HAL open science

Performance and Energy Balance: A Comprehensive Study of State-of-the-Art Sound Event Detection Systems

Francesca Ronchini, Romain Serizel

To cite this version:

Francesca Ronchini, Romain Serizel. Performance and Energy Balance: A Comprehensive Study of State-of-the-Art Sound Event Detection Systems. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr 2024, Seoul, South Korea. pp.1096-1100, ⟨10.1109/ICASSP48485.2024.10445834⟩. ⟨hal-04892368⟩

HAL Id: hal-04892368

https://inria.hal.science/hal-04892368v1

Submitted on 16 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC 4.0 - Attribution - Non-commercial use - International License

PERFORMANCE AND ENERGY BALANCE: A COMPREHENSIVE STUDY OF STATE-OF-THE-ART SOUND EVENT DETECTION SYSTEMS

Francesca Ronchini[1], *Romain Serizel*[2]

[1] Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, Milan, Italy
[2]Université de Lorraine, CNRS, Inria, Loria, Nancy, France

arXiv:2310.03455v2 [eess.AS] 16 Jan 2024

ABSTRACT

In recent years, deep learning systems have shown a concerning trend toward increased complexity and higher energy consumption. As researchers in this domain and organizers of one of the Detection and Classification of Acoustic Scenes and Events challenges task, we recognize the importance of addressing the environmental impact of data-driven SED systems. In this paper, we propose an analysis focused on SED systems based on the challenge submissions. This includes a comparison across the past two years and a detailed analysis of this year's SED systems. Through this research, we aim to explore how the SED systems are evolving every year in relation to their energy efficiency implications [1].

***Index Terms*—** sound event detection, machine listening, energy consumption, carbon footprint

1. INTRODUCTION

Deep learning has yielded incredible achievements in a variety of audio processing applications, including Speech Recognition [1], Machine Listening [2, 3], and Music Generation [4, 5]. Despite the outcome of the remarkable results, the computational overhead of deep learning remains significant and continues to increase [6]. One particular concern is the substantial energy usage and resulting carbon footprint tied to the computational needs of deep learning. This situation is rapidly becoming unsustainable from both technical and environmental perspectives [7–9]. Until very recently, the environmental impact of deep learning models has largely been dominated by the persistent demand for high accuracy and effectiveness [10]. In the last years, there has been a rise of concern about the environmental impact and energy consumption of deep learning within audio signal processing communities [10–14].

However, comparing accurately the energy of different models, possibly trained on different sites is not straightforward [12]. There is no real consensus on the metric to be used and the relation between the different potential metrics (complexity, MACS, energy consumption) is often unclear. In the last years, different open-source Python packages have been introduced by the community in order to face this issue [15–18]. Anyway, none of them provide a complete overview of the environmental impact. In fact, several factors must be taken into account when quantifying the energy consumption of deep learning algorithms [10, 12].

Starting from 2018, the objective of DCASE task 4 has been to investigate SED using a heterogeneous dataset featuring audio soundscapes with varying levels of label detail [19]. Throughout the years, alongside the ranking procedure, the evaluation of diverse submissions in practical scenarios has been crucial to gaining insights into the systems' performance [20–23]. In DCASE task 4, as in many other audio tasks, there has been a consistent trend of models steadily increasing in parameter complexity, frequently incorporating ensemble techniques. As also reported in Serizel et. al [12], as organizers of the DCASE task 4, we recognized a responsibility to raise awareness about the carbon emissions and environmental implications associated with data-driven SED systems.

In 2022, we asked participants to report the energy consumption of their systems both at training and test time [20]. We also introduced a new energy consumption metric [24], based on CodeCarbon toolkit [17]. The metric has been introduced as an optional metric due to possible biases in terms of hardware used and fairness of comparison between systems. Since 2023, energy consumption reporting is mandatory. We also asked every participant to report energy consumption for 10 epoch of training of the baseline on their setup. This aims at normalizing the energy consumption metric by accounting for possible hardware disparities [12]. We further asked participants to report an additional hardware-agnostic metric involving the computation of Multiply-Accumulate operations (MACs) for ten seconds of audio prediction. We employed *THOP: PyTorch-OpCounter* as a framework for MACs computation [18].

This paper presents an analysis of the general evolution trend between 2022 and 2023. Following this initial overview, we focus on SED systems submitted in 2023. The metrics gathered in 2023 include hardware normalization and allow for a fairer comparison. The goal of the analysis is to provide insights for achieving a better balance between performance and energy efficiency in SED system development.

2. ANALYSIS SETUP AND EVALUATION METRICS

The analysis is conducted on DCASE task 4 submissions in 2022 and 2023. For more information on system submissions, the reader is invited to visit the DCASE Challenge website [2]. Within DCASE task 4, systems performance is evaluated with the polyphonic sound event detection scores (PSDS) [25]. PSDS allows users to define parameter sets, defining customized scenarios under which SED systems are evaluated. For DCASE task 4, two scenarios are considered, as described in Ronchini et. al [20]. The PSDS is indicated throughout the paper as **PSDS_1** when evaluated on scenario 1 and **PSDS_2** when evaluated on scenario 2, regardless of the system.

From 2022, we propose a tentative, trivial energy weighted polyphonic sound detection score (EW-PSDS):

[1]The data related to the submissions used for the analysis are available at `https://github.com/RonFrancesca/SED-carbon-footprint`

[2]https://dcase.community/challenge2023/task-sound-event-detection-with-weak-labels-and-synthetic-soundscapes

© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

	System complexity ↓			Energy train (kWh) ↓			Energy test (kWh) ↓		
	25%	Median	75%	25%	Median	75%	25%	Median	75%
2022 Entries	2200000	6676303	18903660	1.815	3.699	17.291	0.010	0.026	0.046
2023 Entries	4804956	14662273	97176570	1.615	4.295	13.975	0.019	0.035	0.283

Table 1. General comparison between DCASE 2022 and DCASE 2023 submissions entries. The table presents the median, 25th percentile, and 75th percentile values for system complexity, training energy, and test energy.

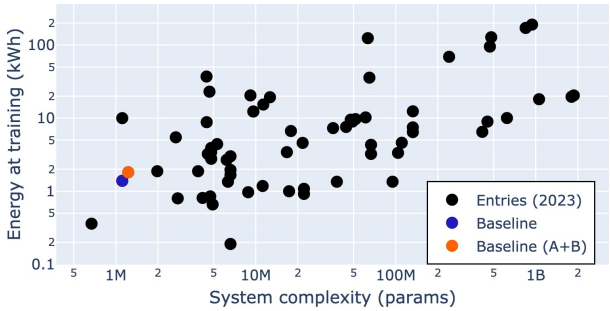

Fig. 1. Relation between system complexity and energy consumption at training for 2023 entries, compared with the two baselines systems.

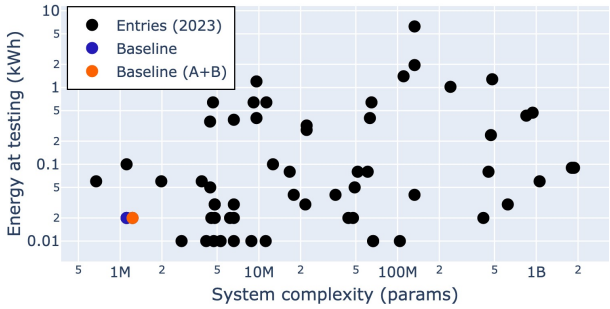

Fig. 2. Relation between system complexity and energy consumption at test for 2023 entries, compared with the two baselines systems.

$$\text{EW} - \text{PSDS} = \text{PSDS} * \frac{\text{kWh}_{\text{baseline}}}{\text{kWh}_{\text{submission}}} \quad (1)$$

where PSDS is the polyphonic sound event detection scores [25], $\text{kWh}_{\text{baseline}}$ is the energy consumption reported for the baseline, and $\text{kWh}_{\text{submission}}$ is the energy consumption of the submitted system [24].

The energy usage during both the training and test stages is reported using kWh (kilowatt-hour) as the unit.

This study presents an analysis of the relation between system performance metrics and energy consumption-related measures [3]. We report a subset of these metrics, along with MACs, system complexity (number of parameters of the deep learning model), and energy consumption. The energy usage is measured during both the training and inference stages.

[3] Additional results are available at `https://github.com/RonFrancesca/SED-carbon-footprint`

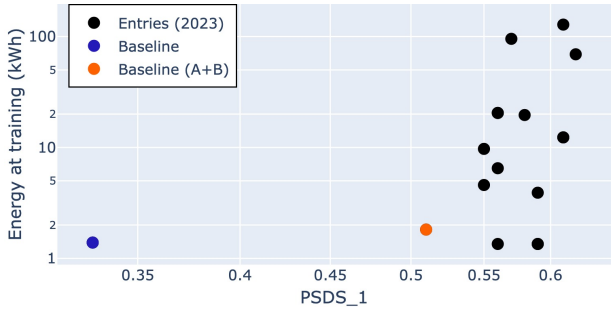

Fig. 3. PSDS_1 and energy consumption at training for best 2023 systems, compared with the two baselines systems.

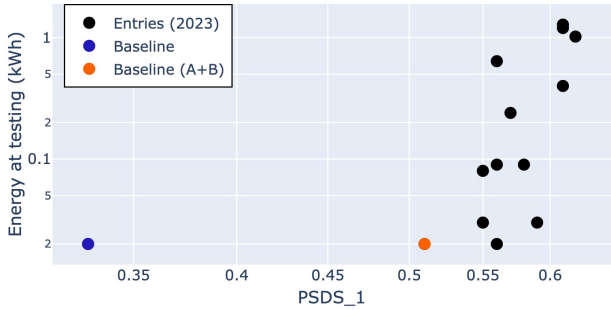

Fig. 4. PSDS_1 and energy consumption at test for best performance 2023 systems, compared with the two baselines systems.

3. GENERAL COMPARISON BETWEEN DCASE 2022 AND DCASE 2023 SYSTEMS

This section analyzes all entries from 2022 and 2023 to understand how energy-related metrics have evolved over the past two years. In this study, we compare median, 25th percentile, and 75th percentile, for system complexity, energy consumption during training, and energy consumption during test. Quartiles and the median were chosen as measures of central tendency instead of the mean and standard deviation due to the presence of significant data variability. We conducted a filtering process in order to remove duplicate entries and entries with erroneous reported metrics. Initially, we had 101 total entries for 2022 and 123 total entries for 2023. After filtering, the analysis focused on 60 entries for 2022 and 64 entries for 2023. Table 1 shows the results for system complexity, training energy consumption, and test energy for both 2022 and 2023 entries. Unsurprisingly, it can be observed a similar trend as the one discussed in Section 1 for deep learning models. There is an inclination towards increased system complexity and energy consumption during both training and test. While this is true, it is interesting to note that the energy consumption at training of the 75th percentile has remained stable or even decreased.

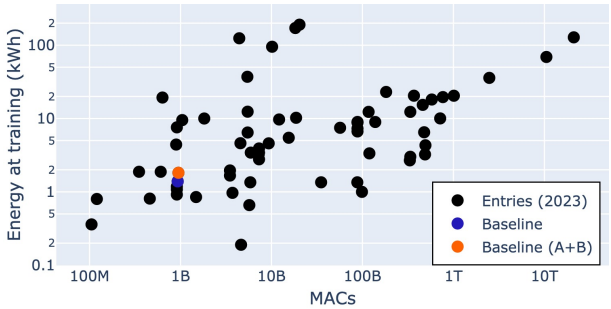

Fig. 5. Relation between MACs and energy consumption at training for 2023 entries, compared with the two baselines systems.

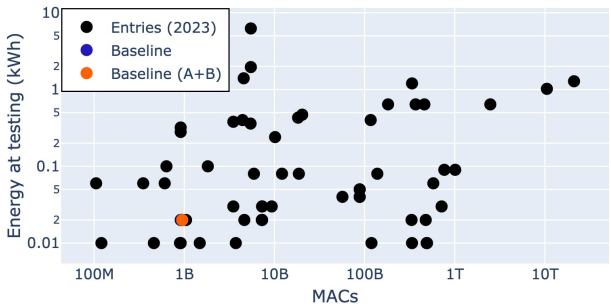

Fig. 6. Relation between MACs and energy consumption at test for 2023 entries, compared with the two baselines systems.

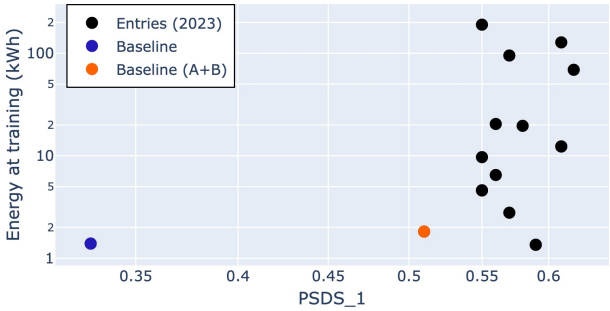

Fig. 7. Relation between PSDS_1 and energy consumption at training for the best ensemble systems.

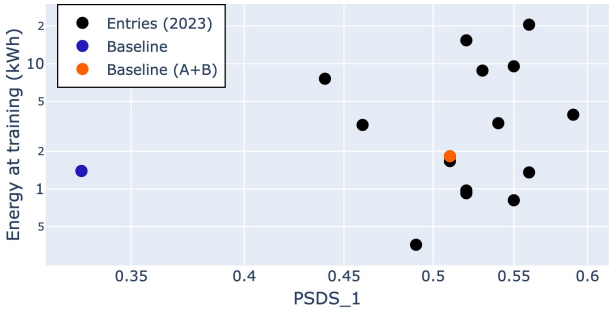

Fig. 8. Relation between PSDS_1 and energy consumption at training for the best not-ensemble systems.

However, it's important to note that the energy consumption data for this general analysis is not normalized, which means we can hardly draw objective conclusions due to potential hardware biases. Additionally, not all 2022 submissions provided energy consumption values. This initial study is only presented to observe general trends. To ensure fairness and a more insightful study, the rest of the analysis is exclusively focused on results related to 2023 submissions.

4. RELATION BETWEEN SYSTEM COMPLEXITY, MACS, AND ENERGY CONSUMPTION

In this section, we present an analysis of the MACs and system complexity in relation to energy consumption during the training and test phases of the DCASE 2023 submissions. To ensure a fair comparison among systems, energy consumption has been normalized by the baseline energy consumption [12].

Figure 1 illustrates the correlation between system complexity and energy consumption during the training stage, Figure 2 presents the corresponding relation for energy consumption during the test stage. Figure 5 and Figure 6 show the relation between MACs and energy consumed at the training phase, and the energy consumed at the test phase, respectively. On all the plots, we also report the performance of the two baselines proposed for the challenge. *Baseline* indicates the simple baseline which does not use external dataset or embeddings, while *Baseline (A+B)* stands for the baseline using Audioset as external dataset [26] and BEATs embeddings [27] [4].

From the results, it is possible to observe that MACs correlate slightly more with energy at training than system complexity. It is surprising that MACs correlate better with training energy than test energy, while the other way around would be expected as MACs are computed at test time. This difference might be related to different system architectures but this would have to be verified in extensive experiments.

The relation between system complexity with the energy consumed by the system at both training and test phases is not straightforward. These three different metrics independently are insufficient to provide a comprehensive understanding of the system's footprint. In fact, as an example, from Figure 1, there are systems with a complexity of 5M parameters that consume more energy than systems with a complexity of 100M parameters. Similar observations apply to MACs counts and energy consumption. However, for simplicity sake, we will focus the analysis mainly on energy consumption for the rest of the paper.

5. RELATION BETWEEN PERFORMANCE AND ENERGY CONSUMPTION

This section presents findings regarding the relation between performance and energy consumption. We focus on the top 15 systems in terms of PSDS_1 performance along with training and test energy consumption. If the same team has more than one submission with values within a range of 1-2 points in terms of PSDS_1, we selected only one submission per team, the best in terms of PSDS_1. We removed duplicate entries for each team based on MACs, keeping the best-performing system for each team. [5]

Figure 3 and Figure 4 illustrate the performance metrics PSDS_1 in relation to training energy consumption, and test energy consumption, respectively. The figures highlight the diversity in terms of sys-

[4] See also the task webpage [2] for more details.

[5] A similar analysis for PSDS_2 is reported on the additional results.

	System complexity		MACs		Energy train (kWh)		System complexity		MACs		Energy train (kWh)	
	Max	PSDS_1	Max	PSDS_1	Max	PSDS_1	Max	PSDS_1	Max	PSDS_1	Max	PSDS_1
All	1B	0.59	492 B	0.59	23.00	0.59	1B	0.62	21 T	0.62	190.00	0.62
25th	5 M	0.55	912 M	0.55	0.99	0.55	25 M	0.61	8 B	0.58	4.59	0.60
Median	6 M	0.59	4 B	0.55	2.33	0.56	67 M	0.61	72 B	0.60	9.34	0.60

Table 2. The table presents PSDS_1 when system complexity MACS and training energy are thresholded to the median value or the 25th percentile. The left side is related to no-ensemble systems, the right is related to ensemble systems.

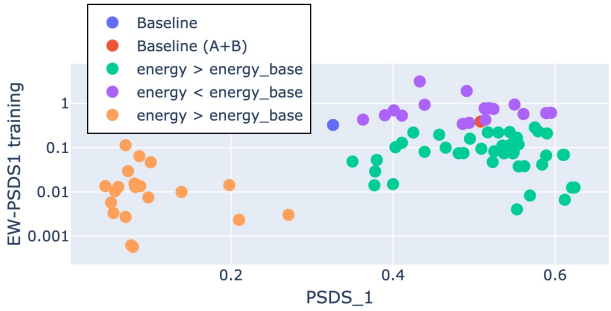

Fig. 9. Relation between PSDS_1 and energy consumption weighted at training for 2023 entries, compared with the two baselines systems.

tem efficiency. Particularly for PSDS_1, some systems manage to outperform the baseline results while consuming less energy. Additionally, should be noted that the top-performing systems are not the systems that consume the most energy. The same observation holds true for energy consumption during test. In fact, from the results, it appears that energy consumed for training and test have a similar distribution. In the remainder of the paper, we will focus on energy at training. We know that energy at test is the most important in terms of the footprint of deployed systems yet it also depends on the hardware target and is difficult to tackle in the current challenge setup. Additionally, according to previous observations energy at training can be considered as a first (gross) indicator of what would happen at the test (even though many other factors are involved).

6. COMPARISON BETWEEN ENSEMBLE/NON-ENSEMBLE SYSTEMS

This section analyzes the relation between PSDS_1 and energy consumption for ensemble and not-ensemble systems. We still focus on the 15 best systems in terms of PSDS_1.

Figure 7 and Figure 8 show the relation between PSDS_1 and energy consumption at training for the best ensemble-based submissions and for the top not-ensembled-based entries, respectively. What can emerge from this analysis is that an ensemble is useful at combining systems that alone are not so good in achieving decent performance (this is not so efficient in terms of energy but these not-so-good systems are already expensive) while a single system can provide a lighter alternative to reach good performance anyway. In fact, the best not-ensemble system is able to achieve a PSDS_1 score of 0.58, while the best ensemble system scores 0.62 for PSDS_1.

7. RELATION BETWEEN EW-PSDS AND PSDS

Figure 9 shows the relation between PSDS_1 and EW-PSDS. In particular, we can image the plot divided into four areas: bottom left corner, showing the system that were outperformed by the baseline and still had potentially higher consumption; bottom right corner: systems that outperform the baseline but consume more energy; top right: systems that outperform the baseline with limited energy consumption increase. The area to which we should aim is the top-right corner, which includes a system able to right high performance, not underestimating the environmental impact they are going to have. Unfortunately, as the figure shows, they are still minorities. This Figure highlights how most of the systems required a higher quantity of energy at training compared to the baseline. Some systems of the top right and the bottom right have pretty close PSDS_1 while having pretty different energy consumption. The systems able to outperform the baseline necessitating less energy at training are three different submissions of the team *Chen_CHT* [28].

8. THRESHOLDING BASED ON ENERGY CONSUMPTION

The last part of the analysis evaluates how much the performances degrade if a footprint cap is set. In order to do so, we define a threshold related to the system complexity, MACs, and energy consumption. For each threshold, we select only the systems that have a lower value of the threshold. For example, when considering the median of the energy consumption at training, we selected the systems with a lower energy consumption than the median consumption and reported the best PSDS_1 score. We applied a similar approach for MACs and system complexity. Due to space limitations, the same analysis for the 75th percentile is reported in the additional results, considering additional metrics [3]. Table 2 reports the degradation of the PSDS_1 for the systems that have been thresholded according to the different metrics. The left side reports the results for non-ensemble systems, while the right reports the results for ensemble systems. The PSDS performance remains rather stable regardless of the threshold cap while the complexity, MACs and energy consumption are substantially decreased. This is even clearer where considering ensembling. This indicates that we are spending a large amount of energy and computation to increase the performance only marginally.

9. CONCLUSIONS

This paper presents a comprehensive examination of energy usage and its correlation with various metrics applied to systems participating in the DCASE task 4 during the years 2022 and 2023. The findings make it evident that relying on a single metric is insufficient for accurately measuring a system's footprint. The paper also highlights that systems consuming the most energy (or having the most MACs) do not necessarily outperform less computationally expressive systems. These observations highlight the pressing need for metric(s) capable of taking into account various factors to accurately estimate the energy consumption of deep learning models while taking into account the task-wise performance of the systems. This would be the first important step to effectively design sustainable SED systems.

10. REFERENCES

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023.

[2] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[3] Ji Won Kim, Sang Won Son, Yoonah Song, Hong Kook Kim, Il Hoon Song, and Jeong Eun Lim, "Semi-supervsied learning-based sound event detection using freuqency dynamic convolution with large kernel attention for dcase challenge 2023 task 4," *arXiv preprint arXiv:2306.06461*, 2023.

[4] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al., "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[5] Antoine Caillon and Philippe Esling, "Rave: A variational autoencoder for fast and high-quality neural audio synthesis," *arXiv preprint arXiv:2111.05011*, 2021.

[6] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos, "Compute trends across three eras of machine learning," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022.

[7] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso, "The computational limits of deep learning," *arXiv preprint arXiv:2007.05558*, 2020.

[8] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu, "Chasing carbon: The elusive environmental footprint of computing," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021.

[9] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni, "Green AI," *Communications of the ACM*, 2020.

[10] Constance Douwes, Giovanni Bindi, Antoine Caillon, Philippe Esling, and Jean-Pierre Briot, "Is quality enough? integrating energy consumption in a large-scale evaluation of neural audio synthesis models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[11] Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat, "Evaluating the carbon footprint of nlp methods: a survey and analysis of existing tools," in *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, 2021.

[12] Romain Serizel, Samuele Cornell, and Nicolas Turpault, "Performance above all? energy consumption vs. performance, a study on sound event detection with heterogeneous data," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[13] Gianmarco Cerutti, Rahul Prasad, Alessio Brutti, Elisabetta Farella, et al., "Neural network distillation on IoT platforms for sound event detection.," in *Interspeech*, 2019.

[14] Titouan Parcollet and Mirco Ravanelli, "The Energy and Carbon Footprint of Training End-to-End Speech Recognizers," in *Interspeech*, 2021.

[15] Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," *arXiv preprint arXiv:2007.03051*, 2020.

[16] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[17] Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni, "Codecarbon: estimate and track carbon emissions from machine learning computing," *Cited on*, 2021.

[18] Ligeng Zhu, "Thop: Pytorch-opcounter," *Lyken17/pytorch-OpCounter*, 2019.

[19] Nicolas Turpault and Romain Serizel, "Training sound event detection on a heterogeneous dataset," *arXiv preprint arXiv:2007.03931*, 2020.

[20] Francesca Ronchini and Romain Serizel, "A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[21] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin 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*, 2019.

[22] Romain Serizel and Nicolas Turpault, "Sound event detection from partially annotated data: Trends and challenges," in *IcETRAN conference*, 2019.

[23] Janek Ebbers, Reinhold Haeb-Umbach, and Romain Serizel, "Post-processing independent evaluation of sound event detection systems," *arXiv preprint arXiv:2306.15440*, 2023.

[24] Francesca Ronchini, Samuele Cornell, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, and Daniel PW Ellis, "Description and analysis of novelties introduced in dcase task 4 2022 on the baseline system," *arXiv preprint arXiv:2210.07856*, 2022.

[25] Çağdaş Bilen, Giacomo Ferroni, and Francesco et al. Tuveri, "A framework for the robust evaluation of sound event detection," in *Proc. of ICASSP*, 2020.

[26] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[27] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[28] Wei-Yu Chen, Chung-Li Lu, Hsiang-Feng Chuang, Yu-Han Cheng Cheng, and Bo-Cheng Chan, "Sound event detection system using pre-trained model for dcase 2023 task 4," Tech. Rep., DCASE2023 Challenge, June 2023.