SONYC-UST-V2: AN URBAN SOUND TAGGING DATASET WITH SPATIOTEMPORAL CONTEXT

Mark Cartwright[1*], *Jason Cramer*[1], *Ana Elisa Mendez Mendez*[1], *Yu Wang*[1]
Ho-Hsiang Wu[1], *Vincent Lostanlen*[1,2], *Magdalena Fuentes*[1], *Graham Dove*[1]
Charlie Mydlarz[1], *Justin Salamon*[3], *Oded Nov*[1], *and Juan Pablo Bello*[1]

[1] New York University, New York, NY, USA
[2] Cornell Lab of Ornithology, Ithaca, NY, USA
[3] Adobe Research, San Francisco, CA, USA

ABSTRACT

We present SONYC-UST-V2, a dataset for urban sound tagging with spatiotemporal information. This dataset is aimed for the development and evaluation of machine listening systems for real-world urban noise monitoring. While datasets of urban recordings are available, this dataset provides the opportunity to investigate how spatiotemporal metadata can aid in the prediction of urban sound tags. SONYC-UST-V2 consists of 18510 audio recordings from the "Sounds of New York City" (SONYC) acoustic sensor network, including the timestamp of audio acquisition and location of the sensor. The dataset contains annotations by volunteers from the Zooniverse citizen science platform, as well as a two-stage verification with our team. In this article, we describe our data collection procedure and propose evaluation metrics for multilabel classification of urban sound tags. We report the results of a simple baseline model that exploits spatiotemporal information.

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

1. INTRODUCTION

Often in machine listening research, researchers work with datasets scraped from the internet, disconnected from real applications, and devoid of relevant metadata such as when and where the data were recorded. However, this is not the case in many real-world sensing applications. In many scenarios, we *do* know when and where the data were recorded, and this spatiotemporal context (STC) metadata may inform us as to what objects or events we may expect to occur in a recording. Computer vision researchers have already shown that STC is helpful in detecting objects such as animals in camera trap images and vehicles in traffic camera images [1]. We believe STC may also aid in sound event detection tasks such as urban sound tagging, e.g Figure 1, by informing us as to what sound events we may expect to hear in sound recordings. For example, in New York City you are more likely to hear an ice cream truck by the park at 3pm on a Saturday in July than you are by a busy street at rush hour on a Tuesday in January; however, you are more likely to hear honking, engines, and sirens on that Tuesday. But, knowledge of a thunderstorm that Saturday afternoon in July would reduce your expectation to hear an ice cream truck and could also help you disambiguate between the noise of heavy rain and that of a large walk-behind saw. However, few works have exploited this information for urban sound tagging [2] or even sound tagging in general. We hypothesize that one of the main reasons for this is the lack of available data with audio and temporal and spatial metadata.

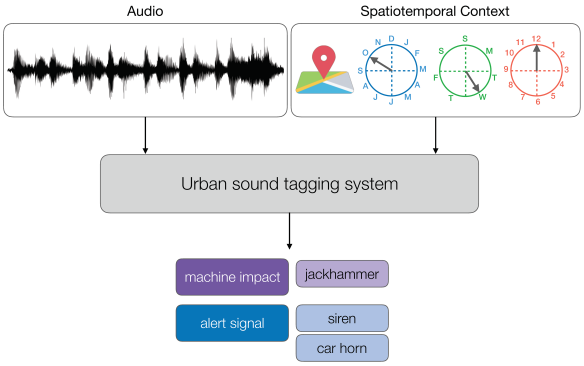

Figure 1: Overview of a system that exploits spatiotemporal information for urban sound tagging.

In this article, we introduce SONYC-UST-V2, a dataset for urban sound tagging with spatiotemporal information,[1] which contains 18510 annotated 10 s recordings from the SONYC acoustic sensor network and which served as the dataset for the DCASE 2020 Urban Sound Tagging with Spatiotemporal Challenge[2]. Each recording has been annotated on a set of 23 "tags", which was developed in coordination with the New York City Department of Environmental Protection (DEP) and represents many of the frequent causes of noise complaints in New York City. In addition to the recording, we provide identifiers for the New York City block (location) where the recording was taken as well as when the recording was taken, quantized to the hour. This information alone can be used to help a tagging model learn the "rhythm" of the city, but it can also be used query and join external datasets that can provide additional contextual information, e.g. weather, traffic, holidays, land use, city permits, and social data—all of which are available through rich, public datasets. We hope this data and task can provide a test bed for investigating these ideas for machine listening.

*We would like to thank all the Zooniverse volunteers who continue to contribute to our project. This work is supported by National Science Foundation awards 1544753 and 1633259.

[1] Download the data at https://doi.org/10.5281/zenodo.3966543.

[2] http://dcase.community/challenge2020/task-urban-sound-tagging-with-spatiotemporal-context.

2. PREVIOUS WORK

SONYC Urban Sound Tagging (SONYC-UST, referred to from here on as SONYC-UST-V1) is a dataset for the development and evaluation of machine listening systems for real-world urban noise monitoring [3]. It was used for the Urban Sound Tagging challenge in DCASE 2019, and consists of 3068 audio recordings from the SONYC acoustic sensor network [4]. This acoustic network consists of more than 50 acoustic sensors deployed around New York City and has recorded 150M+ 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 (see Figure 2). To maintain the privacy of bystanders' conversations and prevent the recording of intelligible conversation, the network's sensors are positioned for far-field recording, 15–25 feet above the ground, and record audio clips at random intervals.

The SONYC-UST-V1 dataset contains annotated training, validation, and test splits (2351 / 443 / 274 recordings respectively). These splits were selected so recordings from the same sensors would not appear in both the training and validation sets, and such that the distributions of labels were similar for both the training and validation sets. Finally, 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 training and validation sets.

The recordings were annotated by citizen volunteers via the Zooniverse citizen science platform [5, 6] and were followed by a two-step verification by our team in the case of the validation and test splits. In Zooniverse, volunteers weakly tagged the presence of 23 fine-grained classes that were chosen in consultation with the New York DEP. These 23 fine-grained classes are then grouped into eight coarse-grained classes with more general concepts: e.g., the coarse *alert signals* category contains four fine-level categories: *reverse beeper*, *car alarm*, *car horn*, *siren*. Recordings that are most similar to a small set of exemplary clips from YouTube for each sound class in our taxonomy were selected for annotation. We refer the interested reader to [3] for further details about the class taxonomy and the similarity measure used for data selection.

3. DATA COLLECTION

Since the release of SONYC-UST-V1, we have continued collecting audio recordings from our acoustic sensor network and Zooniverse volunteers have continued to annotate these recordings. SONYC-UST-V2 includes a total of 18510 annotated recordings from 56 sensors, a small sample of the 150M+ recordings that the SONYC acoustic sensor network has collected. The method for selecting which recordings to annotate has evolved over time. Initially, we sampled recordings as we did for V1, i.e., recordings that were most similar to a small set of exemplary clips from YouTube for each sound class in our taxonomy [3]. Subsequently, we sampled recordings using a batch-based active learning procedure in which a multi-label classifier was trained with all available annotations at that time. The model then predicted the class presence for unlabeled recordings, and recordings with class probabilities above a low fixed threshold were then clustered with minibatch k-means [7]. For each class, recordings were evenly sampled from each cluster to obtain a diverse sample, with more recordings sampled for classes with low representation in the dataset. Batch sizes typically varied between

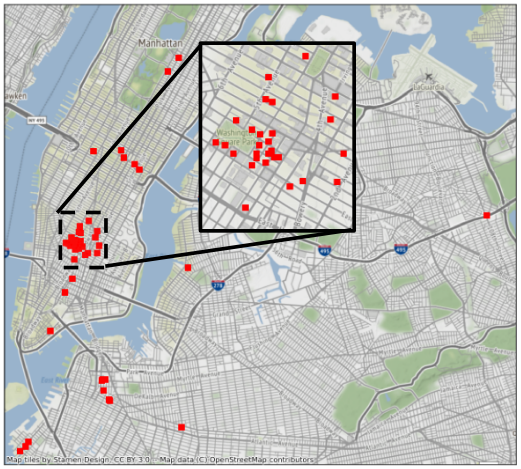

Figure 2: SONYC-UST-V2 sensor locations, many of which are in in Manhattan's Greenwich Village neighborhood (see inset).

1–2k recordings. We sampled the test set with yet another sampling procedure. For this set, a random sample of 10k recordings was selected from the set of unlabeled SONYC recordings. This was reduced to a diverse subset of 1k recordings selected with a determinantal point process (DPP) using the DPPy package [8] and OpenL3 embeddings [9] as the representation. This set was reduced further to adhere to our privacy criteria outlined in Section 4.

Each recording in SONYC-UST-V2 has been annotated by three different Zooniverse volunteers in the same manner as SONYC-UST-V1, i.e., on both the presence and proximity of the 23 fine-level and 8 coarse-level urban sound tags from the SONYC-UST Taxonomy [3].

As in SONYC-UST-V1, a subset of the recordings have annotations verified by the SONYC team in a two-step verification process. 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 Zooniverse volunteer. Two members of the SONYC team then 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. Lastly, a single SONYC team member listened to all of the recordings to ensure consistency across coarse-level categories and to catch any classes overlooked by the crowdsourced annotators. 1380 of the recordings have verified annotations—716 recordings from the SONYC-UST-V1 test and validation sets and 664 new recordings which comprise the SONYC-UST-V2 test set.

In SONYC-UST-V2 we continue our practice of defining training and validation sets that are disjoint by sensor and a test set that is temporally displaced to test generalization in a typical urban noise monitoring scenario. While the dataset contains recordings from 2016–2019, only the test set contains recordings from the latter two thirds of 2019. To capitalize on the effort put into the verified subsets in SONYC-UST-V1, we build upon the existing training and validation sensor split, growing each, while keeping the V1 split still intact. However, the SONYC-UST-V1 test set was not limited to the validation sensor split nor were subsequent crowdsourced an-

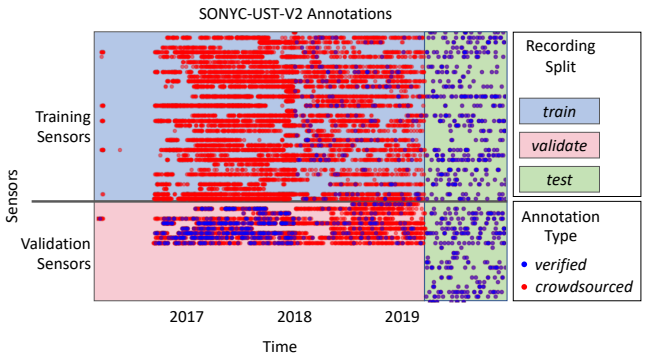

Figure 3: Dataset splits. The sensors in the test set overlap with both the training and validation sets. The test data is temporally dislocated from training and validation to test generalizability in time.

notations limited to recordings in the training sensor split. Thus, we now have verified annotations for recordings in the training sensor split and crowdsourced-only annotations for recordings in the validation sensor split, see Figure 3. All of this data has been included for completeness. However, when training the baseline model (see Section 6), we limit the training set to only the crowdsourced annotations in the training sensor split, and the validation set to only the verified annotations in the validation sensor split. See Figure 4 for the coarse-level class distribution of these recording splits.

Annotating urban sound recordings is a particularly difficult task. Sound events may be very distant with low signal-to-noise ratios, yet still audible. In addition, without visual verification, many sound events can be difficult to disambiguate. To capture this uncertainty, annotators are allowed to provide "incomplete" annotations, providing only the coarse-level class when they are unsure of the fine-level class (e.g. "Other/unknown engine"). Due to this difficult task, the inter-annotator agreement of the crowdsourced annotations as measured by Krippendorff's α [10] is rather low (0.36). Thus, SONYC-UST-V2 includes all of the individual crowdsourced and verified annotations, and we encourage users of the dataset to explore annotation aggregation strategies that model and incorporate annotator reliability. Since that is out of scope of this article, we use a simple approach of minority vote for our baseline model and analysis, i.e., a class is marked as present in the aggregate if at least one annotator marks it present. In previous work with Zooniverse annotators [11], we have found this strategy increases recall without significantly decreasing precision. In Table 1, we evaluate Zooniverse annotations aggregated with minority vote against the verified annotations in the test set using the metrics outlined in Section 5. These results are likely representative of good model performance when only a simplistic annotation aggregation method is used.

4. SPATIOTEMPORAL CONTEXT (STC) INFORMATION

The unique characteristic of this dataset is the inclusion of spatiotemporal context information, which informs where and when each example was recorded. To maintain privacy, we quantized the spatial information to the level of a city block, and we quantized the temporal information to the level of an hour. We also limited the occurrence of recordings with positive human voice annotations to one per hour per sensor. For the spatial information, we have

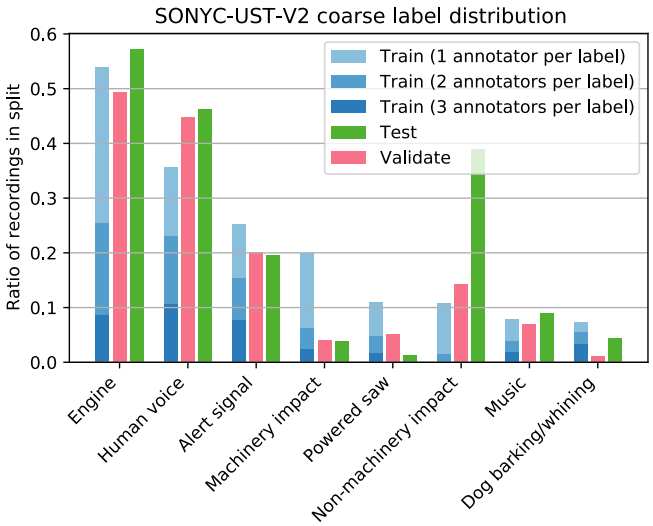

Figure 4: SONYC-UST tag distribution normalized for each recording split, in decreasing order of frequency in the training split. The shades of blue indicate how many annotators tagged the class in a training set recording, i.e., darker shades of blue indicate higher annotator agreement.

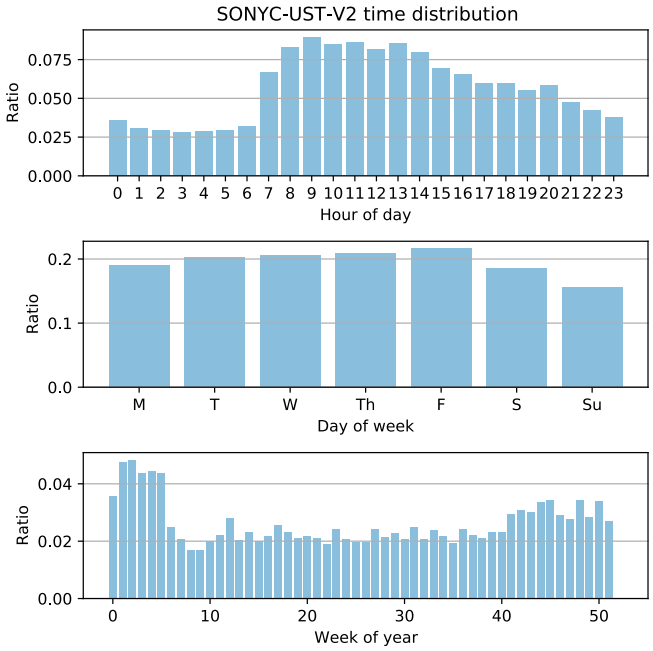

Figure 5: Distribution of dataset recordings per hour of the day, day of the week, and week of the year.

provided borough and block identifiers, as used in NYC's parcel number system known as Borough, Block, Lot (BBL) [12]. This is a common identifier used in NYC datasets, making it easy to relate the sensor data to other city data such as PLUTO [13] and more generally NYC Open Data [14], which contain information regarding land use, construction, transportation, noise complaints, and more. For ease of use with other datasets, we've also included the latitude and longitude coordinates of the center of the block. Figures 5 and 2 are distributions of the recordings in time and space.

5. EVALUATION METRICS

SONYC-UST-V2 includes labels at two hierarchical levels, coarse and fine (cf. [3] for details about the taxonomy), and models are evaluated independently against the labels at each level. Since some of the fine-level classes can be hard to label, even for human experts, a fraction of the samples in SONYC-UST-V2 only have coarse labels for some sound events. For example, a distant engine sound may be too ambiguous to label as a *small engine*, a *medium engine* or a *large enging* (i.e., fine labels), but can still tagged with the coarse label *engine of uncertain size*. For such cases, we use a tag coarsening procedure that leverages the hierarchical relationship between the fine and coarse labels in our taxonomy to obtain performance estimates for fine labels in the face of annotator uncertainty (cf. [3] for further details about this procedure).

For each of the two levels, we compute three metrics: macro-averaged AUPRC, micro-averaged AUPRC, and label-weighted label-ranking average precision (LWLRAP) [15]. We use the first as the primary performance metric, and the second as a secondary metric to gain further insight into the performance of each system. Macro-averaged AUPRC provides a measure of performance across all classes independently of the number of samples per class, while micro-averaged AUPRC is sensitive to class imbalance.

Finally, LWLRAP measures the average precision of retrieving a ranked list of relevant labels for each test clip. It is a generalization of the mean reciprocal rank measure for evaluating multi-label classification, which gives equal weight to each label in the test set (as opposed to each test clip). The metric has been widely adopted in the DCASE community over the past year.

6. BASELINE SYSTEM

For the baseline model [3], we use a multi-label multi-layer perceptron model, using a single hidden layer of size 128 (with ReLU non-linearities), and using AutoPool [16] to aggregate frame level predictions. The model takes in as input *audio content*, *spatial context*, and *temporal context*.

Audio content is given as OpenL3 [9] embeddings (with `content_type="env"`, `input_repr="mel256"`, and `embedding_size=512`), using a window size and hop size of 1.0 second (with centered windows), giving us 11 512-dimensional embeddings for each clip in our dataset. Spatial context is given as latitude and longitude values, giving us two values for each clip in our dataset. Temporal context is given as hour of the day, day of the week, and week of the year, each encoded as a one hot vector, giving us 83 values for each clip in our dataset. We z-score normalize the embeddings, latitude, and longitude values, and concatenate all of the inputs (at each time step), resulting in an input size of 597.

[3] https://github.com/sonyc-project/dcase2020task5-uststc-baseline

Estimator:	Annotators		Model w/ STC		Model w/o STC	
Level:	F	C	F	C	F	C
Overall						
Macro-AUPRC	0.56	0.69	0.44	0.49	0.43	0.49
Micro-AUPRC	0.60	0.75	0.62	0.71	0.62	0.71
LWLRAP	0.62	0.78	0.72	0.83	0.73	0.83
AUPRC						
Engine	0.57	0.82	0.57	0.84	0.59	0.84
Mach. imp.	0.35	0.48	0.19	0.32	0.18	0.30
Non-mach. imp.	0.60	0.60	0.58	0.60	0.59	0.61
Powered saw	0.14	0.37	0.16	0.11	0.12	0.12
Alert signal	0.74	0.82	0.45	0.40	0.44	0.39
Music	0.53	0.75	0.41	0.52	0.41	0.54
Human voice	0.78	0.91	0.88	0.92	0.88	0.93
Dog	0.79	0.79	0.26	0.22	0.24	0.23

Table 1: The performance of the Zooniverse annotations (using minority vote aggregation) and the baseline classifier with and without STC as compared the the ground-truth annotations for the test split on the coarse (C) and fine (F) levels.

We use the weak tags for each audio clip as the targets for each clip. For the training data (which has no verified target), we count a positive for a tag if at least one annotator has labeled the audio clip with that tag (i.e., minority vote). Note that while some of the audio clips in the training set have verified annotations, we only use the crowdsourced annotations. For audio clips in the validation set, we only use annotations that have been manually verified.

We train the model using stochastic gradient descent to minimize the binary cross-entropy loss, using L^2 regularization (weight decay) with a factor of 10^{-5}. For training models to predict tags at the fine level, we modify the loss such that if "unknown/other" is annotated for a particular coarse tag, the loss for the fine tags corresponding to this coarse tag are masked out. We train for up to 100 epochs, using early stopping with a patience of 20 epochs using loss on the validation set. We train one model to predict fine-level tags, with coarse-level tag predictions obtained by taking the maximum probability over fine-tags predictions within a coarse category. We train another model only to predict coarse-level tags.

Table 1 presents the results of the baseline model trained with and without spatiotemporal context. The baseline model's performance is quite low and does not seem to benefit from the inclusion of STC. However, its inclusion of STC and its aggregation of annotations are both rather naive. We hope this simply provides a starting point for researchers to explore more sophisticated approaches that better leverage the unique aspects of this data and incorporate additional contextual data to aid in generalizability.

7. CONCLUSIONS

SONYC-UST-V2 is a multi-label dataset for urban sound tagging with spatiotemporal context information. It consists of 18510 audio examples recorded in New York City between 2016 and 2019 with weak (i.e., tag) annotations on urban sound classes, as well as metadata on where and when each audio example was recorded. We believe STC is a rich source of information for sound tagging that has yet to be adequately explored and could potentially aid models in the challenging task of tagging real-world urban sound recordings. This dataset is the first of its kind that we are aware of and will provide researchers with material for exploring the incorporation of spatiotemporal context (STC) information into sound tagging.

8. REFERENCES

[1] S. Beery, G. Wu, V. Rathod, R. Votel, and J. Huang, "Context r-cnn: Long term temporal context for per-camera object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 075–13 085.

[2] M. Cartwright, J. Cramer, J. Salamon, and J. P. Bello, "Tricycle: Audio representation learning from sensor network data using self-supervision," in *Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 278–282.

[3] 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 *Proceedings of the 2019 Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Oct. 2019, pp. 35–39.

[4] J. P. Bello, C. Silva, O. Nov, R. L. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[5] 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.

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

[7] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 1177–1178.

[8] G. Gautier, G. Polito, R. Bardenet, and M. Valko, "DPPy: DPP Sampling with Python," *Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS)*, 2019, code at http://github.com/guilgautier/DPPy/ Documentation at http://dppy.readthedocs.io/. [Online]. Available: http://jmlr.org/papers/v20/19-179.html

[9] J. Cramer, H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[10] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[11] 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 ACM CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 292.

[12] "Borough, block, lot lookup." [Online]. Available: https://portal.311.nyc.gov/article/?kanumber=KA-01247

[13] "Property Land Use Tax lot Output." [Online]. Available: https://www1.nyc.gov/site/planning/data-maps/open-data.page

[14] "NYC Open Data." [Online]. Available: https://opendata.cityofnewyork.us/

[15] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 69–73.

[16] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 11, p. 2180–2193, Nov. 2018. [Online]. Available: https://doi.org/10.1109/TASLP.2018.2858559