

# STELIN-US: A SPATIO-TEMPORALLY LINKED NEIGHBORHOOD URBAN SOUND DATABASE

*Snehit Chunarkar, Bo-Hao Su, Chi-Chun Lee*

Department of Electrical Engineering, National Tsing Hua University, Taiwan,  
snehit@gapp.nthu.edu.tw, borriusu@gapp.nthu.edu.tw, clee@ee.nthu.edu.tw

## ABSTRACT

Automated acoustic understanding, e.g., sound event detection and acoustic scene recognition, is an important research direction enabling numerous modern technologies. Although there is a wealth of corpora, most, if not all, include acoustic samples of scenes/events in isolation without considering their inter-connectivity with locations nearby in a neighborhood. Within a connected neighborhood, the temporal continuity and regional limitation (sound-location dependency) at distinct locations creates non-iid acoustics samples at each site across spatial-temporal dimensions. To our best knowledge, none of the previous data sources takes on this particular angle. In this work, we present a novel dataset, the Spatio-temporally Linked Neighborhood Urban Sound (STeLiN-US) database. The dataset is semi-synthesized, that is, each sample is generated by leveraging diverse sets of real urban sounds with crawled information of real-world user behaviors over time. This method helps create a realistic large-scale dataset, and we further evaluate it through perceptual listening tests. This neighborhood-based data generation opens up novel opportunities to advance user-centered applications with automated acoustic understanding. For example, to develop real-world technology to model a user’s speech data over a day, one can imagine utilizing this dataset as the user’s speech samples would modulate by diverse sources of acoustics surrounding linked across sites and temporally by natural behavior dynamics at each location over time.

**Index Terms**— Audio Dataset, Sound Synthesis, Urban Sound, Connected

## 1. INTRODUCTION

Understanding acoustic surroundings seamlessly influences our daily life, e.g., recognizing different emergencies by distinct alerting sounds. Besides, acoustic sounds also affect human mental health, e.g., work productivity in a calm/noisy environment [1], and psychological impact on our well-being as the change in stress level [2]. Thus, understanding acoustic sounds plays a crucial role in our life, which provides plentiful information to uplift environmental awareness and life quality. Especially recent advanced techniques and the support of superior hardware in deep learning show a prominent performance on these acoustic contextual tasks.

Basically, these acoustic context tasks can be generally divided into two categories, which are sound event detection (SED) and acoustic scene classification. Specifically, a sound event detection task aims to predict a short-term and precise event, e.g., a dog barking, a car passing by, or a cell phone ringing. Differing from that, an acoustic scene classification task targets an environment-wise contextualization, e.g., on the street and in a coffee shop, which may compound multiple sound events. Recently, for sound event de-

tection tasks, Turpault in [3] proposed to use weakly labeled data where a top-performed system using a convolutional neural network (CNN) model has achieved 42.7% F-measure. Besides, Ronchini et al. [4] integrated non-target events as auxiliary information while training and greatly impacted the SED task. As for acoustic scene classifications, DCASE has been predominantly focusing on scene classification in DCASE challenge Task 1 [5, 6, 7, 8, 9, 10] with constantly evolving their scope of interest within the task. Recently, they have been curious about the scope of this task on low-complexity approach [11] solutions, in which the top system competed with 48 submissions from 19 teams in the challenge and obtained 59.6% accuracy with 1.091 log loss. Both event detection and scene classification tasks manifest great accuracy in understanding the acoustic scenes/events with deep-learning-based models and provide insights for real-world applications.

However, most of them focus solely on scenes and events only. The currently published datasets used for similar tasks; only contain short-term audio from random locations and times isolatedly. None of them consider the inter-connectivity with locations nearby in a neighborhood. For instance, TAU Urban Acoustic Scenes 2020 Mobile [9] is one such designed for the scene classification task, but it lacks consistency in connectivity with its context of surrounding. The UrbanSound dataset [12] presents sound events compound with scrapped urban noises from the internet, which makes it diversely localized but poorly inter-linked. URBAN-SED dataset [13] having 11 events is a synthesized dataset aimed to compensate the sparsity of strongly annotated datasets; however, the same Brownian noise as background for all soundscape with predefined artificial synthesis settings barely justifies the real acoustic variation in an urban surrounding. ESC50 [14] with recordings in 2000 short clips emerge as one of the highest labeled environmental recording datasets bringing distinct 50 classes. Highlighting isolated high-quality sound events, the NIGENS dataset [15] brings 14 distinct sound event classes, including strong annotations. Whereas both ESC50 [14] and NIGENS [15] datasets are designed for SED tasks without the context of surrounding. SINS dataset [16] equipped with 16 activities aimed at activity detection in domestic environments for smart home applications. STARSS22 dataset [17] contains spatial recordings of real sound scenes collected in interiors, including temporal and spatial annotation of 13 sound events. However, both SINS [16] and STARSS22 [17] datasets sound recordings only in the interiors, which limits prominent datasets for diversity in the applications. Unlike the above-mentioned datasets, SONYC-UST [18] has attempted to build a dataset equipped with spatiotemporal metadata. The dataset contains real-world recordings in New York City with annotations defined using 23 tags based on New York City noise code. The highlight of SONYC-UST is the spatiotemporal context information that comforts monitoring the distribution of sound tags. But primarily focused on the events considered to be noise

in the urban environment, eliminating common sound events (e.g., Birds chirping) which are not considered noise. Also, the recordings are intuitively at outdoor locations limiting the SED applications to deal with the events in outdoor scenarios. Alternatively, in this work, we bring a new perspective/angle to this field; we consider an application-wise scenario that can be applied in a user-contextualized, environment-aware closer to our daily life. That is, many applications incorporating speech are published as well, e.g., Speech Enhancement applications, Automatic Speech Recognition (ASR) applications, and Speech Separation Tasks. To mention, an unsupervised federated learning approach proposed by [19] for speech enhancement and separation with a release of LibriFSD50K dataset. And Darius Petermann et al. in [20] introduce the separation of an audio mixture into speech, music, and sound effects using their proposed dataset named Divide and Remaster. However, they integrate acoustic scenes/events into speech but in a non-realistic and artificial manner. Whereas to do so, continuous recording from real-life scenes is required, and even with the recordings subsequently, it needs to be annotated for the event’s presence to be useful for SED tasks. Nevertheless, collecting new and large-scale recordings from the real world and annotating them is expensive, cumbersome, and time-consuming. Synthesis becomes a more feasible way to catch the scalability of existing speech datasets.

Hence, being a preliminary study to implement this idea, we develop a framework for synthesizing a continuous real-world acoustic distributed sound surrounding. Henceforth, we proposed this dataset with the inspiration to equip researchers with variable surrounding sound in an environment closely resembles realistic patterns. The proposed dataset models the small-scale connected surrounding in urban areas. The detail of the work is organized as follows: Section 2 presents the details of the synthesis, Section 3 summarizes the dataset and presents the analysis of the same, and with an end note, Section 4 discuss the dynamic scaling of the dataset with potential applications and concludes the present work.

## 2. METHODOLOGY

The proposed dataset is synthesized to represent a small-scale interconnected urban area. The synthesis framework is divided into Pre-conditions, Traffic, and Scene Synthesis. Here Preconditions deal with the requirements for the synthesis, Whereas the synthesis part is broadly divided into Traffic and Scene Synthesis.

### 2.1. Preconditions

Being an interconnected urban sound database, it is important to map the locations and patterns for scene-specific sound classes to conceptualize. Hence we presented a map in Fig. 1 for the proposed dataset; mapping both indoor and outdoor environments, 5 distinct locations were selected for synthesis representing a small-scale interconnected urban area. Street, Metro Station, Park, School Playground, and Cafe, represented by microphones M1, M2, M3, M4, and M5, respectively, are simulated with 14 acoustic sound classes. Of all classes, 6 represented the background, and 8 were the events. Vehicle, Train, Pedestrian, Cafe Crowd, Children Playing, Urban Park, Street Music, Phone Ring, School Bell, Car Horn, River, Bird, Fountain, and Dog Bark are considered acoustic sound classes. Train, Pedestrian, Cafe Crowd, Urban Park, River, and Fountain are considered as background, and the rest are the events. After a thorough review, the sound recordings for mentioned classes are adopted from a suitable published dataset, as in Table 1. At the same time, the pattern for the appearance of the sound classes Vehicle, Car Horn, Street Music, Pedestrian, and Dog Bark is inspired

by the annotation from real-world distribution of the closely relevant events from the SONYC [18]. And as for the background sound at the synthesized locations Metro Station, Park, and Cafe follows the google maps popular time index using LivePopularTimes<sup>1</sup> python package for the respective sound class. Specifically, searching nearby Manhattan, e.g., ”subway in Manhattan” prompt shows 18 results with the popular time index, which indicates the people’s traffic at that location. That helps relate to the density of background sound of the location, and taking the average for the number of results gives a general idea about the trend of busyness. Since the SONYC [18] data is mainly concentrated around Manhattan, searching for google maps popular times around that area makes the distribution consistent with the base area. This distribution is obtained for a week in an hourly fashion, which makes it convenient to design the density of events or the crowded nature of the background in a similar fashion.

### 2.2. Traffic Synthesis

Temporal connection across microphone locations is shaped by Traffic Synthesis. Autonomous from overall Scene Synthesis, Traffic Synthesis synthesizes a controlled flow of vehicles by tracing each vehicle’s course with calculated time for the appearance of the same vehicle at another microphone that comes under the vehicle’s track. There are 4 entry nodes considered for each vehicle to enter the environment as EN1-EN4. Now considering the more or less busy route, the path for the vehicle is decided with a random distribution till it exits the environment at the diagonally opposite node to its entry node. IDMT dataset [21] enriched with 4 different vehicle sounds at 3 different known speeds is best suited for Traffic Synthesis. Since the map in Fig. 1 is conceived with an approximated distance for microphone locations, hence compiling the information of speed, a good approximation for the timing is achieved by using  $Speed = \frac{Distance}{Time}$ . Following this set of conditions has equipped us with a temporal correlation across the microphone locations.

### 2.3. Scene Synthesis

Audio at 5 different microphone locations is synthesized to assemble a scene that furnishes realistic event patterns with the temporal connection. A brief overview of the acoustic classes at each synthesized location is given in Table 2. Following a realistic distribution described in 2.1, the dense nature of the environment is compiled in the synthesis by adding more audio segments on top of the same

<sup>1</sup><https://github.com/GrocerCheck/LivePopularTimes.git>

Table 1: Sound Classes and Dataset used for the synthesis

Sound Class	Source Dataset
Vehicle	IDMT Traffic [21]
Train, Cafe Crowd, Urban Park	TUT Rare Sound Events 2017 [22]
Pedestrian	TAU Urban Acoustic Scenes 2020 Mobile [9]
Children Playing, Street Music	UrbanSound [12]
Phone Ring	NIGENS [15]
School Bell, River, Fountain	FreeSound.org
Car Horn, Dog Bark	UrbanSound8K [12]
Bird	ESC-50 [14]

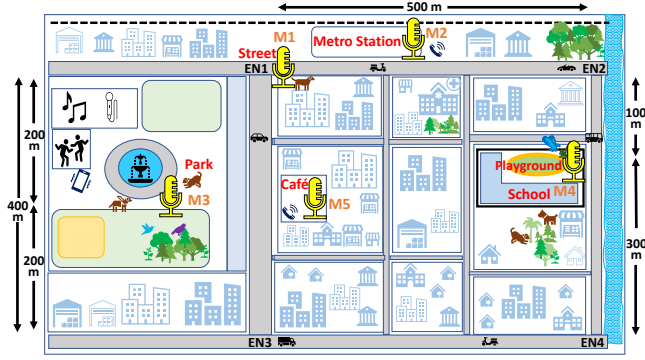


Figure 1: Acoustic Synthesis Map

sound class. A maximum of 3 audio segments have been added to represent the maximum dense structure. For the direct correlation of which popularity index from Google map has scaled down and quantized to an integer value to result in a range of 0-3. The dense scaling factor represents how many same-class audio segments to add for the background. On the other hand, the same factor for the event is inspired by a scaled version of SONYC [18] data for visualization of the event patterns, which indicates the number of sound events from the same classes added. However, events that are outside the SONYC [18] study are designed manually, e.g., Children playing and School bell class distribution in the School Playground microphone location (M4) are designed manually by considering school operating hours. The different sound classes merged to create a scene are scaled with the different intensities which is inspired by the inverse distance relationship with the sound intensity as in eq.(1). Whereas  $I_1, I_2$  are the original and synthesis sound intensities, and  $d_1, d_2$  are the respective distance of the recording from the source. Since the sounds taken from datasets do not contain the information regarding the distance of recording in detail, hence the chosen factor  $d_2$  is scaled in terms of distance  $d_1$  and then verified by manual listening for any resulting scaling change required. The data shown in Table 3 indicate the distance scale for the particular

Table 2: Combination of sound classes present at different locations throughout the week.

Location	Day	Sound Classes
Street	Mon-Sun	Vehicle, Pedestrian, Phone Ring, Car Horn, Dog Bark
	Metro Station	Train, Pedestrian, Phone Ring
Park	Mon-Sun	Vehicle, Pedestrian, Urban Park, Street Music, Phone Ring, Car Horn, Bird, Fountain, Dog Bark
	School Playground	Vehicle, Pedestrian, Children Playing, School Bell, Car Horn, River, Bird, Dog Bark
Cafe	Mon-Sun	Vehicle, Cafe crowd, Phone Ring, Car Horn
	Sat - Sun	Vehicle, Car Horn, River, Bird, Dog Bark

sound class used in synthesis, e.g., intensity scaling factor 2 indicates the audio event or background sound in synthesis audio will be twice as distance with respect to the one in the raw sound itself.

$$I_2 = I_1 \left( \frac{d_1}{d_2} \right)^2, \quad (1)$$

Overall, in the end, all the considered sound classes, after going through dense scaling and distance scaling processes, are added with each other to synthesize the scenario, which has the temporal pattern and interconnection with locations. Hence equipped us with one of its kind acoustic dataset designed to simulate the closely connected neighborhood urban area.

### 3. EXPERIMENTAL RESULTS

A brief assessment of the proposed dataset is presented, divided into a summary and analysis of the dataset. In the following sections, we discuss the summary and distribution of STeLiN-US further; we analyze it from a visual and human listener’s perspective.

#### 3.1. Summary

Following our proposed semi-synthesis procedures, we generate a Spatio-temporally Linked Neighborhood Urban Sound (STeLiN-US) database and is made available online<sup>2</sup>. Containing interconnected acoustic surroundings and scene-specific events, the proposed dataset is equipped with 525 audio clips comprising 43 hr 45 min in total. Synthesized for location-specific scene surrounding and adjunct with strong annotations for the events have reinforced the proposed dataset to be equivalently used in both scene classification and event detection tasks. Besides, embedding the time and day information with the synthesized acoustic scene has lifted the applicability from traditional tasks.

#### 3.2. Analysis

##### 3.2.1. Dataset Distribution

To visualize the distribution of each sound class in the final synthesis with respect to synthesized microphone locations, a series of bar

<sup>2</sup><https://doi.org/10.5281/zenodo.8241539>

Table 3: Scaling  $d_2 = k*d_1$ , considered k values for respective locations and sound class, where  $d_1, d_2$  are the respective distance of the recording from the source.

Sound Class	Locations				
	Street	M.Station	Park	School-P.G.	Cafe
Vehicle	2	-	4	5	5
Train	-	1	-	-	-
Pedestrian	1	1	3	4	-
Cafe Crowd	-	-	-	-	0.5
Children Playing	-	-	-	3	-
Urban Park	-	-	1	-	-
Street Music	-	-	3	-	-
Phone Ring	9	9	9	-	15
School Bell	-	-	-	3	-
Car Horn	3	-	5	6	6
River	-	-	-	4	-
Bird	-	-	2	3	-
Fountain	-	-	2	-	-
Dog Bark	4	-	2	5	-

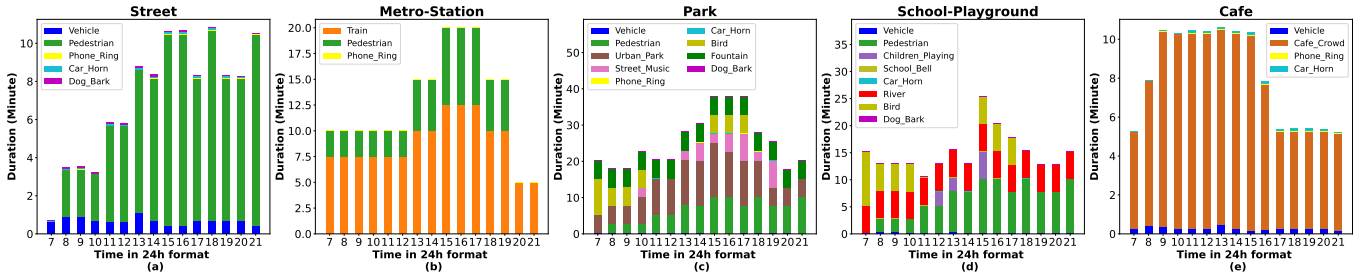


Figure 2: Average distribution of selected sound class in each synthesized location.

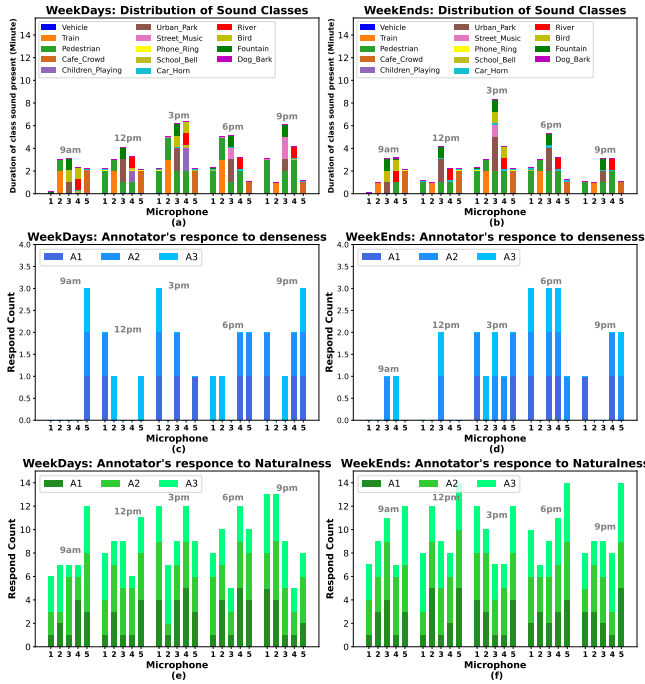


Figure 3: Sound class distribution at each synthesized location on (a) WeekDays (b) WeekEnds with corresponding annotators respond for dense/ sparseness in (c) and (d) and equivalent respond for naturalness in (e) and (f)

graphs are presented in Fig. 2. Following real environment distribution in sound classes, it showcases the distributed imbalance across time, location, and sound class e.g., from Fig. 2 (a), the Street is less busy during the morning than in late afternoon similarly in (e) cafe is busier from morning to afternoon than in late evenings, which are indeed the case in real life. Analogously, Fig. 2 can be compared with Table 2, which explains the presence of each class at synthesized microphone locations.

### 3.2.2. Listening Test

To validate the naturalness and sparseness of the proposed STeLiN-US, we further conducted a listening test by human annotation. In this experiment, a total of 50 audio samples (50 minutes in total) are selected randomly from STeLiN-US but evenly distributed in all synthesized microphone locations and times for this test. Precisely, all locations and time slots should present at least one time in the listening test set. During the listening test, we define two questions for annotators, including naturalness and sparseness. Nat-

uralness is annotated on a 5-Likert scale, where 1 represents strong disagreement on the naturalness of audio (i.e., the audio sounds artificial), and 5 means strong agreement on naturalness (i.e., the audio sounds natural). Similarly, sparseness is labeled by asking whether the audio sounds in rush hour, which is a binary(yes/no) question for them, and 0 for sparse, 1 for dense. In the overall listening test, we include 6 unique annotators (2 females, and 4 males) in total.

Henceforth, to analyze the distribution in a systematic way, we divide them into weekdays and weekends, as shown in Fig. 3 (a) & (b), respectively. Fig. 3 (c) & (d) represent the Dense/Sparseness results from three annotators (A1, A2, and A3) divided into weekdays and weekends, respectively. Similarly, Fig. 3 (e) & (f) is for naturalness result. To have statistical results, we further compute the average annotation among all the annotators and present their standard deviation as well. Notably, we get 0.36 and 3.12 average results for dense and naturalness, respectively, and similarly, we get 0.22 and 0.91 standard deviations. This conveys annotators agree closely for dense and naturalness results with lower deviation at the same time. It is amazing to observe the average naturalness result is more than half of the max on the scale with the least deviation among annotators depicting that even if the dataset is synthesized one is still inclined to feel natural alike.

## 4. DISCUSSION AND CONCLUSION

The proposed synthesis approach cultivated with real-world user behavior can be dynamically scaled to model any required environment. Such wide adaptability can elevate application-specific research solutions. Furnished with the real surrounding pattern distribution of sound classes, the proposed STeLiN-US dataset simulates the acoustic appearance of closely interconnected neighborhoods in urban areas. This help in not only identifying the scenes but also predicting acoustic scenarios. This accommodates the user-centered applications, e.g., If combined with the ASR, the ASR performance can be analyzed based on the location and time more than that possible performance can be predicted beforehand based on the prediction of the scene busyness. Hence this dataset can unveil many possible applications for the researcher. In contrast with previously published datasets, portraying diversity across locations yet interconnected and diverse events to truly justify the surrounding environment and still sound natural alike from the listening test has made the proposed dataset unique and unmatched. Such incorporation of scene-specific events to replicate the real surrounding environments facilitates researchers in testing trailblazing event detection systems.

## 5. ACKNOWLEDGMENT

The work was supported by the National Science and Technology Council (NSTC), Taiwan.

## 6. REFERENCES

- [1] J. Lim, K. Kweon, H. W. Kim, S. W. Cho, J. Park, and C. S. Sim, “Negative impact of noise and noise sensitivity on mental health in childhood,” *Noise Health*, vol. 20, no. 96, pp. 199–211, 2018.
- [2] L. I. Yankoty, P. Gamache, C. Plante, S. Goudreau, C. Blais, S. Perron, M. Fournier, M. S. Ragetti, M. Hatzopoulou, Y. Liu, and A. Smargiassi, “Relationships between long-term residential exposure to total environmental noise and stroke incidence,” *Noise Health*, vol. 24, no. 113, pp. 33–39, 2022.
- [3] N. Turpault, R. Serizel, A. Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” in *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, p. 253.
- [4] F. Ronchini, R. Serizel, N. Turpault, and S. Cornell, “The impact of non-target events in synthetic soundscapes for sound event detection,” in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 115–119.
- [5] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.
- [6] A. Mesaros, T. Heittola, and T. Virtanen, “Assessment of human and machine performance in acoustic scene classification: Dcase 2016 case study,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 319–323.
- [7] —, “Acoustic scene classification: An overview of dcase 2017 challenge entries,” in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 411–415.
- [8] —, “A multi-device dataset for urban acoustic scene classification,” in *Scenes and Events 2018 Workshop (DCASE2018)*, p. 9.
- [9] T. Heittola, A. Mesaros, and T. Virtanen, “Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60. [Online]. Available: <https://arxiv.org/abs/2005.14623>
- [10] I. Martín-Morató, T. Heittola, A. Mesaros, and T. Virtanen, “Low-complexity acoustic scene classification for multi-device audio: analysis of dcase 2021 challenge systems,” 2021.
- [11] I. Martín-Morató, F. Paissan, A. Ancilotto, T. Heittola, A. Mesaros, E. Farella, A. Brutti, and T. Virtanen, “Low-complexity acoustic scene classification in dcase 2022 challenge,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.03835>
- [12] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in *22nd ACM International Conference on Multimedia (ACM-MM’14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.
- [13] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, “Scaper: A library for soundscape synthesis and augmentation,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.
- [14] K. J. Piczak, “Esc: Dataset for environmental sound classification,” in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1015–1018. [Online]. Available: <https://doi.org/10.1145/2733373.2806390>
- [15] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, “The nignens general sound events database,” 2020.
- [16] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, “The SINS database for detection of daily activities in a home environment using an acoustic sensor network,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 32–36.
- [17] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, “STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129. [Online]. Available: <https://dcase.community/workshop2022/proceedings>
- [18] M. Cartwright, A. E. M. Mendez, J. Cramer, V. Lostonlen, 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 *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, October 2019, pp. 35–39. [Online]. Available: [http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop\\_Cartwright.4.pdf](http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Cartwright.4.pdf)
- [19] E. Tzinis, J. Casebeer, Z. Wang, and P. Smaragdis, “Separate but together: Unsupervised federated learning for speech enhancement from non-IID data,” in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, oct 2021. [Online]. Available: <https://doi.org/10.1109%2Fwaspaaw52581.2021.9632783>
- [20] D. Petermann, G. Wichern, Z.-Q. Wang, and J. L. Roux, “The cocktail fork problem: Three-stem audio separation for real-world soundtracks,” 2022.
- [21] J. Abeßer, S. Gourishetti, A. Kátai, T. Clauß, P. Sharma, and J. Liebetrau, “Idmt-traffic: An open benchmark dataset for acoustic traffic monitoring research,” 2021.
- [22] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, “DCASE 2017 challenge setup: Tasks, datasets and baseline system,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.