WILDDESED: AN LLM-POWERED DATASET FOR WILD DOMESTIC ENVIRONMENT SOUND EVENT DETECTION SYSTEM

Yang Xiao and Rohan Kumar Das

Fortemedia Singapore, Singapore
{xiaoyang, rohankd}@fortemedia.com

ABSTRACT

This work aims to advance sound event detection (SED) research by presenting a new large language model (LLM)-powered dataset namely wild domestic environment sound event detection (WildDESED). It is crafted as an extension to the original DESED dataset to reflect diverse acoustic variability and complex noises in home settings. We leveraged LLMs to generate eight different domestic scenarios based on target sound categories of the DESED dataset. Then we enriched the scenarios with a carefully tailored mixture of noises selected from AudioSet and ensured no overlap with target sound. We consider widely popular convolutional neural recurrent network to study WildDESED dataset, which depicts its challenging nature. We then apply curriculum learning by gradually increasing noise complexity to enhance the model's generalization capabilities across various noise levels. Our results with this approach show improvements within the noisy environment, validating the effectiveness on the WildDESED dataset promoting noise-robust SED advancements.

Index Terms— sound event detection, DESED, noisy scenario, noise robust SED, curriculum learning

1. INTRODUCTION

Sounds play a vital role in our lives, helping us understand our surroundings and notice changes. Sound event detection (SED) [1, 2] is essential for interpreting and responding to our environment, with applications ranging from urban noise management to smart-home technologies [3] and security systems [4]. SED has made great strides [5–7], thanks to diverse datasets [8] tailored for specific scenarios. Google AudioSet [9] provides a wide array of sounds, and MAVD [10] focuses on traffic noise. Among various SED datasets, DESED [11, 12] is well known for its focus on domestic environments, which makes it the most utilized dataset for home sound event research. However, DESED faces challenges in comprehensively representing the unpredictable and complex nature of household sounds. Hence, there exists scope for covering a wide range of domestic scenarios with common background noises that can occur in a household.

The quest for noise robustness in SED has led to the development of new methodologies and datasets [13, 14] aimed towards improving performance under challenging conditions such as noisy urban environments. Innovations by researchers like Neri et al. [15], Serizel et al. [16], and Wan et al. [17] have pushed the boundaries of SED systems by integrating deep learning and audio enhancement techniques. These studies, however, predominantly address controlled or semi-controlled environments, leaving a gap for SED systems to effectively detect sound events in the less predictable, 'wild' conditions in domestic environments.

Addressing this gap, our research contributes to the field by introducing a new dataset namely, *wild domestic environment sound event detection* (WildDESED). We proposed carefully selecting noise types from the AudioSet that accurately represent real home environments but are distinct from DESED's target sounds. This artificial selection could be challenging because of the bias and unnatural correlations. Large language models (LLMs) [18] such as GPT-4, ChatGPT, and Llama have demonstrated remarkable potential to perform various tasks [19–21] in recent years. In this regard, we utilized the strong capabilities of LLMs to analyze and summarize acoustic data for selecting specific noises. This helped us to design eight different scenarios that blend the noises with target sounds, simulating authentic domestic environments. The noises are divided into four categories based on their sources and acoustic properties, allowing for a diverse and realistic combination with target sounds. This novel approach has culminated in the creation of WildDESED dataset, specifically designed to enhance SED research in dynamic and natural home environments.

Building on this foundation, our research not only introduces the WildDESED dataset, but also explores the application of curriculum learning in the context of SED to tackle the challenges posed by domestic noisy environments. Curriculum learning [22–24] is a training approach that improves models for noisy speech [25, 26] and audio by starting with simpler, less noisy data and gradually increasing the noise level. This method is similar to the way how the humans learn and helps models adjust from clean to noisy sounds more effectively. In this work, we applied curriculum learning to the baseline convolutional recurrent neural network (CRNN) [27–29] model using the WildDESED dataset for our studies. The novelty of this work lies in the proposal of a new *in-the-wild* dataset for advancing SED research and exploring curriculum learning as an approach to develop noise-robust SED systems. We will publicly release the WildDESED dataset in due course.

2. RELATED WORK

The WildDESED dataset is an extension to the original DESED dataset, which is a foundational resource featuring 10 target sound classes pivotal for understanding the sounds in home environments. The DESED dataset consists of the following subsets: The weak set, with 1,578 real recordings labeled with weak annotations, captures the presence of sound classes without temporal specifics. The unlabeled training set includes 14,412 real, unlabeled recordings. The test set comprises of 1,168 real recordings with strong annotations to assess model performance. These three subsets are real-world recordings from AudioSet. The training synth set contains 10,000 synthetic recordings with strong annotations [30], detailing exact temporal boundaries. The synth validation set has 2,500 syn-

Table 1: A summary of different background noises used in WildDESED dataset.

Noise	Occurrences	Duration (Second)
Bird chirping outside	9,847	7,523
Car passing by outside	311	862
Chair moving	343	359
Clock ticking	2,5777	2,662
Coffee machine	6	30
Door closing	335	196
Fan noise	117	958
Footsteps	6,243	2,101
Light rain	159	1,379
Refrigerator humming	58	456
TV playing in the background	805	7,191
Wind blowing	5,467	48,648
Total	49,468	72,365

thetic recordings with strong annotations for model validation during development. These two synthetic subsets are generated with the Scaper. Their background files are extracted from SINS [31], TUT [32], MUSAN [33], or YouTube and have been selected because they contain a very low amount of our sound event classes. We propose to simulate more diverse and complex noisy scenarios that are not covered by the original DESED dataset and also introduce a controlled variability for testing.

3. WILDDESED

We extend DESED to the WildDESED for in-the-wild scenarios for domestic environments by considering three primary set of questions to address as follows:

- What type of background noises do we use?
- What are the domestic scenarios we choose?
- How do we mix the background noises to the scenarios?

GPT-4 is an advanced language model that builds on the GPT-3 architecture but uses a larger amount of training data. It includes the latest techniques to enhance understanding of natural language. In the following subsections, we will detail how we leverage GPT-4 to address each of these questions, outlining the methodology behind the creation of the WildDESED dataset. This new dataset aims to bridge the gap between the controlled environment of existing datasets and the dynamic, often unpredictable nature of real-world domestic soundscapes, thus expanding the potential for noise-robust SED research in truly 'wild' home scenarios.

3.1. What type of background noises do we use?

To construct the WildDESED dataset, we initiated our process with the foundational DESED dataset, which identifies 10 distinct sound events in 10-second audio clips. The events in DESED include diverse household sounds like alarms, appliances, pets, and running water. We input the total 356 classes from the strongly annotated subset of AudioSet to the GPT-4 together with the 10 DESED classes. Then we guide GPT-4 by the following prompt:

"Select noise classes from the 356 strongly annotated AudioSet classes, alongside the 10 DESED classes ensuring clear delineation and no overlap with DESED's sound events. Further, apply thorough filtering to exclude any AudioSet classes similar to DESED target classes, preserving the distinctiveness of the dataset."

Considering the output of GPT-4, we enhanced DESED with selected events from the strongly annotated subset of AudioSet, ensuring clear delineation and no overlap with DESED's sound events. A thorough filtering process was applied to exclude any AudioSet classes that are very similar to target classes of DESED dataset, preserving the distinctiveness of our dataset. Table 1 displays the outcome of our selection process, listing the types and quantities of noise clips integrated into WildDESED. We included a spectrum of sounds both indoor, like the clock ticking, and outdoor, such as birds chirping that capture the essence of a domestic environment. The 'clock ticking' class, for instance, has the largest event count, while 'wind blowing' spans the greatest duration, together reflecting the continuous and transient nature of home sounds.

This dataset construction ensures WildDESED encompasses a rich and authentic array of domestic noises, ready to challenge and advance SED systems in recognizing the events under complex acoustic home environments.

3.2. What are the domestic scenarios we choose?

For the WildDESED dataset, we still have to map the selected 12 noise classes with our 10 target classes. We input them to GPT-4 and use the following prompt:

"Create eight different domestic scenarios so that they should map 12 selected noise classes to the 10 target classes from the DESED dataset, crafting authentic household soundscapes. Ensure the scenarios reflect typical sounds one would encounter in a household environment."

Considering the output of LLM, we crafted eight different domestic scenarios, each mapping to target classes from the DESED dataset to create authentic soundscapes one would encounter in a household. These scenarios are constructed to reflect the typical activities and the accompanying sounds in a domestic environment.

- **Morning Routine:** Associated with *'Blender'* target sounds, this scenario captures the essence of the morning with 'Light rain', 'Refrigerator humming', 'Clock ticking', and 'TV playing in the background'.
- **Home Office:** Linked to *'Speech'* as the target class, it includes background sounds of 'Car passing by', 'Fan noise', and 'Footsteps', emulating a work-from-home setting.
- **Household Chores:** Representing *'Vacuum cleaner'* noises as the target, this scenario combines 'Door closing', 'Chair moving', and 'Footsteps' as background to depict cleaning activities.
- **Late-night:** Tied to the *'Electric shaver toothbrush'* target sound, offering the 'Clock ticking' and 'Light rain' as a backdrop for night-time routines.
- **Cooking:** Merging the target sounds of *'Frying'* and *'Dishes'* with 'Coffee machine' buzzes and 'Refrigerator humming', this scenario is bustling with culinary activity.
- **Pet Care:** Incorporating target sounds of *'Cat'* and *'Dog'*, this setting is further brought to life with 'Bird chirping outside' and 'TV playing in the background'.
- **Bathroom Routine:** Linked to *'Running water'* as the target sound, with added 'Fan noise' and 'Wind blowing', simulating personal care sounds.
- **Emergency:** Associated with the *'Alarm bell ringing'* target sound, it layers urgent sounds like 'Refrigerator humming' and 'Fan noise' with 'Clock ticking' and 'Car passing by'.

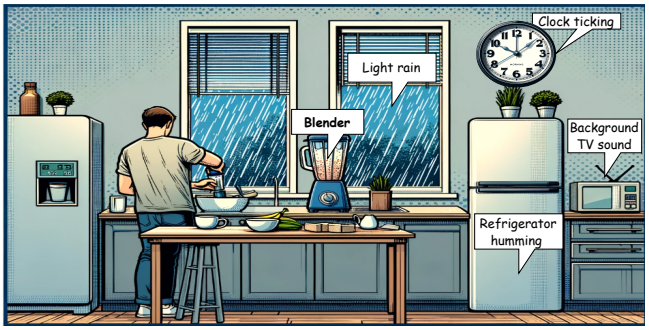

Figure 1: Illustration[2] of Morning Routine Scenario out of the total eight scenarios in WildDESED dataset. In the scenario, key target sound events are written in bold fonts, along with added different background noises to simulate real-life settings.

Each scenario's sound design is a thoughtful blend of target and noise classes, chosen to challenge the detection capabilities of SED systems within the rich and varied auditory context of a home environment. To illustrate our scenarios, we present a Figure 1 that showcases two typical scenarios out of the eight: the 'Pet Care Scenario' and the 'Morning Routine Scenario'. This figure highlights key target sound events within each scenario, incorporating strategically placed background noises to simulate the real-life acoustic challenges found in domestic settings.

3.3. How do we mix the background noises to the scenarios?

In the WildDESED dataset, the integration of background noises into the selected domestic scenarios is meticulously structured around a quadrant based on the acoustic characteristics of the noises. The quadrant categorizes noises into four groups: Ambient Environmental Sounds, Human-related and Intermittent Sounds, Mechanical and Electronic Sounds, and Nature and Outdoor Sounds, as illustrated in Figure 2.

- For **Ambient Environmental Sounds**, such as 'Light rain' and 'Wind blowing', we repeated these sounds to cover the entire duration of the audio clip from the original DESED dataset. These sounds are mixed at a low intensity to ensure they provide a consistent background atmosphere without overpowering the primary sound events. The rationale behind this is to create an unobtrusive ambient layer that emulates the continuous presence of these sounds in a typical home environment.
- Sounds like 'Footsteps', 'Door closing', and 'Chair moving' fall into the **Human-Related and Intermittent Sounds** category. These are inserted at random intervals to simulate the sporadic nature of human movement and activities within a home. The volume and frequency of these sounds are varied to reflect the realistic and unpredictable nature of their occurrence in daily life.
- **Mechanical sounds**, including 'Clock ticking' and 'Coffee machine', are inserted at specific points to coincide with the actions they represent, such as a coffee machine being used during morning routines. The volume is set to be noticeable but not overwhelming, ensuring the sound is recognized as a part of the scenario without becoming a large distraction.
- Lastly, **Nature and Outdoor Sounds** like 'Car passing by outside' and 'Bird chirping outside' are incorporated randomly to enhance the realism of external environmental influences. The volume may fluctuate to mimic the variable volume of these sounds in real settings, contributing to the unpredictability and diversity of the overall soundscape.

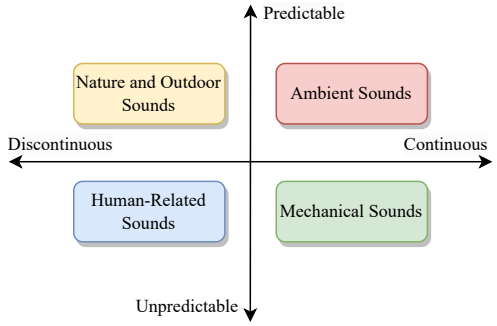

Figure 2: Quadrant showing four groups of noise types based on their acoustic characteristics considered in the WildDESED.

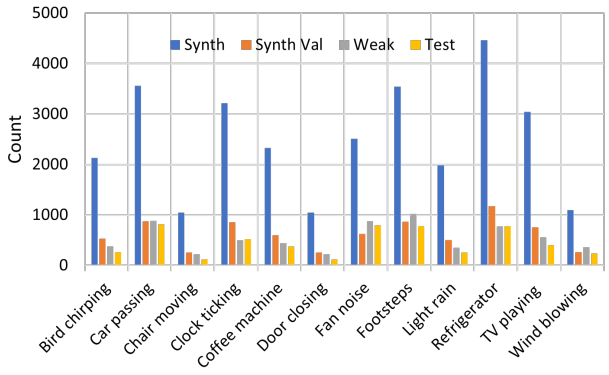

Figure 3: Statistics of noises in the WildDESED subsets.

Each noise type and its corresponding mixing approach are tailored to maintain the authenticity of the domestic scenarios. This methodical and scenario-specific approach to mix noises ensures that the WildDESED dataset not only presents a challenge for SED systems but also closely reflects the complex acoustic environments of actual domestic settings.

In finalizing the composition of the WildDESED dataset, special consideration was given to the representation of the 'speech' sound class due to its prevalence and significance in domestic environments. For the 'Home Office' scenario in synth set and synth val set, we exclusively selected clips that featured the 'speech' class in isolation, omitting any clips where 'speech' occurred alongside other sound events.

Figure 3 displays class-wise statistics for different background noises in each subset of the WildDESED dataset, indicating the prevalence of each noise type, within synth, synth val, weak, and test subsets. Figure 4 shows scenario-wise statistics for the scenarios in the WildDESED dataset, quantifying how frequently each scenario appears in each subset. Through this detailed dataset structure, WildDESED positions itself as a crucial resource for developing and evaluating SED systems, equipping researchers with the means to advance the field of SED in naturalistic home environments.

[2]Figures generated using DALL-E-2 (https://openai.com/dall-e-2)

Table 2: Performance in PSDS1 (P1), PSDS2 (P2) and PSDS1 + PSDS2 (P1 + P2) of the proposed curriculum learning (CL) approach on the DESED devtest set and our proposed WildDESED (W) dataset with SNR in dB.

Model	Performance on DESED			Performance on WildDESED											
				10dB			5dB			0dB			-5dB		
	P1	P2	P1 + P2	P1	P2	P1 + P2	P1	P2	P1 + P2	P1	P2	P1 + P2	P1	P2	P1 + P2
CRNN	0.344	0.543	**0.887**	0.222	0.409	0.631	0.148	0.302	0.450	0.064	0.174	0.238	0.017	0.078	0.095
CRNN (W)	0.200	0.329	0.529	0.175	0.337	0.512	0.135	0.303	0.438	0.087	0.242	0.329	0.048	0.174	0.222
CRNN (W+ CL)	0.265	0.461	0.726	0.212	0.443	**0.655**	0.175	0.390	**0.565**	0.114	0.317	**0.431**	0.049	0.211	**0.260**

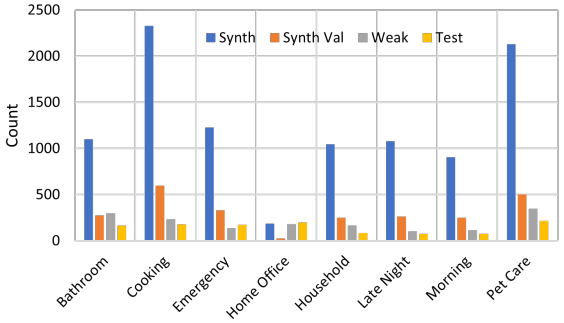

Figure 4: Statistics of the scenarios in the WildDESED subsets.

4. CURRICULUM LEARNING FOR NOISE-ROBUST SED

We use a curriculum learning [22, 26] method to develop noise-robust SED systems. This approach introduces complexity in stages, starting with simple tasks and gradually integrating noise at various signal-to-noise ratios (SNR), aligning with our goal to augment the model's resilience to noise.

We have five stages in our methodology, each with an increasing level of noise difficulty. Initially, the model learns from clean audio samples. This foundational step is crucial for establishing an understanding of the sound events without the confounding presence of noise. We then incrementally introduce noise, decreasing the SNR by 5dB in subsequent stages. Let N be the total number of training samples. Given k noise levels $L = [L_1, L_2, \ldots, L_k]$, the dataset D is composed as follows:

$$D = \bigcup_{i=1}^{k} \{D_i\}, D_i = \frac{N}{k} \text{ samples at noise level } L_i \quad (1)$$

The k in our experiment here is 5 including the clean DESED, and noise levels 10dB, 5dB, 0dB, and -5dB are considered. The model's progress is meticulously monitored, and a validation metric c is used to evaluate learning at each epoch. In our approach, the c is the intersection f1-score. If c fails to improve for ten consecutive epochs [34], the best-performing model state is reloaded, and the training progresses to the next noise level.

5. EXPERIMENTAL SETTINGS

5.1. Dataset and Evaluation Metric

We considered the DESED dataset and our proposed WildDESED dataset, featuring 10-second audio clips across various subsets. All clips were resampled to 16 kHz mono and segmented using a 2048-sample window and 256-sample hop length for spectrogram extraction and log-mel spectrogram generation. Our systems were evaluated using the threshold-independent polyphonic sound event detection scores (PSDS) [35] in two scenarios following DCASE 2023 Challenge Task 4A protocol. Scenario-1 focuses on prompt reaction and temporal localization, while Scenario-2 emphasizes on reducing class confusion for SED.

5.2. Implementation Details

For our experiments, following the DCASE 2023 Task 4A baseline [28], we utilized a batch size of 48 and employed the Adam optimizer with an initial learning rate of 0.001, coupled with an exponential warmup scheduler applied across the first 50 epochs out of a total 200 epochs. To stabilize training, we implemented a mean teacher model with an exponential moving average [36] factor set at 0.999. We consider the CRNN [28] baseline system from DCASE 2023 Task 4A, featuring approximately 1.2 million parameters, ensuring a robust comparison for our curriculum learning approach.

6. RESULTS AND DISCUSSION

Table 2 shows the results of our studies on DESED and newly created WildDESED datasets. It is observed that the performance of the baseline CRNN model trained using DESED dataset drops significantly as the noise levels are increased on WildDESED dataset compared to that on the original DESED dataset. We then explore the baseline CRNN model trained using WildDESED data, which we refer to as CRNN (W). We find that CRNN (W) performs better than the original CRNN model when the noise levels on WildDESED are on the higher end (0 dB and -5 dB). However, the performance is comparable for both models when noise level is 5 dB and then the original CRNN model performs better for less noisy scenario of 10dB on WildDESED and on the clean DESED dataset.

We now focus on the studies for curriculum learning approach applied on the CRNN model trained using WildDESED dataset. We refer this model as CRNN (W+CL) and find that it outperforms both CRNN as well as CRNN (W) models for all noise levels on the WildDESED dataset. This highlights the scope of curriculum learning approach for developing noise-robust SED systems using WildDESED dataset for unseen complex domestic settings. We also note that the CRNN model trained on the clean DESED performs the best on the DESED test due to the matched conditions. However, the model CRNN (W+CL) with curriculum learning certainly helps to boost the performance of the CRNN (W) model trained on WildDESED dataset to bring it closer that of the CRNN model on DESED test set. The future work will focus on reducing this performance gap on the clean scenario for noise-robust SED models.

7. CONCLUSION

In this work, we have presented a new dataset referred to as WildDESED to advance SED research under noisy home settings and also explored a preliminary curriculum learning method to develop noise-robust SED systems. We used 12 noises from Audioset to craft the WildDESED dataset considering 8 different scenarios depicting complex home environments by considering assistance from an LLM. The studies conducted showed the scope of curriculum learning approach for developing noise-robust SED systems using the WildDESED dataset. We believe this WildDESED dataset will be useful for future horizons of noise-robust SED research.

8. REFERENCES

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound Event Detection: A Tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[2] T. Khandelwal, R. K. Das, and E. S. Chng, "Sound Event Detection: A Journey Through DCASE Challenge Series," *APSIPA Transactions on Signal and Information Processing*, vol. 13, 2024.

[3] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio Analysis for Surveillance Applications," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, 2005, pp. 158–161.

[4] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A System for Monitoring, Analyzing, and Mitigating Urban Noise Pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[5] Y. Xiao and R. K. Das, "Dual Knowledge Distillation for Efficient Sound Event Detection," *arXiv:2402.02781*, 2024.

[6] Y. Xiao, T. Khandelwal, and R. K. Das, "FMSG Submission for DCASE 2023 Challenge Task 4 on Sound Event Detection with Weak Labels and Synthetic Soundscapes," DCASE 2023 Challenge, Tech. Rep., 2023.

[7] F. Ronchini and R. Serizel, "Performance and Energy Balance: A Comprehensive Study of State-of-the-art Sound Event Detection Systems," *arXiv:2310.03455*, 2023.

[8] H. Dinkel, M. Wu, and K. Yu, "Towards Duration Robust Weakly Supervised Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 887–900, 2021.

[9] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[10] P. Zinemanas, P. Cancela, and M. Rocamora, "MAVD: A Dataset for Sound Event Detection in Urban Environments." in *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

[11] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," in *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

[12] N. Turpault and R. Serizel, "Training Sound Event Detection on a Heterogeneous Dataset," in *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

[13] Y. Choi, O. Atif, J. Lee, D. Park, and Y. Chung, "Noise-robust Sound-event Classification System with Texture Analysis," *Symmetry*, vol. 10, no. 9, p. 402, 2018.

[14] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[15] M. Neri, F. Battisti, A. Neri, and M. Carli, "Sound Event Detection for Human Safety and Security in Noisy Environments," *IEEE Access*, vol. 10, pp. 134 230–134 240, 2022.

[16] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound Event Detection in Synthetic Domestic Environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 86–90.

[17] T. Wan, Y. Zhou, Y. Ma, and H. Liu, "Noise Robust Sound Event Detection Using Deep Learning and Audio Enhancement," in *Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2019, pp. 1–5.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language Models are Few-shot Learners," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

[19] "(toolqa: A dataset for llm question answering with external tools."

[20] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, "Listen, Think, and Understand," in *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[21] J. Bai, H. Yin, M. Wang, D. Shi, W.-S. Gan, J. Chen, and S. Rahardja, "AudioLog: LLMs-Powered Long Audio Logging with Hybrid Token-Semantic Contrastive Learning," *arXiv preprint:2311.12371*, 2023.

[22] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum Learning," in *Proc. Annual International Conference on Machine Learning (ICML)*, 2009, pp. 41–48.

[23] X. Wang, Y. Chen, and W. Zhu, "A Survey on Curriculum Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[24] A. Pentina, V. Sharmanska, and C. H. Lampert, "Curriculum Learning of Multiple Tasks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5492–5500.

[25] S. Braun, D. Neil, and S.-C. Liu, "A Curriculum Learning Method for Improved Noise Robustness in Automatic Speech Recognition," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2017, pp. 548–552.

[26] D. Ng, Y. Xiao, J. Q. Yip, Z. Yang, B. Tian, Q. Fu, E. S. Chng, and B. Ma, "Small Footprint Multi-channel Network for Keyword Spotting with Centroid Based Awareness," in *Proc. INTERSPEECH*, 2023, pp. 296–300.

[27] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[28] F. Ronchini, S. Cornell, R. Serizel, N. Turpault, E. Fonseca, and D. P. W. Ellis, "Description and Analysis of Novelties Introduced in DCASE Task 4 2022 on the Baseline System," in *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2022.

[29] T. Khandelwal and R. K. Das, "A Multi-Task Learning Framework for Sound Event Detection using High-level Acoustic Characteristics of Sounds," in *Proc. INTERSPEECH*, 2023, pp. 1214–1218.

[30] F. Ronchini, R. Serizel, N. Turpault, and S. Cornell, "The Impact of Non-Target Events in Synthetic Soundscapes for Sound Event Detection," in *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021, pp. 115–119.

[31] 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," in *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2017, pp. 1–5.

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

[33] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint:1510.08484*, 2015.

[34] L. Prechelt, "Early Stopping-but When?" *Neural Networks: Tricks of the trade*, pp. 55–69, 2002.

[35] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Post-Processing Independent Evaluation of Sound Event Detection Systems," in *Proc. the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2023.

[36] A. Tarvainen and H. Valpola, "Mean Teachers are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.