SONYC URBAN SOUND TAGGING (SONYC-UST): A MULTILABEL DATASET FROM AN URBAN ACOUSTIC SENSOR NETWORK

Mark Cartwright[1,2,3*], *Ana Elisa Mendez Mendez*[1], *Jason Cramer*[1], *Vincent Lostanlen*[1,2,4], *Graham Dove*[6], *Ho-Hsiang Wu*[1], *Justin Salamon*[5], *Oded Nov*[6], *and Juan Pablo Bello*[1,2,3]

[1] Music and Audio Resarch Lab, New York University, NY, USA
[2] Center for Urban Science and Progress, New York University, NY, USA
[3] Department of Computer Science and Engineering, New York University, NY, USA
[4] Cornell Lab of Ornithology, Cornell University, NY, USA
[5] Adobe Research, San Francisco, CA, USA
[6] Department of Technology Management and Innovation, New York University, NY, USA

ABSTRACT

SONYC Urban Sound Tagging (SONYC-UST) is a dataset for the development and evaluation of machine listening systems for real-world urban noise monitoring. It consists of 3068 audio recordings from the "Sounds of New York City" (SONYC) acoustic sensor network. Via the Zooniverse citizen science platform, volunteers tagged the presence of 23 fine-grained classes that were chosen in consultation with the New York City Department of Environmental Protection. These 23 fine-grained classes can be grouped into eight coarse-grained classes. In this work, we describe the collection of this dataset, metrics used to evaluate tagging systems, and the results of a simple baseline model.

Index Terms— Audio databases, Urban noise pollution, Sound event detection

1. INTRODUCTION

Noise pollution is a major concern for urban residents and has negative effects on residents' health [1, 2, 3] and learning [2, 4]. To mitigate the recurrence of harmful sounds, the City of New York employs a legal enforcement strategy guided by a "noise code". For example, jackhammers can only operate on weekdays; pet owners are held accountable for their animals' noises; ice cream trucks may only play their jingles while in motion; blasting a car horn is restricted to situations of imminent danger. After a city resident complains about noise, the New York City Department of Environmental Protection (DEP) sends an inspector to investigate the complaint. If the inspector is able to confirm that the offending noise violates the noise code, they incentivize the manager of the noise source to reduce their noise footprint in compliance with the code. Unfortunately, this complaint-driven enforcement approach results in a mitigation response biased to neighborhoods who complain the most, not necessarily the areas in which noise causes the most harm. In addition, due to the transient nature of sound, the offending noise source may have already ceased by the time an inspector arrives on site to investigate the complaint.

Sounds of New York City (SONYC) is a research project investigating data-driven approaches to mitigating urban noise pollution. One of its aims is to map the spatiotemporal distribution of noise at the scale of a megacity like New York City, in real time, and throughout multiple years. With such a map, city officials could better understand noise in the city; more effectively allocate city resources for mitigation; and develop informed mitigation strategies while alleviating the biases inherent to complaint-driven approaches. To this end, SONYC has designed an acoustic sensor for noise pollution monitoring that combines relatively high quality sound acquisition with a relatively low production cost [5]. Between 2016 and 2019, over 50 different sensors have been assembled and deployed in various areas of New York City.

Each SONYC sensor measures the sound pressure level (SPL) of its immediate vicinity, but it does not infer and report the causes of changes in SPL. From a perceptual standpoint, not all sources of outdoor noise are equally unpleasant, nor are they equally enforcible with respect to the noise code. Therefore, it is necessary to resort to computational methods for detection and classification of acoustic scenes and events (DCASE) in the context of automated noise pollution monitoring. To address this, the sensors have also been collecting non-contiguous 10 s audio recordings during deployment and have collectively gathered over 100 M recordings.

There are several attributes of urban sound event detection that make it a challenging task. Sound sources of interest are often far away from the sensors. Several sources of interest may occur simultaneously. Many sound classes seem quite similar, yet are distinct in the noise code and so should be identified as such. Many other distractor sounds occur within urban sound recordings. And lastly, the acoustic environment changes by location and by time within seasonal cycles. Due to the complexity of this problem, it is important to evaluate machine listening systems for monitoring urban noise in realistic scenarios, using actual recordings from urban noise sensors and a label space that matches the needs of city agencies.

In this article, we present the SONYC Urban Sound Tagging (SONYC-UST) dataset[1], which contains 3068 annotated 10 s recordings from the SONYC acoustic sensor network and which served as the dataset for the DCASE 2019 Urban Sound Tagging Challenge[2]. Each recording has been annotated using a set of 23 "tags", which was developed in coordination with the New York City Department of Environmental Protection (DEP) and represents

*This work was partially funded by National Science Foundation awards 1633259 and 1544753

[1] Download the data at https://doi.org/10.5281/zenodo.3338310
[2] http://dcase.community/challenge2019/task-urban-sound-tagging

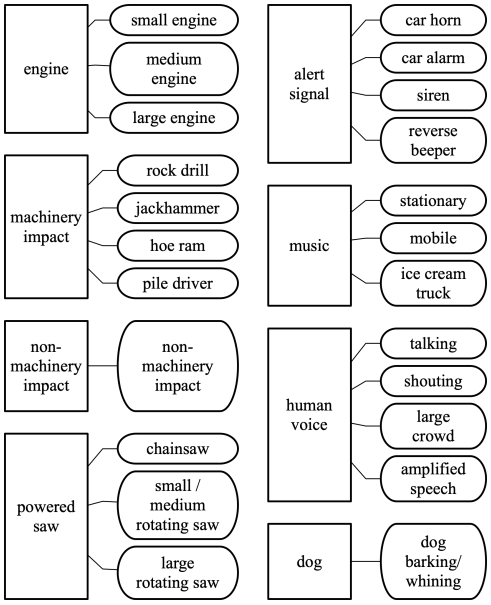

Figure 1: Hierarchical taxonomy of the SONYC Urban Sound Tagging (SONYC-UST) dataset. Rectangular and round boxes respectively denote coarse and fine urban sound tags.

many of the frequent causes of noise complaints in New York City.

Existing datasets for urban noise monitoring do not accurately represent the problem of urban noise monitoring. The freefield1010 [6], UrbanSound [7], UrbanSound8k, [7], and Urban-SED [8] datasets contain recordings curated from Freesound [9] rather than recorded in a realistic noise monitoring scenario. In addition, these datasets are multi-class, in which only the predominant sound class is labeled. The exception is Urban-SED [8], which does have strong, multi-label annotations, but it is a synthetic dataset that is not representative of actual urban soundscapes. The TUT Sound Events 2016 [10, 11, 12] and 2017 [13, 14] datsets consists of audio recordings in real urban environments as well as providing strong, multi-label annotations. However, these datasets have label sets limited to human presence and traffic, and their spatiotemporal context is limited to a handful of times and locations. SONYC-UST addresses these limitations by providing recordings from urban noise sensors across a variety of times and locations, and by more closely matching the label set to the needs of noise enforcement agencies.

2. SONYC-UST TAXONOMY

Through consultation with the New York Department of Environmental Protection (DEP) and the New York noise code, we constructed a small, two-level urban sound taxonomy (see Figure 1) consisting of 8 coarse-level and 23 fine-level sound categories, e.g., the coarse *alert signals* category contains four fine-level categories: *reverse beeper*, *car alarm*, *car horn*, *siren*. Unlike the Urban Sound Taxonomy [7], this taxonomy is not intended to provide a framework for exhaustive description of urban sounds. Instead, it was scoped to provide actionable information to the DEP, while also being understandable and manageable for novice annotators. The chosen sound categories map to categories of interest in the noise code; they were limited to those that seem likely discernible by novice annotators; and we kept the number of categories small enough so that they can all be visible at once in an annotation interface.

3. DATA COLLECTION

The SONYC acoustic sensor network consists of more than 50 acoustic sensors deployed around New York City and has recorded over 100M 10-second audio clips since its launch in 2016. The sensors are located in the Manhattan, Brooklyn, and Queens boroughs of New York, with the highest concentration around New York University's Manhattan campus. To maintain the privacy of bystanders' conversations, the network's sensors are positioned for far-field recording, 15–25 feet above the ground, and record audio clips at random intervals, rather than continuously.

To annotate the sensor recordings, we launched an annotation campaign on Zooniverse [15, 16], the largest citizen-science platform. In a previous study comparing multiple types of weak annotation tasks, we found that full multi-label annotation (i.e., an annotation task in which all classes are annotated at once by each annotator) with at least three annotators per recording resulted in high quality annotations and high throughput with citizen science volunteers [17]. In another previous study, we found that spectrogram visualizations aided annotators in producing high quality annotations [18]. Given these findings, we configured the annotation task as a multi-label, weak annotation (i.e., tagging) task in which the annotators were presented with a spectrogram visualization of the audio clip along with the audio playback.

After presenting volunteers with instructions explaining the task and a field guide describing the SONYC-UST classes, we asked them to annotate the presence of all of the fine-level classes in a recording. For every coarse-level class (e.g., *alert signal*) we also included a fine-level *other/unknown* class (e.g., *other/unknown alert signal*) with the goal of capturing an annotator's uncertainty in a fine-level tag while still annotating the coarse-level class. If an annotator marked a sound class as present in the recording, they were also asked to annotate the proximity of the sound event (*near*, *far*, *not sure*). Volunteers could annotate as many recordings as were available.

Manually annotating all 100M+ of the unlabeled sensor recordings is not feasible, but annotating a random sample is not efficient since many of them may not contain sound events of interest. To address this, we sample sensor recordings that are most similar to a small set of exemplary clips for each sound class in our taxonomy. The exemplary clips were curated from YouTube and selected based on the presence of the target class in the audio along with visual confirmation from the video. Similarity to the exemplary clips was computed using a distance function D, which compares a sensor recording to M exemplary clips for a particular class:

$$D(\boldsymbol{X}^{(c)}, \boldsymbol{y}_n) = \sum_m \frac{1}{K_m} \sum_k^{K_m} \min_j d(x_{m,k}^{(c)}, y_{n,j})^2 \qquad (1)$$

where $x_{m,k}^{(c)}$ is the k^{th} VGGish [19] embedding frame of the m^{th} example clip (from class c) with K_m frames, $\boldsymbol{X}^{(c)}$ represents the M exemplary clips from class c, $y_{n,j}$ is the j^{th} VGGish embedding frame of the n^{th} sensor recording $\boldsymbol{y}_n$, and d is the Euclidean distance function.

The SONYC-UST dataset contains annotated train, validate, and test splits (2351 / 443 / 274 recordings respectively). We selected these splits such that recordings from the same sensors would

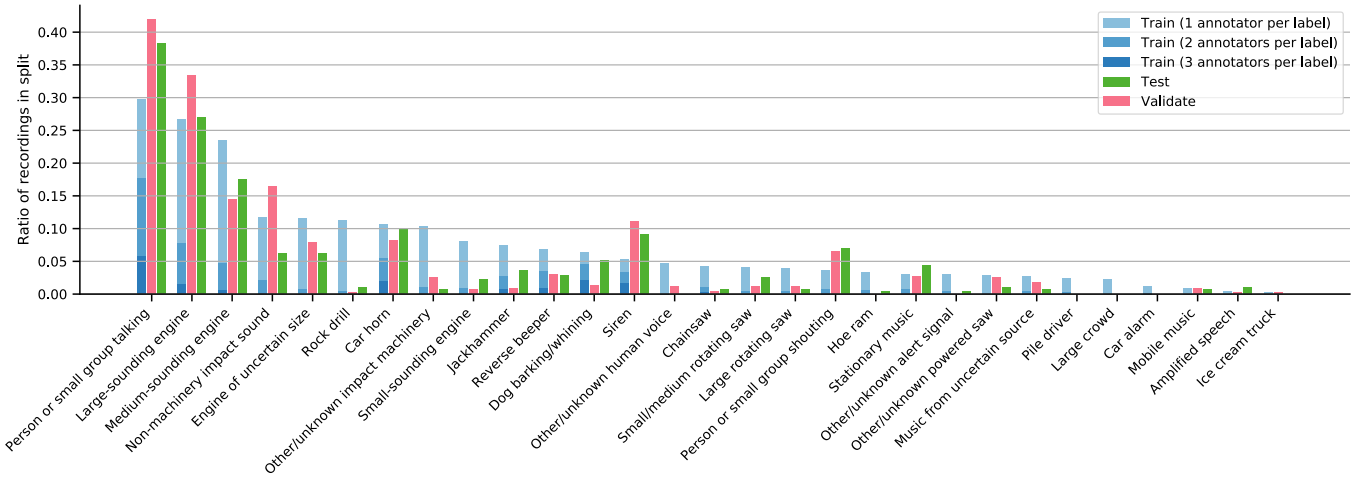

Figure 2: SONYC-UST tag distribution normalized for each split, in decreasing order of frequency in the train split. The shades of blue indicate how many annotators tagged the class in a training set recording, i.e. darker shades of blue indicate higher annotator agreement.

not appear in both the train and validate sets, and such that the distributions of citizen-science-provided labels were similar for both the train and validate sets (see Figure 2). While doing so, we considered a class present in a sample if at least one of the three volunteers annotated it as such. 35 sensors were assigned to the training set and 8 sensors assigned to the validate set. Unlike the train/validate sets, the test set is not disjoint in terms of sensors, but rather it is disjoint in time—all recordings in the test set are posterior to those in the train/validate sets. This allows us to evaluate model generalization to known locations at unseen times.

In addition to the crowdsourced annotations from Zooniverse, we include "SONYC-team-verified" labels for both the validation and test splits. To create verified labels, we first distributed recordings based on coarse-level sound category to members of the SONYC research team for labeling. To determine whether a recording belonged to a specific category for the validation process, we selected those that had been annotated by at least one volunteer. Next, two members of our team labeled each category independently. Once each member had finished labeling their assigned categories, the two annotators for each class discussed and resolved label disagreements that occurred during the independent annotation process. We use these agreed-upon "SONYC-team-verified" labels as the "ground truth" when evaluating models. We also use these labels to evaluate the annotations from Zooniverse, aggregated using minority vote, which we have previously shown to be an effective aggregation strategy in this context [17]. To aggregate with minority vote, we simply count a positive for a tag if at least one annotator has labeled the audio clip with that tag. In Table 1, we present annotation accuracy results using the metrics described Section 4. When examining the class-wise F1 scores, we see that crowdsourced annotations score well against the ground-truth for many classes, but it seems the Zooniverse annotators have difficulty identifying impact sounds and powered saws, especially when discriminating between fine-level classes.

In the SONYC-UST dataset, we include the Zooniverse volunteers' fine-level multi-label class-presence and proximity annotations for all the audio recordings in all three data splits. We also provide the SONYC-team-verified multi-label class-presence annotations and the agreed-upon ground-truth class-presence labels for the validate and test sets. All annotations also include identifiers for both the annotator and the sensor from which the clip was recorded. The coarse-level indicators of class presence are also included and are computed by logical disjunction over the fine-level class-presence tags associated with the coarse-level category.

Estimator:	Annotators				Baseline Model			
Split:	Validate		Test		Validate		Test	
Level:	F	C	F	C	F	C	F	C
Overall								
AUPRC	.73	.87	.75	.90	.67	.77	.62	.76
F1@0.5	.68	.83	.68	.84	.50	.70	.43	.67
Class F1@0.5								
Engine	.64	.94	.64	.94	.37	.79	.29	.76
Mech. imp.	.25	.24	.32	.35	.29	.36	.62	.58
Non-mech. imp.	.49	.49	.60	.60	.05	.02	.00	.11
Powered saw	.47	.70	.28	.62	.30	.66	.45	.66
Alert signal	.88	.95	.87	.92	.48	.67	.35	.48
Music	.59	.76	.52	.76	.07	.07	.00	.00
Human voice	.82	.95	.82	.96	.77	.84	.63	.77
Dog	.74	.74	.96	.96	.00	.00	.66	.66

Table 1: The performance of the Zooniverse annotations (using minority vote aggregation) and the baseline classifier as compared the the ground-truth annotations for both validate and test splits on the coarse (C) and fine (F) levels. AUPRC and F1 are both micro-averaged.

4. MULTILABEL CLASSIFICATION METRICS

Due to the presence of *other/unknown* tags, SONYC-UST has an incomplete ground truth at the fine taxonomical level. Such incompleteness poses a problem for evaluating multilabel classifiers. We propose a pragmatic solution to this problem; the guiding idea behind our solution is to evaluate the prediction at the fine level only when possible, and fall back to the coarse level if necessary.

Let a coarse-level category contain K fine-level tags. We de-

note by $t_1 \dots t_K$ the indicators of presence of these tags in the ground truth. For $k \in \{1 \dots K\}$, the integer t_k is equal to 1 if the fine-level tag k is present in the ground truth and equal to 0 otherwise. Furthermore, we denote by t_0 the indicator of presence of the *other/unknown* tag in the ground truth for the coarse category at hand. In the following, we adopt the bar notation $\overline{t_k}$ as a shorthand for the logical negation $(1 - t_k)$. Whereas the fine-level composition of the coarse category cannot be assessed with certainty, taking the product of all integers $\overline{t_0} \dots \overline{t_k}$ yields a coarse indicator of *certain absence*, equal to 1 if and only if none of the fine-level tags is present, even the uncertain one. This operation of tag coarsening allows to evaluate any prediction $\boldsymbol{y}$ against the ground truth $\boldsymbol{t}$. In each coarse category, the comparison of $\boldsymbol{y}$ and $\boldsymbol{t}$ results in, either, a true positive (TP), a false positive (FP), or a false negative (FN):

$$\begin{aligned}
\mathrm{TP}_{\mathrm{coarse}} &= \left(1 - \prod_{k=0}^{K} \overline{t_k}\right) \times \left(1 - \prod_{k=0}^{K} \overline{y_k}\right) \\
\mathrm{FP}_{\mathrm{coarse}} &= \left(\prod_{k=0}^{K} \overline{t_k}\right) \times \left(1 - \prod_{k=0}^{K} \overline{y_k}\right) \\
\mathrm{FN}_{\mathrm{coarse}} &= \left(1 - \prod_{k=0}^{K} \overline{t_k}\right) \times \left(\prod_{k=0}^{K} \overline{y_k}\right). \qquad (2)
\end{aligned}$$

The three numbers above are equal to zero or one, and sum to one in each coarse category. Although they are resilient to the incompleteness of tags, this comes at the cost of them being insensitive to permutations of complete fine-level tags within the same coarse category. Therefore, we propose the alternative definitions below:

$$\begin{aligned}
\mathrm{TP}_{\mathrm{fine}} &= \left(\sum_{k=1}^{K} t_k y_k\right) + t_0 \times \left(1 - \prod_{k=1}^{K} t_k y_k\right) \times \left(1 - \prod_{k=0}^{K} \overline{y_k}\right), \\
\mathrm{FP}_{\mathrm{fine}} &= \overline{t_0} \times \left(\sum_{k=1}^{K} \overline{t_k} y_k\right) + \overline{t_0} y_0 \times \left(\prod_{k=1}^{K} \overline{t_k}\right) \times \left(1 - \prod_{k=1}^{K} y_k\right), \\
\mathrm{FN}_{\mathrm{fine}} &= \left(\sum_{k=1}^{K} t_k \overline{y_k}\right) + t_0 \times \left(\prod_{k=1}^{K} \overline{t_k}\right) \times \left(\prod_{k=0}^{K} \overline{y_k}\right). \qquad (3)
\end{aligned}$$

In contrast to their coarse counterparts, these counters range from 0 to K. In the simple case where the ground truth is complete (i.e., $t_0 = 0$), they boil down to a one-to-one comparison of complete fine-level predicted tags y_k with the complete fine-level ground truth tags t_k, with the incomplete prediction y_0 being counted as a false positive if present. However, if the ground truth contains the incomplete tag (i.e., $t_0 = 1$), $\mathrm{FP}_{\mathrm{fine}}$ falls to zero. If, in addition, no complete fine-level ground truth tag t_k matches a complete fine-level prediction (i.e., $t_k y_k = 0$ for all $k > 0$), then the number of true positives is set to one if the coarsened predicted tag is present (i.e., $y_k = 0$ for any $k \geq 0$) and zero otherwise. Lastly, if the coarsened predicted tag is absent (i.e., $y_k = 0$ for all $k \geq 0$) and if the ground truth does not contain any complete tag (i.e., $t_k = 0$ for all $k > 0$), then the number of false negatives is set to t_0.

For example, if a small engine is present in the ground truth and absent in the prediction but an *other/unknown engine* is predicted, then it is a true positive in the coarse-grained sense, but a false negative in the fine-grained sense. However, if a small engine is absent in the ground truth and present in the prediction, then the outcome of the evaluation will depend on the completeness of the ground truth for the coarse category of engines. If this coarse category is complete (i.e. if the tag "engine of uncertain size" is absent from the ground truth), then we may evaluate the small engine tag at the fine level, and count it as a false positive. Conversely, if the coarse category of engines is incomplete (i.e. the tag "engine of uncertain size" is present in the ground truth), then we fall back to coarse-level evaluation for the sample at hand, and count the small engine prediction as a true positive, in aggregation with potential predictions of medium engines and large engines.

In each coarse category, these integer counts can then be translated into well-known information retrieval metrics: precision, recall, F1 score, and area under the precision recall curve (AUPRC). Furthermore, they can be micro-averaged across coarse categories to yield an overall F1 score and an overall AUPRC. The repository of our baseline system contains an open-source implementation of these metrics, both for "coarse" and "fine" formulas[3].

5. BASELINE SYSTEM

For the baseline classifier (cf. footnote 3) we use a multi-label logistic regression model. The model uses VGGish embeddings [19] as its input representation, which are computed from non-overlapping 0.96-second windows, giving us ten frames of 128-dimensional embeddings for each clip in our dataset. We train the model at the frame-level and use the weak tags from each audio clip as the targets for each of the 10 frames in a clip. To aggregate the crowdsourced annotations for training, we count a positive for a tag if at least one annotator has labeled the audio clip with that tag, i.e. minority vote.

We trained the model using stochastic gradient descent to minimize binary cross-entropy loss. To train the model to predict fine-level tags, the loss is modified such that if "other/unknown" is annotated for a particular coarse tag, the loss for the fine tags corresponding to this coarse tag is masked out. We use early stopping based on the validation set loss to mitigate overfitting. We trained one model to only predict fine-level tags, and we trained another model to only predict coarse-level tags.

For clip-level inference, we predict tags at the frame level and take the average of output tag probabilities as the clip-level tag probabilities. The resulting summary and class-wise metrics are presented in Table 1. Overall the baseline models achieved an AUPRC of 0.62 and 0.76 on the test split's fine and coarse levels respectively, and performed poorly on *music* and *non-machinery impact* sounds, leaving considerable room for improvement.

6. CONCLUSION

SONYC-UST is a multi-label dataset for urban sound tagging, recorded from an urban acoustic sensor network and annotated by crowdsourced volunteers. This dataset addresses deficiencies in the current set of urban sound datasets by providing real-world recordings and a label set that more closely matches the needs of city agencies in charge of noise abatement. In this work, we present the process used to collect this data; a taxonomy of urban sound tags informed by the New York City noise code and consultation with noise enforcement agents; metrics to evaluate tagging systems with uncertain ground-truth data; and a baseline model demonstrating that this is a challenging task with considerable room for improvement. We hope this dataset will encourage researchers to focus on this problem and advance the state of the art in urban sound event detection, helping build tools to make cities quieter.

[3] https://github.com/sonyc-project/urban-sound-tagging-baseline

7. REFERENCES

[1] M. S. Hammer, T. K. Swinburn, and R. L. Neitzel, "Environmental noise pollution in the united states: developing an effective public health response," *Environmental Health Perspectives*, vol. 122, no. 2, pp. 115–119, 2013.

[2] A. Bronzaft, "Neighborhood noise and its consequences," *Survey Research Unit, School of Public Affairs, Baruch College, New York*, 2007.

[3] World Health Organization, "Burden of disease from environmental noise: Quantification of healthy life years lost in europe," in *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*, 2011, pp. 126–126.

[4] M. Basner, W. Babisch, A. Davis, M. Brink, C. Clark, S. Janssen, and S. Stansfeld, "Auditory and non-auditory effects of noise on health," *The Lancet*, vol. 383, no. 9925, pp. 1325–1332, 2014.

[5] C. Mydlarz, M. Sharma, Y. Lockerman, B. Steers, C. Silva, and J. P. Bello, "The life of a new york city noise sensor network," *Sensors*, vol. 19, no. 6, p. 1415, 2019.

[6] D. Stowell and M. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, Jan 2014. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=17095

[7] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[8] 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)*. IEEE, 2017, pp. 344–348.

[9] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 411–412.

[10] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[11] ——, "Tut sound events 2016, development dataset," 2016. [Online]. Available: https://zenodo.org/record/45759

[12] ——, "Tut sound events 2016, evaluation dataset," 2017. [Online]. Available: https://zenodo.org/record/996424

[13] ——, "Tut sound events 2017, development dataset," 2017. [Online]. Available: https://zenodo.org/record/814831

[14] ——, "Tut sound events 2017, evaluation dataset," 2017. [Online]. Available: https://zenodo.org/record/1040179

[15] R. Simpson, K. R. Page, and D. De Roure, "Zooniverse: Observing the world's largest citizen science platform," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14 Companion. New York, NY, USA: ACM, 2014, pp. 1049–1054.

[16] "Zooniverse." [Online]. Available: www.zooniverse.org

[17] M. Cartwright, G. Dove, A. E. Méndez Méndez, J. P. Bello, and O. Nov, "Crowdsourcing multi-label audio annotation tasks with citizen scientists," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 292.

[18] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J. P. Bello, and O. Nov, "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, p. 29, 2017.

[19] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430