Sound event detection in domestic environments with weakly labeled data and soundscape synthesis

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, Justin Salamon

SOUND EVENT DETECTION IN DOMESTIC ENVIRONMENTS WITH WEAKLY LABELED DATA AND SOUNDSCAPE SYNTHESIS

Nicolas Turpault[1], *Romain Serizel*[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

This paper presents Task 4 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge and provides a first analysis of the challenge results. The task is a follow-up to Task 4 of DCASE 2018, and involves training systems for large-scale detection of sound events using a combination of weakly labeled data, i.e. training labels without time boundaries, and strongly-labeled synthesized data. The paper introduces Domestic Environment Sound Event Detection (DESED) dataset mixing a part of last year dataset and an additional synthetic, strongly labeled, dataset provided this year that we'll describe more in detail. We also report the performance of the submitted systems on the official evaluation (test) and development sets as well as several additional datasets. The best systems from this year outperform last year's winning system by about 10% points in terms of F-measure.

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

1. INTRODUCTION

Sound conveys important information in our everyday lives and we depend on sounds to better understand changes in our physical environment and to perceive events occurring around us. We perceive the sound scene (the overall soundscape of e.g. an airport or inside a house) as well as individual sound events (e.g. car honks, footsteps, speech, etc.). Sound event detection within an audio recording refers to the task of detecting and classifying sound events, that is, temporally locating the occurrences of sound events in the recording and recognising which object or category each sound belongs to. Sound event detection has potential applications in noise monitoring in smart cities [1, 2], surveillance [3], urban planning [1], multimedia information retrieval [4, 5]; and domestic applications such as smart homes, health monitoring systems and home security solutions [6, 7, 8] to name a few. In recent years the field has gained increasing interest from the broader machine learning and audio processing research communities.

Sound event detection (SED) systems trained using weak labels have have seen significant interest [6, 9, 10, 11, 12] in the research community, as they address some of the challenges involved in developing models that require strongly labeled data for training. In particular, strongly labeled data is time-consuming and difficult to annotate as it requires annotating the temporal extent of event occurrences in addition to their presence or absence. Strong label annotations are also more likely to contain human errors/disagreement given the ambiguity in the perception of some sound event onsets and offsets. In the case of weakly labeled data, we only have information about whether an event is present in a recording or not. We have no information about how many times the event occurs nor the temporal locations of the occurrences within the audio clip. For real-world applications it is critical to build systems that generalize over a large number of sound classes and a variety of sound event distributions. In such cases, it may be more feasible to collect large quantities of weakly labeled data as opposed to strongly labeled data which is significantly more costly in time and effort.

We propose to follow up on DCASE 2018 Task 4 [6] and investigate the scenario where large-scale SED systems can exploit the availability of a small set of weakly annotated data, a larger set of unlabeled data and an additional training set of synthetic soundscapes with strong labels. Given these data, the goal of this task is to train SED models that output event detections with time boundaries (i.e. strong predictions) in domestic environments. That is, a system has to detect the presence of a sound event as well as predict the onset and offset times of each occurrence of the event. We generate strongly annotated synthetic soundscapes using the Scaper library [13]. Given a set of user-specified background and foreground sound event recordings, Scaper automatically generates soundscapes containing random mixtures of the provided events sampled from user-defined distributions. These distributions are defined via a sound event specification including properties such as event duration, onset time, signal-to-noise ratio (SNR) with respect to the background and data augmentation (pitch shifting and time stretching). This allows us to generate multiple different soundscape instantiations from the same specification which is set based on our general requirements for the soundscapes. Since generating such strongly labeled synthetic data is feasible on large scale, we provide a strongly labeled synthetic dataset in order to explore if it can help improving SED models. We believe insights learned from this task will be beneficial to the community as such an exploration is novel and will provide a pathway to developing scalable SED systems.

The remainder of this manuscript is organized as follows. Section 2 provides a brief overview of the task definition and how the development and evaluation datasets were created. Section 3 describes the baseline system and the evaluation procedure for Task 4. Section 4 gives an overview of the systems submitted to the challenge for this task. Finally, conclusions from the challenge are provided in section 5.

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).

Class	Unique events	Dev set Clips	Events
Alarm/bell/ringing	190	392	755
Blender	98	436	540
Cat	88	274	547
Dishes	109	444	814
Dog	136	319	516
Electric shaver/toothbrush	56	221	230
Frying	64	130	137
Running water	68	143	157
Speech	128	1272	2132
Vacuum cleaner	74	196	204
Total	1011	2045	6032

Table 1: Class-wise statistics for the synthetic development subset.

Class	Unique events	Synth set 1 Clips	Events
Alarm/bell/ringing	63	101	184
Blender	27	84	95
Cat	26	113	197
Dishes	34	161	293
Dog	43	124	217
Electric shaver/toothbrush	17	113	117
Frying	17	52	52
Running water	20	67	73
Speech	47	471	803
Vacuum cleaner	20	92	93
Total	314	1378	2124

Table 2: Class-wise statistics for the synthetic evaluation subsets

2. TASK DESCRIPTION AND DESED DATASET

2.1. Task description

This task is the follow-up to DCASE 2018 Task 4 [6]. Systems are expected to produce strongly labeled output (i.e. detect sound events with a start time, end time, and sound class label), but are provided with weakly labeled data (i.e. sound recordings with only the presence/absence of a sound included in the labels without any timing information) for training. Multiple events can be present in each audio recording, including overlapping events. As in the previous iteration of this task, the challenge entails exploiting a large amount of unbalanced and unlabeled training data together with a small weakly annotated training set to improve system performance. However, unlike last year, in this iteration of the challenge we also provide an additional training set with strongly annotated synthetic soundscapes. This opens the door to exploring scientific questions around the informativeness of real (but weakly labeled) data versus strongly-labeled synthetic data, whether the two data sources are complementary or not, and how to best leverage these datasets to optimize system performance.

2.2. DESED development dataset

The development (training) part of DESED dataset is composed of 10-sec audio clips recorded in domestic environment or synthesized to simulate a domestic environment. The task focuses on the same 10 classes of sound events used in Task 4 of DCASE 2018 [6]. The DESED dataset is comprised of a subset of real recordings taken from AudioSet [14] and a subset of synthetic soundscapes generated using Scaper. The subset of real recordings is the same as in DCASE 2018 Task 4: the training set remains the same [6] and the validation set is the combination of the validation and evaluation sets from DCASE 2018 Task 4 [10].

2.2.1. Synthetic soundscape generation procedure

The subset of synthetic soundscapes is comprised of 10 second audio clips generated with Scaper [13], a python library for soundscape synthesis and augmentation. Scaper operates by taking a set of foreground sounds and a set of background sounds 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. The foreground events are obtained from the Freesound Dataset (FSD) [15, 16]. 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 if 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 Table 1. It also presents the number of clips containing a class and the number of events per class.

The background textures are obtained from the SINS dataset (activity class "other") [17]. This particular activity class was selected because it contains a low amount of sound events from the 10 target foreground sound event classes. However, there is no guarantee that these sound event classes are completely absent from the background clips. A total of 2060 unique background clips are used to generate the synthetic subset.

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 synthetic soundscapes are annotated with strong labels automatically generated by Scaper [13].

2.3. DESED evaluation dataset

The evaluation dataset is composed of two subsets: a subset with real recording and a subset with synthetic soundscapes.

2.3.1. Real recordings

The first subset is comprised of audio clips extracted from YouTube and Vimeo videos under creative common licenses. This subset contains 1,013 audio clips and is used for ranking purposes.

2.3.2. Synthetic soundscapes

The second subset is comprised of synthetic soundscapes generated with Scaper[1]. This subset is used for analysis purposes and its design is motivated by the analysis of last year's results [10]. In particular, most submission from last year were perform badly in terms

[1]The JAMS [18] annotation files corresponding to these soundscapes will be released on:

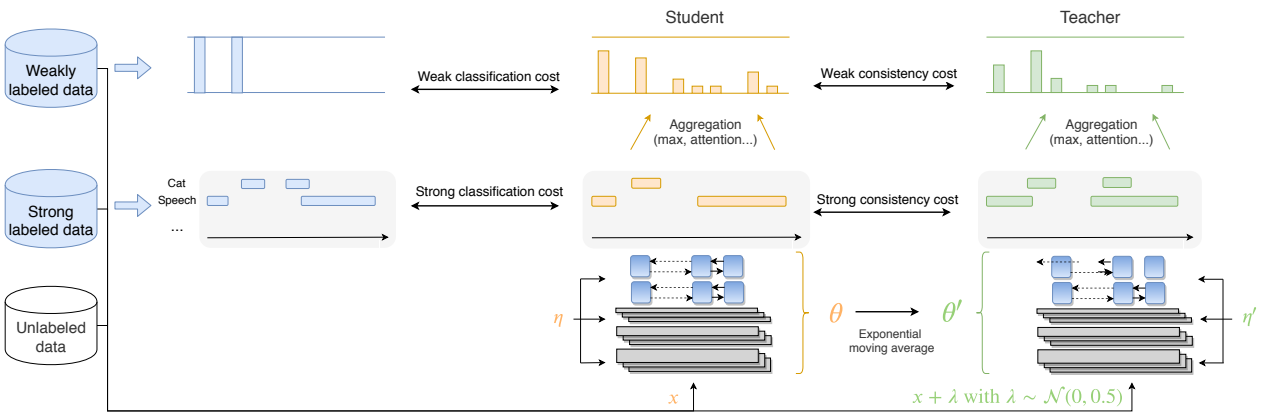

Figure 1: Mean-teacher model. η and η' represent noise applied to the different models (in this case dropout).

of segmentation. One of the goal of this subset is to analyze to which extent strongly labeled data in the training set helped refining the segmentation. The foreground events are obtained from the FSD [15, 16]. The selection process was the same as described for the development dataset. Background sounds are extracted from YouTube videos under a Creative Common license and from the Freesound subset of the MUSAN dataset [19]. The synthetic subset is further divided into several subsets (described below) for a total of 12,139 audio clips synthesized from 314 isolated events. The isolated sound event distribution per class is presented in Table 2.

Varying foreground-to-background SNR: A subset (denoted Synthetic set 1) of 754 soundscapes is generated with a sound event distribution similar to that of the training set. Four versions of this subset are generated varying the value of the foreground events' SNR with respect to the background: 0 dB, 6 dB, 15 dB and 30 dB.

Audio degradation: Six alternative versions of the previous subset (with SNR=0 dB) are generated introducing artificial degradation with the Audio Degradation Toolbox [20]. The following degradations are used (with default parameters) : "smartPhonePlayback", "smartPhoneRecording", "unit_applyClippingAlternative", "unit_applyDynamicRangeCompression", "unit_applyHighpassFilter" and "unit_applyLowpassFilter".

Varying onset time: A subset of 750 soundscapes is generated with uniform sound event onset distribution and only one event per soundscape. The sound event SNR parameter is set to 0 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.

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 similar sound event scripts and varying values of the sound event SNR parameter (0 dB, 15 dB and 30 dB).

3. BASELINE

The baseline system is inspired by the winning system from DCASE 2018 Task 4 by Lu [21][2]. It uses a mean-teacher model which is a combination of two models: a student model and a teacher model (both have the same architecture). Our implementation of the mean-teacher model is based on the work of Tarvainen and Valpola [22]. The student model is the final model used at inference time, while the teacher model is aimed at helping the student model during training and its weights are an exponential moving average of the student model's weights. A depiction of the baseline model is provided in Figure 1.

The models are a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN) followed by an aggregation layer (in our case an attention layer). The output of the RNN gives strong predictions (the weights of this model are denoted θ_s) while the output of the aggregation layer gives the weak predictions (the weights of this model are denoted θ).

The student model is trained on the synthetic and weakly labeled data. The loss (binary cross entropy) is computed at the frame level for the strongly labeled synthetic data and at the clip level for the weakly labeled data. The teacher model is not trained, rather, its weights are a moving average of the student model (at each epoch). During training, the teacher model receives the same input as the student model but with added Gaussian noise, and helps train the student model via a consistency loss (mean-squared error) for both strong (frame-level) and weak predictions. Every batch contains a combination of unlabeled, weakly and strongly labeled samples.

This results in four loss components: two for classification (weak and strong) and two for consistency (weak and strong), which are combined as follows:

$$\begin{aligned} L(\theta) = & L_{class_w}(\theta) + \sigma(\lambda) L_{cons_w}(\theta) \\ & + L_{class_s}(\theta_s) + \sigma(\lambda) L_{cons_s}(\theta_s) \end{aligned} \quad (1)$$

4. SUBMISSION EVALUATION

DCASE 2019 Task 4 obtained 57 submissions from 18 different teams involving 60 researchers overall.

[2]The code for the baseline model is open source and available on: https://github.com/turpaultn/DCASE2019_task4/tree/public/baseline

Rank	System	Classifier	Real recordings Event-based **Eval**	Youtube	Vimeo	Valid	Segment-based Eval	Synthetic Event-based Set 1
1	**Lin, ICT**	CNN	**42.7%**	47.7%	29.4%	45.3%	64.8%	47.6%
2	**Delphin, OL**	CRNN	**42.1%**	45.8%	33.3%	43.6%	71.4%	59.8%
3	**Shi, FRDC**	CRNN	**42.0%**	46.1%	31.5%	42.5%	69.8%	53.2%
4	**Pellegrini, IRIT**	CRNN	**39.7%**	43.0%	30.9%	39.9%	64.7%	50.8%
5	**Yan, USTC**	CRNN	**36.2%**	38.8%	28.7%	42.6%	65.2%	41.8%
6	**Lim, ETRI**	CRNN, Ensemble	**34.4%**	38.6%	23.7%	40.9%	66.4%	42.5%
7	**Kiyokawa, NEC**	ResNet, SENet	**32.4%**	36.2%	23.8%	36.1%	65.3%	42.3%
8	**CTK, NU**	NMF, CNN	**31.0%**	34.7%	21.6%	30.4%	58.2%	46.7%
9	**ZYL, UESTC**	CNN,ResNet,RNN	**30.8%**	34.5%	21.1%	35.6%	60.9%	49.2%
10	**Kothinti, JHU**	CRNN, RBM, CRBM, PCA	**30.7%**	33.2%	23.8%	34.6%	53.1%	35.6%
11	**bolun, NWPU**	CNN, RNN, ensemble	**27.8%**	30.1%	21.7%	31.9%	61.6%	32.9%
12	**Lee, KNU**	CNN	**26.7%**	28.1%	22.9%	31.6%	50.2%	33.0%
	Baseline 2019	CRNN	**25.8%**	29.0%	18.1%	23.7%	53.7%	40.6%
13	**Agnone, PDL**	CRNN	**25.0%**	27.1%	20.0%	59.6%	60.4%	46.7%
14	**Rakowski, SRPOL**	CNN	**24.2%**	26.2%	19.2%	24.3%	63.4%	29.7%
15	**Kong, SURREY**	CNN	**22.3%**	24.1%	17.0%	21.3%	59.4%	23.6%
16	**Mishima, NEC**	ResNet	**19.8%**	21.8%	15.0%	24.7%	58.7%	33.0%
17	**Wang, NUDT**	CRNN	**17.5%**	19.2%	13.3%	22.4%	63.0%	14.0%
18	**Yang, YSU**	CMRANN-MT	**6.7%**	7.6%	4.6%	19.4%	26.3%	7.5%

Table 3: F1-score performance on the evaluation sets

4.1. Evaluation metrics

Submissions were evaluated according to an event-based F1-score with a 200 ms collar on the onsets and a collar on the offsets that is the greater of 200 ms and 20% of the sound event's length. The overall F1-score is the unweighted average of the class-wise F1-scores (macro-average). In addition, we provide the segment-based F1-score on 1 s segments as a secondary measure. The metrics are computed using the sed_eval library [23].

4.2. System performance

The official team ranking (best system from each team) along with some characteristics of the submitted systems is presented in Table 3. Submissions are ranked according to the event-based F1-score computed over the real recordings in the evaluation set. For a more detailed comparison, we also provide the event-based F1-score on the YouTube and Vimeo subsets and the segment-based F1-score over all real recordings. The event-based F1-score on the validation set is reported for the sake of comparison with last year's results (75% of the 2019 validation is comprised of the 2018 evaluation set). The performance on synthetic recordings is not taken into account in the ranking, but the event-based F1-score on Synthetic set 1 (0 dB) is presented here as well.

Twelve teams outperform the baseline with the best systems [24, 25, 26] outperforming the baseline by 16% points and the best system from 2018 by over 10 % points. While the ranking on the YouTube subset is similar to the official ranking, there rankings based on the Vimeo and synthetic subsets are notably different. Performance on the Vimeo set is in general considerably lower than on the YouTube set and Synthethic set 1. The fact that no data from Vimeo was used during training (unlike data from YouTube and synthetic data) suggests that the submitted systems struggle to generalize to an entirely unseen set of recording conditions.

All three top performing teams used a semi-supervised mean-teacher model [22]. Lin and Wang [24] focused on the importance of semi-supervised learning with a guided learning setup [27] and on how synthetic data can help within this setup when used together with a sufficient amount of real data. Delphin-Poulat and Plapous [25] focused on data augmentation and Shi [26] focused on a specific type of data augmentation where both audio files and their labels are mixed. Cances et al. [28] proposed a multi-task learning setup where audio tagging (producing weak predictions) and the sound event localization in time (strong predictions) are treated as two separate subtasks [29]. The latter was also the least complex of the-performing systems.

Most of the top-performing systems also demonstrate the importance of employing class dependent post-processing [24, 25, 28], which improves performance significantly compared to e.g. using a fixed median filtering approach. This highlights the benefits of applying dedicated segmentation post-processing [28, 30].

5. CONCLUSION

This paper presents DCASE 2019 Task 4 and the DESED dataset, which focus on SED in domestic environments. The goal of the task is to exploit a small dataset of weakly labeled sound clips together with a larger unlabeled dataset to perform SED. An additional training dataset composed of synthetic soundscapes with strong labels is provided in order to explore the gains achievable with simulated data. The best submissions from this year outperform last year's winning submission by over 10 % points, representing a notable advancement. Evaluation on different subsets, and in particular the Vimeo subset, suggests there is still a significant challenge in generalizing to unseen recording conditions.

6. ACKNOWLEDGMENT

The authors would like to thank the Hamid Eghbal-Zadeh from Johannes Kepler University (Austria) who participated to the initial discussions about this task as well as all participants to the task.

7. 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," *Communications of the ACM*, In press, 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. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *Proc. WASPAA*. IEEE, 2005, pp. 158–161.

[4] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[5] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Proc. Interspeech*, 2012.

[6] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," July 2018, proc. DCASE Workshop.

[7] 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.

[8] Y. Zigel, D. Litvak, and I. 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.

[9] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. DCASE Workshop*, 2017.

[10] R. Serizel and N. Turpault, "Sound Event Detection from Partially Annotated Data: Trends and Challenges," in *Proc. IcETRAN conference*, June 2019.

[11] 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.

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

[13] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. WASPAA*. IEEE, 2017, pp. 344–348.

[14] J. F. Gemmeke, D. P. W. 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 *Proc. ICASSP*, 2017.

[15] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. ACMM*. ACM, 2013, pp. 411–412.

[16] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proc. ISMIR*, Suzhou, China, 2017, pp. 486–493.

[17] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, 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," in *Proc. DCASE Workshop*, November 2017, pp. 32–36.

[18] E. J. Humphrey, J. Salamon, O. Nieto, J. Forsyth, R. Bittner, and J. P. Bello, "JAMS: A JSON annotated music specification for reproducible MIR research," in *15th Int. Soc. for Music Info. Retrieval Conf.*, Taipei, Taiwan, Oct. 2014, pp. 591–596.

[19] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[20] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. ISMIR*, 2013, pp. 83–88.

[21] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," DCASE2018 Challenge, Tech. Rep., September 2018.

[22] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NeurIPS*, 2017, p. 10.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, May 2016.

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

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

[26] Z. Shi, "Hodgepodge: Sound event detection based on ensemble of semi-supervised learning methods," Fujitsu Research and Development Center, Beijing, China, Tech. Rep., June 2019.

[27] L. Lin, X. Wang, H. Liu, and Y. Qian, "What you need is a more professional teacher," *arXiv preprint arXiv:1906.02517*, 2019.

[28] L. Cances, T. Pellegrini, and P. Guyot, "Multi task learning and post processing optimization for sound event detection," IRIT, Universit de Toulouse, CNRS, Toulouse, France, Tech. Rep., June 2019.

[29] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[30] S. Kothinti, G. Sell, S. Watanabe, and M. Elhilali, "Integrated bottom-up and top-down inference for sound event detection," Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, USA, Tech. Rep., June 2019.