A Strongly-Labelled Polyphonic Dataset of Urban Sounds with Spatiotemporal Context

Kenneth Ooi, Karn N. Watcharasupat, Santi Peksi, Furi Andi Karnapi, Zhen-Ting Ong,
Danny Chua, Hui-Wen Leow, Li-Long Kwok, Xin-Lei Ng, Zhen-Ann Loh, and Woon-Seng Gan

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Emails: {wooi002, karn001, speksi, furi, ztong, jchua056, leow0121, e180115, xng033, zloh011, ewsgan}@ntu.edu.sg

Abstract—This paper introduces SINGA:PURA, a strongly labelled polyphonic urban sound dataset with spatiotemporal context. The data were collected via several recording units deployed across Singapore as a part of a wireless acoustic sensor network. These recordings were made as part of a project to identify and mitigate noise sources in Singapore, but also possess a wider applicability to sound event detection, classification, and localization. This paper introduces an accompanying hierarchical label taxonomy, which has been designed to be compatible with other existing datasets for urban sound tagging while also able to capture sound events unique to the Singaporean context. This paper details the data collection, annotation, and processing methodologies for the creation of the dataset. We further perform exploratory data analysis and include the performance of a baseline model on the dataset as a benchmark.

I. INTRODUCTION

The collection of high-quality data for domain-specific problems is a major challenge in deep learning, and particularly so for tasks in the audio domain. Publicly available audio datasets captured in real-world contexts are few and far in between, owing to the high monetary and technical requirements needed to embark on audio data acquisition projects. As a result, synthetic data generation is often employed using software libraries such as Scaper [1] and Pyroomacoustics [2] in order to meet the training requirements of deep learning systems. Although these software libraries have been extremely helpful in the experimentation and development of many audio-based deep learning systems, synthetic data can never fully replicate the nuances of sounds recorded in real acoustic environments. The resulting covariate shift is fatal in many deep learning systems, thus leading to well-known problems of performance degradation when systems trained on synthetic data are deployed in the real world. Beyond generic sound event detection or classification system, urban sound tagging systems are amongst the applications in which real data are of extreme importance.

Audio data have unique distributions which are highly dependent not only on the sound sources themselves, but also on factors including the sonic environments and recording equipment. Moreover, when the sound itself is dependent on human activities, as is the case in urban sounds, further context (like the location and time of the recording) is often useful for both humans and machines to fully understand the sonic environment [3]. Unsurprisingly, the complex and layered nature of urban sounds means that synthetic samples are unable to perfectly replicate the real-world distribution of the frequency and timing of a given collection of urban sound sources. Hence, synthetic training samples are inadequate as training data for neural urban sound tagging systems. To develop systems which are able to cope with the complexity of the real world — such as those in wireless acoustic networks deployed in Singapore [4], Germany [5, 6], Spain [7], and the United States [8] — training with real-world data is vital.

At the time of writing, there exists no publicly available, strongly labelled urban sound dataset consisting of recordings in real-world contexts. This paper introduces the SINGA:PURA (*Singapore: Polyphonic Urban Audio*) dataset[1], which is an urban sound dataset of 6547 strongly labelled recordings made in real acoustic environments with the corresponding spatiotemporal metadata attached to each recording. These recordings were made by single-channel sensors and multi-channel sensor arrays deployed across the island of Singapore via a wireless acoustic sensor network described in [4]. In addition to urban sound tagging, this dataset would provide new grounds for applications of more complex tasks, such as sound event localization and detection, in urban settings.

II. EXISTING URBAN SOUND DATASETS

Although there are a number of datasets for generic sound classification [9–13] and synthetic datasets for urban sound classification [14, 15], there are currently only two main non-synthetic datasets of urban sound recordings that are publicly available. Specifically, they are the UrbanSound datasets [16], and the SONYC Urban Sound Tagging datasets [3, 17]. We briefly discuss these datasets, as well as the related FSD50k dataset, in this section.

A. Freesound FSD50k Dataset

Although not an urban sound dataset by design, FSD50k [10] is to date one of the most comprehensive publicly available audio datasets. The dataset consists of over 100 hours

---

[1]The dataset and detailed documentation are publicly available at https://doi.org/10.21979/N9/Y8UQ6F. In addition to the strongly labelled recordings, we have also included 72406 unlabelled recordings as part of the dataset.

of sounds drawn from Freesound, which is a database of audio clips licensed under various Creative Commons licenses. The recordings were annotated based on the AudioSet ontology [9]. Many smaller and task-specific datasets have been derived from Freesound, including the UrbanSound datasets [16].

However, since the audio data were crowdsourced and users on the Freesound platform are free to enter anything (or nothing) to describe their recordings, there is no proper documentation on the recording conditions, devices, and methods for these data. Moreover, the audio formats and number of channels vary across the entire dataset. Hence, these limitations apply to all datasets derived from FSD50k, such as the UrbanSound datasets, as well.

B. UrbanSound and UrbanSound8k Datasets

UrbanSound and UrbanSound8k are amongst the first datasets designed specifically for urban sound tagging [16]. They were curated using recordings from Freesound [10]. Both variants of UrbanSound consist of 10 coarse-level sound classes, namely air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. The main UrbanSound dataset consists of 1302 potentially polyphonic recordings with only events from one class labelled strongly in each recording. UrbanSound8k contains 8732 labeled monophonic sound excerpts.

C. SONYC Urban Sound Tagging Datasets

The SONYC Urban Sound Tagging (SONYC-UST) dataset is a weakly labelled audio dataset of urban sound recordings. The audio was recorded by an urban acoustic sensor network under the Sounds of New York City (SONYC) project [8]. Early versions of the dataset only included the audio data [17]. Starting from version 2, the dataset also includes coarse spatiotemporal context (STC) of the location and time that each recording was made [3].

The IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) featured urban sound tagging as Task 5 in the 2019 and 2020 editions. The 2019 challenge utilized an early version (v0.2.0 and v0.3.0) of the dataset without spatiotemporal context. Recognizing the importance of spatiotemporal context in urban sound tagging, the 2020 challenge featured the second version (v2.2.0 and v2.3.0) of the dataset. Top-ranked systems from the 2020 challenge [18, 19] demonstrated the efficacy of incorporating spatiotemporal features into deep learning systems for the classification of audio events whose distributions are context-dependent, such as that in urban sound tagging.

III. THE SINGA:PURA DATASET

This section will describe the data acquisition, annotation, and processing methodologies for the creation of the SINGA:PURA dataset.

A. Data Acquisition

Our data collection effort utilizes the wireless acoustic sensor network first introduced and detailed in [4]. Each sensor

Fig. 1. Sensor units used to make recordings for the SINGA:PURA dataset. *Left*: A unit with a single-channel SPH0645 microphone. *Right*: A unit with a seven-channel UMA-8 v2 microphone array.

TABLE I
DATA COLLECTION UNIT DEPLOYMENT

Neighborhood	Microphone	No. of Units
East 1	SPH0645	2
East 2	SPH0645	3
West 1	UMA-8 v2	4
West 2	SPH0645	5

unit is located in a public location, and houses either a single-channel Knowles SPH0645 microphone or a seven-channel MiniDSP UMA-8 v2 microphone array. The microphone in each sensor unit is controlled by a Raspberry Pi 3 Model B, which doubles a data collection unit for spatiotemporal information. All was recorded at a 44.1 kHz sampling rate. Pictures of the sensor units in their actual deployment contexts can be seen in Fig. 1.

The recordings in the SINGA:PURA dataset are collected from 14 individual sensors deployed across four neighborhoods in Singapore, and span the duration between 3 August 2020 and 31 October 2020. The four neighborhoods consisted of two in the eastern region of Singapore and two in the western region of Singapore. The breakdown of the deployment by neighborhood, as well as the type of microphone is shown in Table I.

B. Label Taxonomy

The label taxonomy was derived from the taxonomy used in the SONYC-UST datasets [3, 17], but has been adapted to fit the local context while retaining compatibility with the SONYC-UST ontology. Hence, this allows the SINGA:PURA dataset to be used in conjunction with the SONYC-UST datasets when training urban sound tagging models by simply omitting the labels that are absent in the SONYC-UST taxonomy from the recordings in the SINGA:PURA dataset.

Our adapted taxonomy retains the original 8 coarse classes from SONYC-UST and introduces 6 new coarse classes, each possibly subsuming new fine classes as well. In addition, we have included a catch-all "Others" class to capture sound events that exist in the SINGA:PURA dataset, but do not fit well into any of the coarse class labels. Our taxonomy also
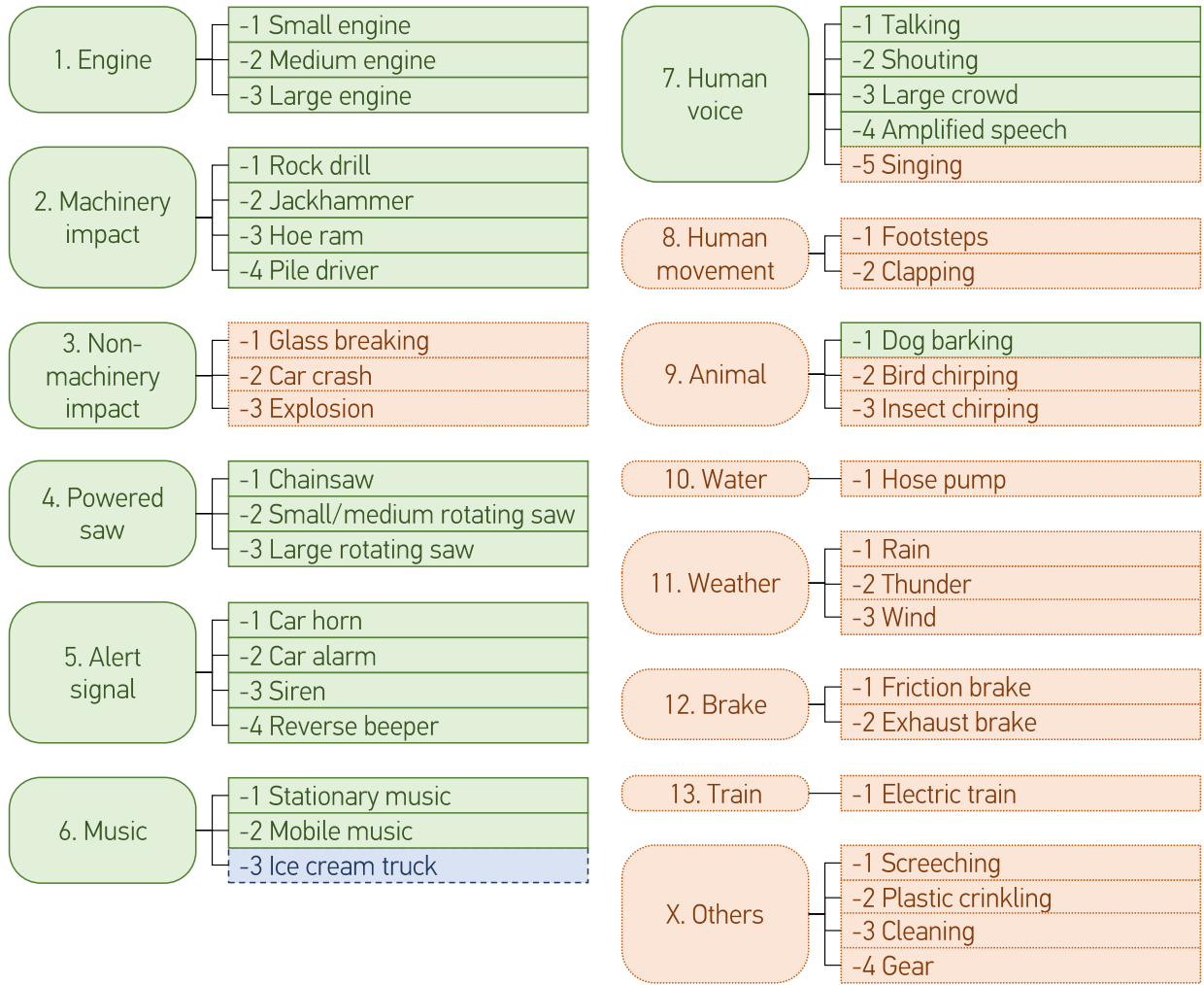
Fig. 2. Label taxonomy for the SINGA:PURA dataset. Green boxes with solid outlines are classes common to both SINGA:PURA and SONYC-UST ontology. Red boxes with dotted outlines are classes newly introduced in SINGA:PURA ontology. The blue box with dashed outline indicates the class included in SONYC-UST ontology but excluded from SINGA:PURA ontology

preserves all the fine classes in the SONYC-UST ontonlogy except for the "ice-cream truck" class, since ice cream trucks do not exist in Singapore. In summary, our label taxonomy consists of 14 coarse classes and 40 fine classes, as illustrated in Fig. 2.

C. Annotation

The recorded audio data were split into 10-second chunks, before being individually annotated by a team of five annotators. Recordings with audible sensor faults were firstly excluded from the published dataset before being sent to the annotators to undergo the annotation process. The annotation process included identification of sound events, their fine class labels, their onset times, their offset times, and their estimated proximity from the sensor, thus resulting in polyphonic strong labels of all identifiable sound events in each recording. Each 10-second recording was annotated by at least one annotator, but some recordings were independently annotated by multiple annotators. In the case of recordings where the class labels of the sound events were ambiguous sound events, all annotators listened and a consensus decision was made to determine the sound classes, onsets, and offsets of the sound events.

All listening was done via Audio Technica ATH-M50x headphones, with playback levels matching that of white noise at an A-weighted equivalent sound pressure level of 65 dB. Calibration to the playback levels was performed on a GRAS 45BB KEMAR Head and Torso Simulator using the automatic calibration method described in [20].

D. Dataset structure

The audio recordings are provided in the Free Lossless Audio Codec (FLAC) format at the original sampling rate of 44.1 kHz and are organized into folders according to the dates when the recordings were made.

The sound event labels are provided as a comma-separated values (CSV) file. The CSV file is organized in a similar manner to the label files used in DCASE sound event detection tasks, such as that used for DCASE 2021 Task 4 [21]. Each

row in the label file represents exactly one sound event. The columns indicate the name of the recording, sound class, proximity, onset, and offset. A recording may contain multiple sound events and these are represented by multiple rows in the label file accordingly.

The spatiotemporal metadata is also provided as a separate CSV file, with each row containing the spatiotemporal context for each recording. Specifically, the columns in this CSV file indicate the sensor unit identifier, the date and time of recording, the day of the week, and the neighborhood the sensor was deployed in.

Table II shows a comparison between the SINGA:PURA dataset and other existing non-synthetic datasets.

IV. EXPLORATORY DATA ANALYSIS

In total, 6547 recordings of 10 seconds were annotated, totaling to 18.2 hours of audio data. Of these, 38 recordings were found to have no sound events in the taxonomy illustrated in Fig. 2. All fine-level class labels have at least one example in the other 6509 recordings in the dataset, *except* for the "car crash" class (class 3-2). Hence, only 39 of the 40 fine-level sound classes in Fig. 2 will be represented in the current version of the SINGA:PURA dataset.

The distribution of the individual sound events by class is shown in Fig. 3. Sound events are considered as continuous intervals of time, so it is possible for a single recording to contain two sound events of the same class, even if the time intervals when they are active are overlapping. We can see that the "non-machinery impact" class (class 3) has the highest frequency of occurrence, followed by the "animal" class (class 9). The high occurrence of the "animal" class is largely due to the presence of cicadas and crickets chirping in a large number of recordings, since they are very commonly found in Singapore.

In contrast, the distribution of the number of recordings with at least one sound event by coarse-level class is shown in Fig. 4. We can see that certain sound classes, such as the "engine" class (class 1) and the "non-machinery impact" class (class 3), appear in many recordings. Since a given recording may have multiple sound events, the total number shown in all the bars exceeds the total number of recordings (6547) in the SINGA:PURA dataset.

The distribution of the total active duration of sound events by coarse-level class across the entire dataset is shown in Fig. 5. We can see that the distribution is unbalanced across the coarse classes and follows an inverse power law, which is typical of real recordings in urban environments. In a similar manner as Fig. 4, the total duration represented by all the bars in Fig. 5 is longer than that of all the recordings in the SINGA:PURA dataset because each recording may have multiple sound events that overlap.

Fig. 6 shows the average active duration of sound events by coarse-level class. As expected, sound classes that are characterized by stationary or ambient sounds, such as the "engine" class (due to the presence of traffic) and the "weather" class (due to the presence of rain), naturally have longer average
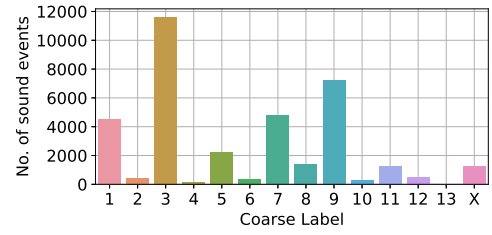

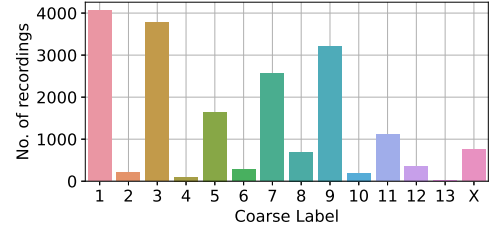Fig. 3. Distribution of the number of disjoint sound events by coarse-level class

Fig. 4. Distribution of the number of recordings with at least one sound event by coarse-level class
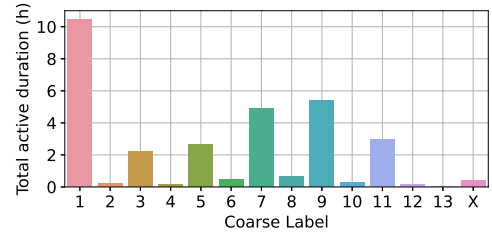

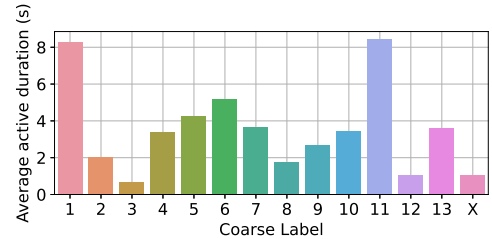Fig. 5. Distribution of the total duration of active sound events by class

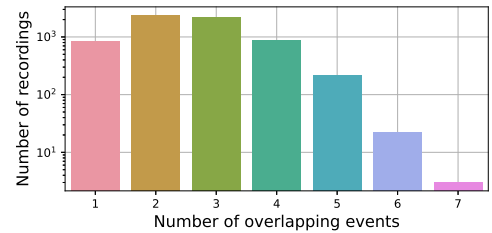Fig. 6. Average sound event duration by coarse-level class

Fig. 7. Distribution of the maximum number of overlapping sound events at any point in time in a recording.

durations compared to other classes. On the other hand, sound classes characterized by impulsive sounds, such as the "non-machinery impact" class (class 3) and the "brake" class (class 12) naturally have shorter average durations compared to other classes.

Fig. 7 shows the distribution of the maximum number of overlapping sound events at any point in time in a recording. The most common degree of polyphony in the dataset is two, while the maximum degree of polyphony is seven. Beyond a degree of polyphony of three, the distribution shows an

| Dataset | Location | No. of classes | | STC | No. of Channels | Label Type | Duration (hours) |
		Coarse	Fine				
SINGA:PURA	Singapore	14	40	●	1 or 7	Strong polyphonic	18.2
SONYC-UST v2.3 [3]	New York City	8	23	●	1	Weak polyphonic	51.4
SONYC-UST v1.0 [17]	New York City	8	23	×	1	Weak polyphonic	8.5
UrbanSound [16]	Varies, unknown	10	×	×	varies	Strong monophonic	18.5
UrbanSound8k [16]	Varies, unknown	10	×	×	varies	Weak monophonic	9.7
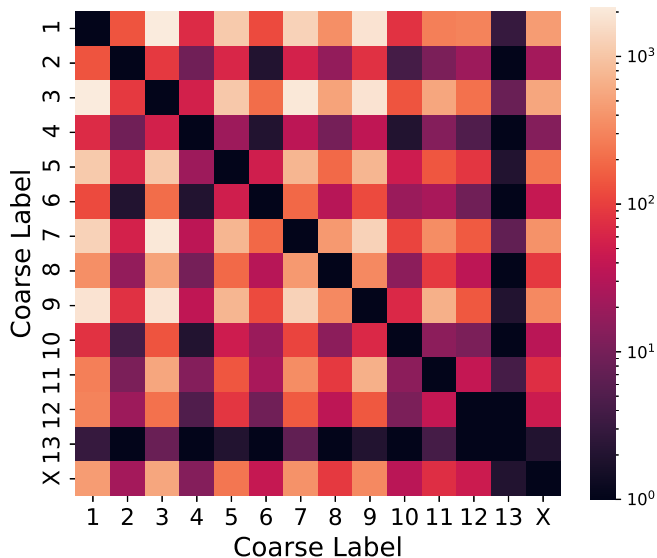
Fig. 8. Coarse-level co-occurrence matrix computed at recording level. Note that the color bar is in the logarithmic scale.

Metrics	SINGA:PURA	SONYC-UST
Micro AUPRC	0.7064	0.8352
Micro F1	0.4811	0.6736
Macro AUPRC	0.5825	0.6370
Class-wise AUPRC		
1. Engine	0.8706	0.8500
2. Machinery impact	0.2284	0.6021
3. Non-machinery impact	0.8553	0.4192
4. Powered saw	0.0220	0.7200
5. Alert signal	0.7679	0.8518
6. Music	0.4890	0.6145
7. Human voice	0.8708	0.9593
8. Human movement	0.5354	
9. Animal/Dog*	0.7906	0.0463
10. Water	0.1682	
11. Weather	0.7698	
12. Brake	0.1649	
13. Train	0.0000	
X. Others	0.2029	

* The SONYC-UST taxonomy does not have an *animal* coarse-level class but has a coarse-level *dog* class. In the SINGA:PURA taxonomy, the *dog* class is a fine-level class under the *animal* class. We include the result from the *animal* class for SINGA:PURA and from the *dog* class for SONYC-UST.

exponential decrease, as can be seen form the logarithmic scale used in the y-axis of Fig. 7.

Lastly, Fig. 8 shows the co-occurrence matrix of the coarse-level sound classes at the recording level. It can be seen that certain classes of sound events tend to occur together in a recording (e.g. engine, non-machinery impact, alert signal, human voice, and animal), while some classes of sound events rarely or never occur together in the same recording (e.g. water with either machinery impact or powered saw).

V. BASELINE MODEL

As a baseline model, we adapted the baseline model from Task 5 (Urban Sound Tagging with Spatiotemporal Context) of the DCASE 2020 Challenge [22] for the SINGA:PURA dataset. The adapted system and training process are detailed in [23] and we replicate the results here for ease of reference. The system in [23] is identical to the model in [22] except for the prediction layer which was changed to accommodate the additional sound classes in the SINGA:PURA dataset. The model was initially trained on the SONYC-UST v2.2 dataset [3] with coarse labels. The original output layer was then detached and replaced with a 14-class output layer. The model was then fine-tuned for coarse-level prediction on the

SINGA:PURA dataset. The overall and class-wise performances of the model are shown in Table III. For the summary of results, we report the micro- and marco-aggregated areas under the precision-recall curve (AUPRC) as well as a micro-aggregated F1 score. For the class-wise results, we report the AUPRC. Although the results cannot (and should not) be directly compared in a conventional sense, we also include the performance of the coarse-level model trained on the SONYC-UST v2.2 dataset from [22] as a rough reference.

VI. CONCLUSIONS

This paper introduces a non-synthetic polyphonic urban sound dataset with spatiotemporal context and strong labels. The data collection was done via a wireless acoustic sensor network deployed across Singapore. The multi-channel nature of a number of the recordings from the dataset also opens the possibility of localizing sound events within the dataset on top of detecting or classifying them. Lastly, an expanded label taxonomy for urban sounds compatible with the SONYC-UST taxonomy was introduced, thus allowing for benchmarking of urban sound tagging systems across multiple cities.

REFERENCES

[1] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, vol. 2017-Octob. Institute of Electrical and Electronics Engineers Inc., 12 2017, pp. 344–348.

[2] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2018-April. Institute of Electrical and Electronics Engineers Inc., 9 2018, pp. 351–355.

[3] M. Cartwright, J. Cramer, A. E. M. Mendez, 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. 5th Workshop Detect. Classif. Acoust. Scenes Events (DCASE)*, 9 2020.

[4] E. L. Tan, F. A. Karnapi, L. J. Ng, K. Ooi, and W. S. Gan, "Extracting Urban Sound Information for Residential Areas in Smart Cities Using an End-to-End IoT System," *IEEE Internet Things J.*, vol. 4662, no. c, 2021.

[5] T. C. J. Abeßer, H. Lukashevich, R. Gräfe, F. Häuser, C. Kühn, and T. Sporer, "Stadtlärm - a distributed system for noise level measurement and noise source identification in a smart city environment," in *Proc. Deutsche Jahrestagung für Akustik (DAGA)*, 2018, pp. 285–288.

[6] J. Abeßer, M. Götze, T. Clauß, D. Zapf, C. Kühn, H. Lukashevich, S. Kühnlenz, and S. I. Mimilakis, "Urban Noise Monitoring in the Stadtlärm Project - A Field Report," in *Proc. 4th Workshop Detect. Classif. Acoust. Scenes Events(DCASE)*. New York University, 10 2019, pp. 1–4.

[7] D. Salvo, G. Piñero, P. Arce, and A. Gonzalez, "A Low-cost Wireless Acoustic Sensor Network for the Classification of Urban Sounds," in *Proc. 17th ACM Symp. Perform. Eval. Wirel. Ad Hoc, Sensor, Ubiquitous Netw..* Association for Computing Machinery, Inc, 11 2020, pp. 49–55.

[8] M. Cartwright, A. E. M. Mendez, J. 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," in *Proc. 4th Workshop Detect. Classif. Acoust. Scenes Events (DCASE)*. New York University, 10 2019, pp. 35–39.

[9] 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. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 776–780.

[10] 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. 18th Int. Soc. for Music. Inf. Retr. Conf. ISMIR 2017*. International Society for Music Information Retrieval (ISMIR), 2017, pp. 486–493.

[11] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 2015 ACM Multimed. Conf..* Association for Computing Machinery, Inc, 10 2015, pp. 1015–1018.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. 3rd Workshop Detect. Classif. Acoust. Scenes Events (DCASE)*, 7 2018.

[13] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The NIGENS General Sound Events Database," 2 2019. [Online]. Available: http://arxiv.org/abs/1902.08314

[14] J. Abeßer, "USM-SED - A Dataset for Polyphonic Sound Event Detection in Urban Sound Monitoring Scenarios," 2021. [Online]. Available: http://arxiv.org/abs/2105.02592

[15] D. S. Johnson, W. Lorenz, M. Taenzer, S. Mimilakis, S. Grollmisch, J. Abeßer, and H. Lukashevich, "DESED-FL and URBAN-FL: Federated Learning Datasets for Sound Event Detection," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, 2021.

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

[17] J. P. Bello, C. Silva, O. Nov, R. Luke Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution," *Commun. ACM*, vol. 62, no. 2, pp. 68–77, 2 2019.

[18] A. Arnault and N. Riche, "CRNNs for Urban Sound Tagging with spatiotemporal context," IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events 2020, Tech. Rep., 2020.

[19] T. Iqbal, Y. Cao, M. D. Plumbley, and W. Wang, "Incorporating Auxiliary Data for Urban Sound Tagging," IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events 2020, Tech. Rep., 2020.

[20] K. Ooi, Y. Xie, B. Lam, and W. S. Gan, "Automation of binaural headphone audio calibration on an artificial head," *MethodsX*, vol. 8, p. 101288, 1 2021.

[21] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis," in *Proc. 4th Workshop Detect. Classif. Acoust. Scenes Events (DCASE)*, 2019, pp. 253–257.

[22] A. Cramer and M. Cartwright, "Baseline system for Task 5 (Urban Sound Tagging with Spatiotemporal Context) of the DCASE 2020 Challenge," 6 2020. [Online]. Available: https://github.com/sonyc-project/dcase2020task5-uststc-baseline

[23] L.-L. Kwok, "Urban sound tagging with spatiotemporal context," Nanyang Technological University, Tech. Rep., 2021. [Online]. Available: https://dr.ntu.edu.sg/handle/10356/149470