SALT: STANDARDIZED AUDIO EVENT LABEL TAXONOMY

Paraskevas Stamatiadis, Michel Olvera, Slim Essid

LTCI, Télécom Paris, Institut Polytechnique de Paris, France
{paraskevas.stamatiadis, olvera, slim.essid}@telecom-paris.fr

ABSTRACT

Machine listening systems often rely on fixed taxonomies to organize and label audio data, key for training and evaluating deep neural networks (DNNs) and other supervised algorithms. However, such taxonomies face significant constraints: they are composed of application-dependent predefined categories, which hinders the integration of new or varied sounds, and exhibits limited cross-dataset compatibility due to inconsistent labeling standards. To overcome these limitations, we introduce *SALT: Standardized Audio event Label Taxonomy*. Building upon the hierarchical structure of AudioSet's ontology, our taxonomy extends and standardizes labels across 24 publicly available environmental sound datasets, allowing the mapping of class labels from diverse datasets to a unified system. Our proposal comes with a new Python package designed for navigating and utilizing this taxonomy, easing cross-dataset label searching and hierarchical exploration. Notably, our package allows effortless data aggregation from diverse sources, hence easy experimentation with combined datasets.

Index Terms— Machine listening, DCASE, sound taxonomy, sound categorization, data aggregation

1. INTRODUCTION

Machine listening systems support a wide range of audio applications, including urban sound analysis [1, 2], industrial acoustic monitoring [3, 4, 5], music analysis [6, 7, 8], and speech recognition [9, 10, 11]. The key success of these systems, typically relying on supervised machine learning approaches, especially using deep neural networks (DNNs), lies in the systematic annotation of training data, using predefined class labels from hierarchical sound ontologies and taxonomies [12, 13, 14, 15, 16, 17, 18].

Sound ontologies and taxonomies serve as foundational frameworks to categorize everyday sound scenes and events [19]. Developed across several research fields—for instance auditory cognition [20, 21], soundscape research [22], sound design [23]—they are particularly instrumental in machine listening. Notable examples stand out for their wide adoption in the DCASE[1] community: UrbanSound8K [12], SONYC-UST [15, 24], MAVD-traffic [16] for urban sound analysis and ESC-50 [25] and AudioSet [15] for broader sound event recognition.

As evident from the previous examples, categorization of sounds in these systems are context-specific and tailored to desired applications. Their static nature often entails significant, overhauls for updates or extensions, especially when combining audio events from different environments, even for the same application. An exemplary case of this, is the adaptation of the SONYC-UST's taxonomy. Originally developed to classify urban sounds in New York City, this taxonomy required an expansion to accommodate the unique sounds of Singapore city's soundscapes, while maintaining compatibility with the base categorization. This adaptation allowed for bench-marking of urban sound tagging systems across different cities [17, 26].

While recent initiatives such as *mirdata* [27] and *Soundata* [28] have simplified the use of major datasets for Music Information Research (MIR) and DCASE, by standardizing data loading, these efforts primarily focus on addressing issues related to data management and accessibility through open-source software packages. As such, these packages promote reproducibility and flexible data-processing pipelines. However, the development of adaptable and extensive sound categorization frameworks capable of integrating new audio event labels from diverse datasets while maintaining compatibility with existing taxonomies remains largely unaddressed.

In this work, we tackle such challenges by introducing *SALT: a Standardized Audio event Label Taxonomy*. Leveraging the hierarchical structure of AudioSet, *SALT* extends and standardizes labels across 24 publicly available environmental sound datasets. Such a large collection of datasets covers diverse audio analysis tasks including audio tagging, sound event detection and acoustic scene classification. By standardizing labels, SALT enables mapping them across diverse datasets, ensuring compatibility and easing dataset aggregation. Alongside our proposed taxonomy of standard dataset labels, we present *py-salt*, an open-source python package designed to navigate through its content. This tool allows users to easily navigate through the hierarchical label taxonomy at any level of granularity. It turns out to be quite valuable when performing experiments considering various existing datasets whose particular labelling schemes can be seamlessly represented in our unified taxonomy.

We posit that our contribution is timely in a research context where large-scale training of audio models is fueled by the availability of (labelled) training data and computational resources. Our taxonomy with standardized audio event labels simplifies data aggregation, complementing tools like *Soundata* to develop audio classification models at scale.

The remainder of this work is organized as follows: Section 2, introduces the motivation and design principles behind SALT. Section 3 presents the functionalities and applications of py-salt, our proposed Python package, and Section 4 concludes the article.

2. SALT

The motivation behind creating SALT is the development of a new solution leveraging existing taxonomies to facilitate experimentation across different environmental sound datasets. The key feature of this solution is label aggregation, which allows unified categorization of sound events. This approach necessitates a standardized

[1] Detection and Classification of Acoustic Scenes and Events

set of labels applicable to multiple environmental sound datasets. Consequently, with SALT we aim to expand AudioSet, the largest general-purpose sound event taxonomy, and use it as a common frame of reference to represent the annotations of all major publicly available DCASE datasets.

2.1. Design Principles

We aim to establish a general-purpose sound taxonomy with label aggregation capabilities at the core of its design. To achieve this, we adapt existing sound event taxonomies from diverse domains, including but not limited to urban sound analysis, acoustic scene classification, domestic sound event detection, among others, using AudioSet's taxonomy as our basis. A key principle is to integrate labels from diverse sound collections, prioritizing datasets that are independent from each other rather than subsets of others, leading to a natural expansion of AudioSet's taxonomy. This integration into a unified taxonomy entails a standardization process to ensure label consistency across datasets.

Label standardization. The standardization process involves a mapping of original (*i.e.*, default) category names from different datasets that describe the same acoustic event to a standardized label. For example, labels such as "car horn" in *UrbanSound8K*, "car_horn" in *ESC-50* and "Vehicle horn, car horn, honking" in *AudioSet*, all refer to the sound produced by a car horn. To aggregate labels effectively, we map them to the standard label *car_horn* in SALT. Our notation for denoting standard labels uses lowercase characters and underscores instead of white spaces.

Mapping for accurate aggregation. In cases where a dataset label indicates more than one acoustic event, or sources producing sound, the mapping depends on the nature of the sounds. If the events or sources have similar acoustic properties, the word "or" is introduced in the standard label to preserve both sounds in the label. For example, the label "Railroad car, train wagon" in AudioSet is mapped to the standard label *railroad_car_or_train_wagon* as both sources produce the same type of sound. On the contrary, when a dataset label indicates multiple sound events, each of them entailing unique acoustic signatures, the mapping selects the most specific (*i.e.,*finest-grained) standard label that avoids incorrect associations in the aggregation process. For example, the label "dog-barking-whining" in *SONYC-UST* is mapped to the broader standard label *dog* to ensure accurate aggregation. This principle prevents mistakenly including unrelated events into more specific standard labels such as *dog_barking* or *dog_whining*. In Figure 1 we present a clear depiction of our label standardization procedure.

Hierarchy expansion. Additionally, our objective is to preserve the base hierarchy of AudioSet while integrating new standard labels when strictly necessary. This design principle serves two main purposes. First, it facilitates label aggregation across multiple hierarchical levels by mapping dataset labels not only to a standard label, but also to its hierarchical ancestors (also standardized labels). For example, the label "Bird" in AudioSet is mapped to the standard label *bird* in our taxonomy, as well as to its standard ancestors *wild_animal* and *animal*. Second, it refines the AudioSet taxonomy by incorporating new or rare sound event labels coming from a wide variety of environmental sound datasets serving different audio analysis tasks. When a dataset contains class labels which do not fit neatly into the AudioSet taxonomy

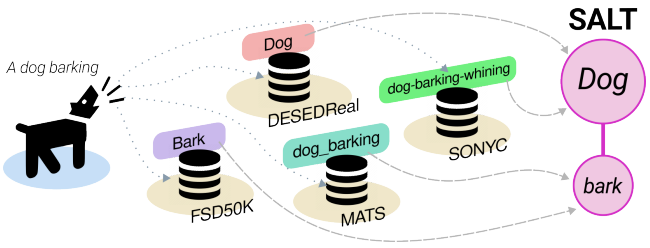

Figure 1: Illustration of SALT's standardization process. Dataset labels are systematically mapped to a standard label that ensures cross-dataset compatibility.

or cannot be covered by any existing node in the structure, new standardized labels are introduced to accomodate such labels. For instance, labels such as "truck/compressor" from the *MAVD-traffic* and "Friction brake" from the *SINGA:PURA* datasets, represent cases where AudioSet's existing labels are insufficient to fully capture the diverse set of sounds encountered in publicly available datasets.

2.2. Taxonomy Structure

Our proposed extension to Audioset's taxonomy, is structured into multiple hierarchical levels, each representing a different granularity of sound categories. Starting from AudioSet's seven broad sound categories — "Human sounds", "Animal", "Music", "Source-ambiguous sounds", "Sounds of things", "Natural sounds", and "Channel, environment and background" — and 616 sound labels (out of the 632 provided in the taxonomy), we expand to 734 sound labels. These labels are categorized under the original seven AudioSet categories, with the addition of two new categories: *Water* and *Other*. Figure 2 illustrates the contribution of the original labels from all considered datasets to compose the standard labels in SALT.

With careful examination of the video clips available in AudioSet and their associated labels, we refined the hierarchical structure of AudioSet to clarify the placement of labels within the taxonomy. For example, when examining the label "Water", we found that 37% of clips include tags related to water sounds occurring in domestic environments *e.g., "Water faucet"*, while, only 2% of them are related to outdoor and/or natural landscapes. This distribution indicates that the "Water" tag does not exclusively belong under "Natural sounds", but also frequently appears in domestic settings. Additionally, we conducted refinements by examining children within categories such as "Vehicle" and "Engine". For example, clips tagged with the label "Accelerating, revving, vroom", categorized under "Engine", primarily pertain to vehicle sounds, accounting for approximately 93% of its instances. Therefore, "Accelerating, revving, vroom" is additionally categorized under "Vehicle". For a complete list of all such refinements, we refer the reader to our companion repository[2].

3. PY-SALT

In this section, we give an overview of the functionalities and applications of SALT, designed to unify event labels through standard

[2] https://github.com/tpt-adasp/salt

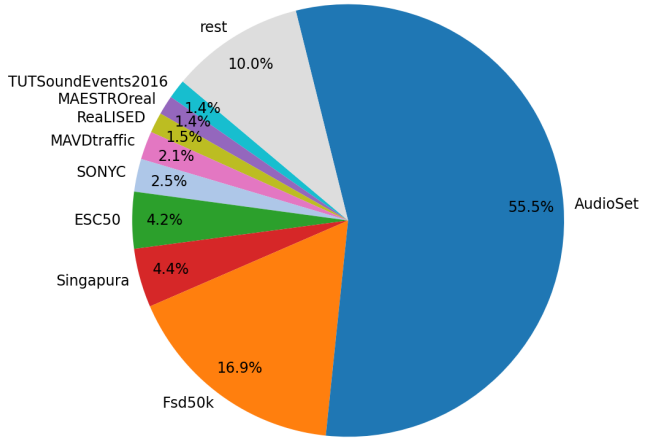

Figure 2: Contribution of dataset's original (default) labels to SALT after the standardization process.

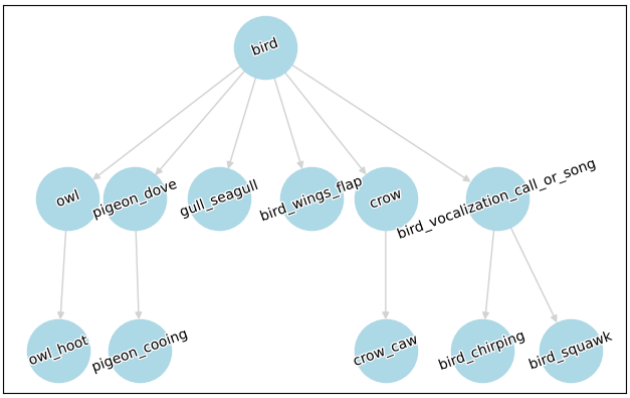

Figure 3: Example of standard label mapping for the standardized label *bird*.

labels, allowing for label search, data exploration, and hierarchical parsing. To exploit the benefits of our proposed taxonomy with standard labels, we developed *py-salt*, a Python package that provides tools for navigating and utilizing the taxonomy.

3.1. Functionalities

Label searching. This functionality allows two searching modes. First, standard labels can be employed to look for corresponding original (default) dataset labels across all those integrated in SALT, *e.g., motorcycle* → "motorbike" coming from *Urbansas*, "Motorcycle" from *AudioSet/FSD50K*, "motorcycle/wheel_rolling", "motorcycle/engine_idling", "motorcycle/engine_accelerating" from *MAVD-traffic*, etc. Secondly, original dataset labels can be employed to identify their counterparts across all datasets within the taxonomy. This dual approach offers comprehensive coverage and consistency for cross-dataset retrieval, *e.g., ReaLISED's* "water tap" → "Water_tap_and_faucet", "Water tap, faucet", "water tap running" coming from *FSD50K*, *AudioSet* and *TUT Sound Events 2016*, respectively.

Hierarchical exploration and expansion. This functionality allows browsing the taxonomy at any level of the hierarchy and easily locate superodinate (parent), subordinate (child) and coordinate (sibling) categories. Additionally, SALT supports *mapping expansion*, a functionality useful to incorporate new datasets and label categories into the taxonomy. The mapping process can be performed using the existing standard labels or by defining new ones suiting the user's requirements.

Visualization and searching tools. Graph plotting utilities are included in `py-salt`, which allows users to explore SALT visually. The python library contains methods to plot graphs showing the hierarchical structure of a given standard label in SALT, and also to depict all original (default) class names in the aggregated datasets that mapped to a SALT label. For example, the function `plot_hierarchical_tree_graph('bird')` serves to generate a graphical representation of the hierarchical structure for the standard label *bird* as illustrated in Figure 3. This functionality provides a clearer depiction of the relationships between parent, child and sibling categories. Similarly, the function `plot_std_label_mapping('car_horn')` serves to show the mapping of dataset labels to the standard label *car_horn* as illustrated in Figure 4. This is useful for identifying the potential datasets needed for a specific application of interest.

An additional example, is illustrated in Listing 1, which involves retrieving the dataset labels mapped to the standard label *reverse_beeper*. The function returns a Python dictionary where dataset names serve as keys and their corresponding labels as values. This example highlights the package's feature to provide detailed and well-organized information about class labels, which is essential for analysis and integration of data.

Furthermore, the package comes with extensive documentation including a tutorial notebook and practical examples to demonstrate all functionalities discussed in this section. The interested reader is referred to the corresponding repository for more information about `py-salt`.

3.2. Applications

SALT can serve diverse applications and use cases through its functionalities. The provided python library, facilitates exploration of mapped datasets, both individually and in combination. An interesting use case comprises the compilation of data from multiple datasets to compose new datasets or collections. This is achieved by the use of a series of methods provided in `py-salt`, that allows gathering the desired labels from specific datasets or domains of interest. For example, to develop an audio classifier specialized in the detection of emergency signals, all relevant labels from different datasets *e.g., AudioSet, SINGA:PURA, ESC-50, etc.* can be easily accessed through the standard label *alarm_signal*. Similarly, to develop an urban sound monitoring system, various dataset labels can be aggregated through a set of standard labels such as *vehicle*, *engine*, and *outdoor_urban_or_manmade* from datasets *MAVD-traffic, AudioSet, FSD50K* and *SONYC-UST*, respectively. To give another example, for a system targeting the detection of domestic sound events, labels such as *kitchen* (*i.e. kitchen sounds*), *bell* and *television* can be aggregated to create a specialized classifier for recognizing common household sounds. Figure 5 illustrates the significant benefits of label aggregation in augmenting the amount of data

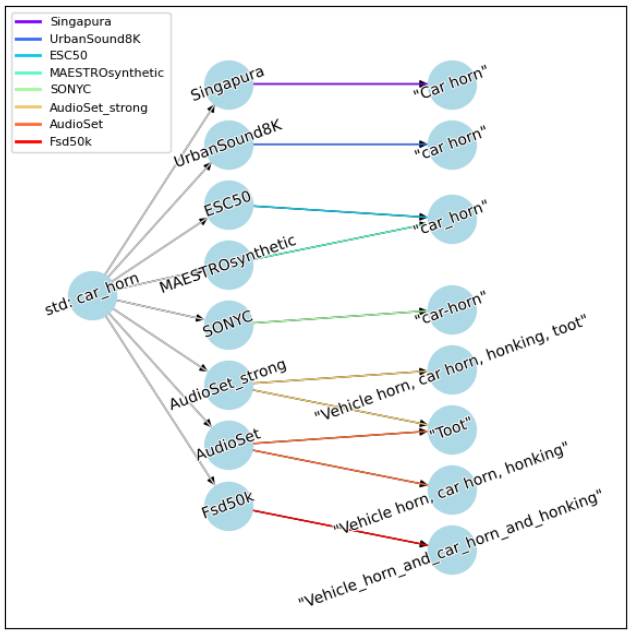

Figure 4: Example of dataset label mapping for the standardized label *car_horn*.

available for minority classes.

Another use case involves defining a common set of standard labels for cross-dataset evaluation purposes, *e.g.*, training on *SONYC* and testing on the same set of labels on *UrbanSound8K*. This approach is particularly useful for bench-marking audio analysis systems and assess their generalization capabilities. Overall, SALT can diminish inconsistencies and discrepancies between different datasets and promotes fair comparison of model performance.

```python
from py_salt.event_mapping import EventExplorer()

# Init taxonomy explorer
e = EventExplorer()

# Get dataset mapping dictionary
e.get_mapping_for_std_label('reverse_beeper')

{'SONYC': ['reverse-beeper'],
 'Singapura': ['Reverse beeper'],
 'AudioSet_strong': ['Reversing beeps'],
 'AudioSet': ['Reversing beeps']}
```

Listing 1: Label search using the standardized label *reverse_beeper*

4. CONCLUSION

In this paper, we introduced *SALT: Standardized Audio event Label Taxonomy* to unify existing sound taxonomies into a global one through the standardization of labels, while also addressing some of their limitations. Built upon AudioSet's hierarchical structure, SALT standardizes and extends labels across 24 environmental sound datasets, enhancing clarity and precision and enabling cross-dataset label compatibility. Furthermore, we support the use of SALT, by introducing a Python package that provides robust

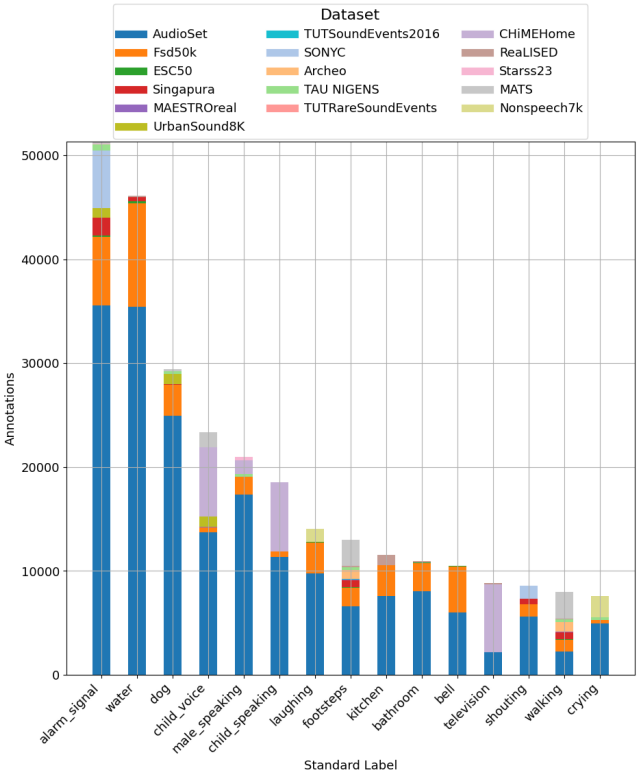

Figure 5: The benefit of label aggregation in selected standardized labels targeting domestic sound events.

tools to perform cross-dataset label aggregation, explore hierarchical relationships and visualize label mappings. These capabilities streamlining data aggregation and analysis, make SALT a valuable resource for developing machine listening systems at scale.

5. ACKNOWLEDGMENT

This work was supported by the Audible project, funded by French BPI.

6. REFERENCES

[1] E. Vidaña-Vila, J. Navarro, D. Stowell, and R. M. Alsina-Pagès, "Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors," *Sensors*, vol. 21, no. 22, p. 7470, 2021.

[2] F. Angulo, S. Essid, G. Peeters, and C. Mietlicki, "Cosmopolite sound monitoring (cosmo): A study of urban sound event detection systems generalizing to multiple cities," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[3] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. ICASSP*. IEEE, 2020, pp. 271–275.

[4] G. Wichern, A. Chakrabarty, Z.-Q. Wang, and J. Le Roux, "Anomalous sound detection using attentive neural processes," in *Proc. WASPAA*. IEEE, 2021, pp. 186–190.

[5] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. ICASSP*. IEEE, 2022, pp. 816–820.

[6] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. ICASSP*. IEEE, 2008, pp. 169–172.

[7] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie, "An end-to-end machine learning system for harmonic analysis of music," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1771–1783, 2012.

[8] S. Oramas, F. Barbieri, O. Nieto Caballero, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[9] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*. IEEE, 2018, pp. 4904–4908.

[10] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2letter++: A fast open-source speech recognition system," in *Proc. ICASSP*. IEEE, 2019, pp. 6460–6464.

[11] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *Proc. ICASSP*. IEEE, 2015, pp. 4295–4299.

[12] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Multimedia*, 2014, pp. 1041–1044.

[13] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*. IEEE, 2017, pp. 776–780.

[14] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2021.

[15] M. Cartwright, A. E. M. Mendez, A. Cramer, V. Lostanlen, G. Dove, H.-H. Wu, J. Salamon, O. Nov, and J. Bello, "Sonyc urban sound tagging (sonyc-ust): A multilabel dataset from an urban acoustic sensor network," 2019.

[16] P. Zinemanas, P. Cancela, and M. Rocamora, "Mavd: a dataset for sound event detection in urban environments," 2019.

[17] K. Ooi, K. N. Watcharasupat, S. Peksi, F. A. Karnapi, Z.-T. Ong, D. Chua, H.-W. Leow, L.-L. Kwok, X.-L. Ng, Z.-A. Loh, *et al.*, "A strongly-labelled polyphonic dataset of urban sounds with spatiotemporal context," in *Proc. APSIPA ASC*. IEEE, 2021, pp. 982–988.

[18] 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. DCASE*, 2019.

[19] C. Guastavino, "Everyday sound categorization," *Computational analysis of sound scenes and events*, pp. 183–213, 2018.

[20] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[21] C. Guastavino, "Categorization of environmental sounds." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 61, no. 1, p. 54, 2007.

[22] R. M. Schafer, *Our sonic environment and the soundscape: The tuning of the world*. Destiny Books, 1994.

[23] D. Moffat, D. Ronan, J. D. Reiss, *et al.*, "Unsupervised taxonomy of sound effects," *context*, vol. 6, no. 7, 2017.

[24] M. B. Cartwright, J. Cramer, A. E. M. Méndez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, O. Nov, and J. P. Bello, "Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context," in *Proc. DCASE*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221641055

[25] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proc. ACM Multimedia*, 2015, pp. 1015–1018.

[26] F. Angulo, S. Essid, G. Peeters, and C. Mietlicki, "Cosmopolite sound monitoring (cosmo): A study of urban sound event detection systems generalizing to multiple cities," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[27] R. M. Bittner, M. Fuentes, D. Rubinstein, A. Jansson, K. Choi, and T. Kell, "mirdata: Software for reproducible usage of datasets." in *ISMIR*, 2019, pp. 99–106.

[28] M. Fuentes, J. Salamon, P. Zinemanas, M. Rocamora, G. Paja, I. R. Román, M. Miron, X. Serra, and J. P. Bello, "Soundata: A python library for reproducible use of audio datasets," *arXiv preprint arXiv:2109.12690*, 2021.