of the sound clip. The F1-score performance is presented for the 3 top performing systems (on this particular task) together with the performance of the baseline system [10].

For short sound event classes, the F1-score performance is more or less the same wherever the sound event is located within the segment so we will focus on long sound event classes. In Figure 3, we present the F1-score in term of segmentation, onset and offset detection for long sound event classes (but regardless of the sound event class).

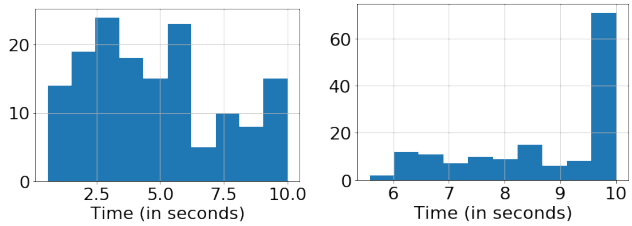
For the long sound event classes, the sound event position within the clip seems to have a large impact as performance dramatically decreases when the sound event is located towards the end of the audio clip. One possible explanation could be that in the training set, long sound events have onsets and offsets are mostly located at the beginning of the audio clips. However, as shown in Figure 4, the onset distribution over time is similar for long and short sound event classes. However, the offsets of long sound event classes are often located toward the end of the sound clip. Therefore, if any bias was introduced by the training it should probably have led to a better offset detection for long sound event classes toward the end of the sound clips. This is confirmed by the fact that all systems are able to detect quite accurately the offsets of long sound event classes when the sound event onset is located toward the middle of the sound clip. However, in this case the sound event offsets are located toward the end of the sound clip in most of the time (see Figure 5).

One alternative explanation is that the submitted systems are simply not able to detect a long sound event class toward the end of the sound clip. For example, median filtering with variable length that are used in most of the submissions (more than 0.5 s for long sound event classes in some cases [19, 20]) would make it unlikely to detect a long sound event class at the end of the sound clip. This hypothesis tends to be confirmed by the fact that the baseline that is using fixed length median filtering as post-processing performs similarly wherever the sound event is located.



(a) Long sound event classes.

Fig. 4: Time distribution of the onsets and the offsets in the synthetic soundscapes subset of DESED training set.



(a) Distribution for the **500ms** subset.(b) Distribution for the **5500ms** subset.

Fig. 5: Time distribution of the offsets for long event classes in the subsets **500ms** and **5500ms** of DESED Evaluation set.

5. CONCLUSION

In this paper, we present an analysis of the performance of the state-of-the-art SED systems. All submissions to DCASE 2019 task 4 were evaluated on a subset composed of synthetic soundscapes. The analysis shows that training SED on sound clips extracted from internet video makes the systems robust towards degradation related to recording and playing sound on a smartphone. Additionally, we emphasize that even though performance has drastically improved since DCASE 2018 task 4, SED systems still rely on biases (in particular for segmentation) that would probably prevent from generalizing to real case conditions. A first step towards solving this problem was taken in DCASE 2019 where the evaluation set included real recording from an unseen source (Vimeo). A solution regarding the segmentation problem would be to design the evaluation set such that it includes the limit cases exhibited in the paper together with providing the isolated sound events to the participants (instead of the soundscapes) such that they could design more diverse training examples that would allow for removing so of the bias introduced during the design and learning phases of their systems.

6. REFERENCES

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

[2] 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," *Communications of the ACM*, In press, 2018.

[3] Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon, "Sound analysis in smart cities," in *Computational Analysis of Sound Scenes and Events*, pp. 373–397. Springer, 2018.

[4] Regunathan Radhakrishnan, Ajay Divakaran, and A Smaragdis, "Audio analysis for surveillance applications," in *Proc. WASPAA*. IEEE, 2005, pp. 158–161.

[5] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[6] Qin Jin, Peter Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze, "Event-based video retrieval using audio," in *Proc. Interspeech*, 2012.

[7] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," Submitted to DCASE2018 Workshop, July 2018.

[8] Christian Debes, Andreas Merentitis, Sergey Sukhanov, Maria Niessen, Nikolaos Frangiadakis, and Alexander Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[9] Yaniv Zigel, Dima Litvak, and Israel Gannot, "A method for automatic fall detection of elderly people using floor vibrations and soundproof of concept on human mimicking doll falls," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 12, pp. 2858–2867, 2009.

[10] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," working paper or preprint, July 2019.

[11] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo 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.

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

[13] Matthias Mauch and Sebastian Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 2013, pp. 83–88.

[14] Frederic Font, Gerard Roma, and Xavier Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 411–412.

[15] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 486–493.

[16] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[17] Justin Salamon and Juan Pablo Bello, "Feature learning with deep scattering for urban sound analysis," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 724–728.

[18] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162, May 2016.

[19] Liwei Lin and Xiangdong Wang, "Guided learning convolution system for dcase 2019 task 4," Tech. Rep., Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, June 2019.

[20] Lionel Delphin-Poulat and Cyril Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," Tech. Rep., Orange Labs Lannion, France, June 2019.