A Dataset and Taxonomy for Urban Sound Research

Justin Salamon[1,2], Christopher Jacoby[1], Juan Pablo Bello[1]
[1]Music and Audio Research Laboratory, New York University
[2]Center for Urban Science and Progress, New York University
{justin.salamon, cbj238, jpbello}@nyu.edu

ABSTRACT

Automatic urban sound classification is a growing area of research with applications in multimedia retrieval and urban informatics. In this paper we identify two main barriers to research in this area – the lack of a common taxonomy and the scarceness of large, real-world, annotated data. To address these issues we present a taxonomy of urban sounds and a new dataset, *UrbanSound*, containing 27 hours of audio with 18.5 hours of annotated sound event occurrences across 10 sound classes. The challenges presented by the new dataset are studied through a series of experiments using a baseline classification system.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing; H.5.5 [**Information Systems**]: Sound and Music Computing

Keywords

Urban sound; dataset; taxonomy; classification

1. INTRODUCTION

The automatic classification of environmental sound is a growing research field with multiple applications to large-scale, content-based multimedia indexing and retrieval (e.g. [13, 16, 10]). In particular, the sonic analysis of *urban* environments is the subject of increased interest, partly enabled by multimedia sensor networks [15], as well as by large quantities of online multimedia content depicting urban scenes. However, while there is a large body of research in related areas such as speech, music and bioacoustics, work on the analysis of urban acoustics environments is relatively scarce. Furthermore, when existent, it mostly focuses on the classification of auditory scene type, e.g. street, park, as opposed to the identification of sound sources in those scenes, e.g. car horn, engine idling, bird tweet. See [5] for an example.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2655045.

One of the main challenges and hindrances to urban sound research is the lack of labeled audio data. Previous work has focused on audio from carefully produced movies or television tracks [3], from specific environments such as elevators or office spaces [13, 8]; and on commercial or proprietary datasets [11, 4]. The large effort involved in manually annotating real-world data means datasets based on field recordings tend to be relatively small (e.g. the event detection dataset of the IEEE AASP Challenge [8] consists of 24 recordings per each of 17 classes). A second challenge faced by the research community is the lack of a common vocabulary when working with urban sounds. This means the classification of sounds into semantic groups may vary from study to study, making it hard to compare results.

The goal of this paper is to address the two aforementioned challenges. In Section 2 we propose a taxonomy for urban sound sources to facilitate a common framework for research. Then, in Section 3 we present *UrbanSound*, a dataset of 27 hours of field recordings containing thousands of labeled sound source occurrences. To the best of the authors' knowledge this is the largest free dataset of labelled urban sound events available for research. To understand the complexity and challenges presented by this new dataset, we run a series of baseline sound classification experiments, described in Section 4. The paper concludes with a summary in Section 5.

2. URBAN SOUND TAXONOMY

The taxonomical categorization of environmental sounds, a common first step in sound classification, has been extensively studied in the context of perceptual soundscape research [14]. Specific efforts to describe urban sounds have often been limited to subsets of broader taxonomies of acoustic environments [2], and thus only partially address the needs of systematic urban sound analysis. For an exhaustive review of previous work the reader is referred to [12].

In our view, an urban sound taxonomy should satisfy the following three requirements: (1) it should factor in previous research and proposed taxonomies, (2) it should aim to be as detailed as possible, going down to low-level sound sources such as "car horn" (versus "transportation") and "jackhammer" (versus "construction"), (3) it should, in its first iteration, focus on sounds that are of specific relevance to urban sound research, such as sounds that contribute to urban noise pollution. To address (1), we decided to base our taxonomy on the subset of [2] dedicated to the urban acoustic environment. We define 4 top level groups: human, nature, mechanical and music, which are common to most previ-

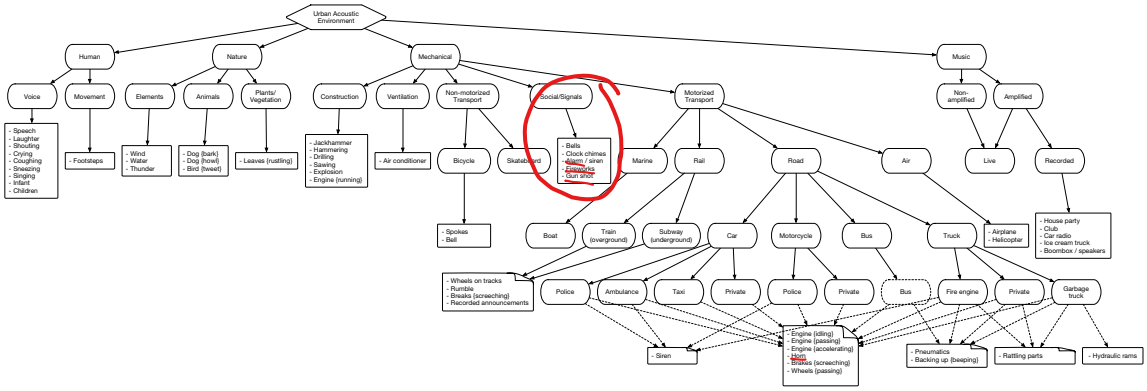

Figure 1: Urban Sound Taxonomy.

ously proposed taxonomies [12]. To address (2), we specify that the leaves of the taxonomy should be sufficiently low-level to be unambiguous – e.g. car "brakes", "engine" or "horn", instead of simply "car". Finally to address (3), we examined all the noise complaints filed through New York City's 311 service from 2010 to date[1] (over 370,000), and built the taxonomy around the most frequently complained about sound categories and sources – construction (e.g. jackhammer), traffic noise (car and truck horns, idling engines), loud music, air conditioners and dog barks to name a few.

The resulting taxonomy is provided in Figure 1. Further information about the principles and process behind the construction of the taxonomy, as well as a scalable digital version, are available online[2]. Rounded rectangles represent high-level semantic classes (e.g. mechanical sounds). The leaves of the taxonomy (rectangles with sharp edges) correspond to classes of concrete sound sources (e.g. siren, footsteps). For conciseness, leaves can be shared by several high-level classes (indicated by an earmark). Since the number of possible sound sources in an urban setting is very large (potentially infinite), we consider the taxonomy to be a constant work in progress rather than fixed. We plan to continue expanding and reformulating the taxonomy as we increase the scope of sounds covered by our research, by engaging the international research community and promoting a collaborative effort via (for instance) dedicated workshops.

3. THE URBANSOUND DATASET

In addition, we have collected a dataset of annotated urban sounds including 10 low-level classes from the taxonomy: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music. With the exception of "children playing" and "gun shot" which were added for variety, all other classes were selected due to the high frequency in which they appear in urban noise complaints, as noted in the previous section. The dataset is called *UrbanSound*. Due to the manual annotation effort required (see below) we limited the number of classes to 10, which we consider a good starting point. We intend to extend it to more classes in future iterations. For a review of existing datasets and related literature cf. footnote 2.

Before compiling the dataset, we set three main goals: (1) it should contain sounds that occur in an urban environment, (2) all recordings must be real field-recordings, (3) the dataset should be sufficiently large and varied in terms of sounds and recording conditions such that it will be useful for training scalable algorithms capable of analyzing real data from sensor networks or multimedia repositories. To accomplish these goals, we turned to Freesound[3], an online sound repository containing over 160,000 user-uploaded recordings under a creative commons license. Freesound contains a large amount of field recordings, many of which are in urban settings. Using the Freesound API we were able to search and download a subset of the repository, and exploit the user-provided metadata (title, description and tags) to significantly speed up the annotation process.

For each class, we started by downloading all sounds returned by the Freesound search engine when using the class name as a query (e.g. "jackhammer"), resulting in over 3000 recordings summing to just over 60 hours of audio. We then manually checked every recording by listening to it and inspecting the user-provided metadata, only keeping those that were actual field recordings where the sound class of interest was present somewhere in the recording. After this first filtering stage we were left with 1314 recordings summing to just over 27 hours of audio. Next, given all the recordings for a specific sound class, we used Audacity[4] to label the start and end times of every occurrence of the sound in each recording, with an additional *salience* description indicating whether the occurrence was subjectively perceived to be in the foreground or background of the recording. This resulted in a total of 3075 labeled occurrences amounting to 18.5 hours of labeled audio. The distribution of total occurrence duration per class and per salience is in Fig. 2(a).

The resulting collection of 1314 full length recordings with corresponding sound occurrence and salience annotations, *UrbanSound*, is freely available online (cf. footnote 2) for research purposes. The audio is provided in the same format in which it was originally uploaded to Freesound. Note that the duration of source occurrences in the set can vary from 1-2 s (e.g. gun shot sounds) to over 30 s (continuous sounds such as jackhammers or idling engines).

For research on sound source identification, we created an additional subset of short audio snippets which we call the *UrbanSound8K* subset, also available online. In [5] the authors conducted a listening test and found that 4 seconds were sufficient for subjects to identify environmental sounds with 82% accuracy, and consequently use a 4 s clip duration in their experiments with automatic classification. Following

[1] https://nycopendata.socrata.com/data
[2] http://serv.cusp.nyu.edu/projects/urbansounddataset/
[3] http://www.freesound.org
[4] http://audacity.sourceforge.net/

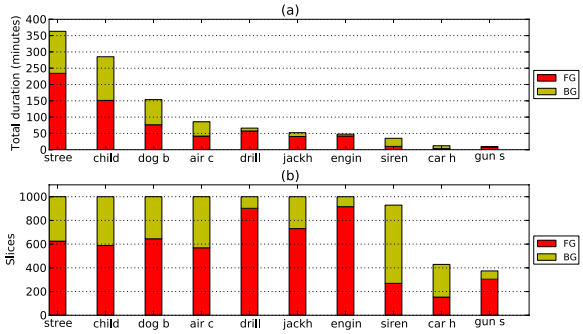

Figure 2: (a) Total occurrence duration per class in UrbanSound. (b) Slices per class in UrbanSound8K. Breakdown by foreground (FG) / background (BG).

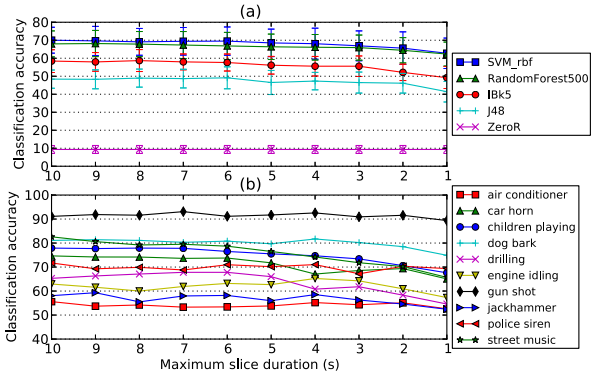

Figure 3: Classification accuracy vs maximum slice duration: (a) by classifier, (b) by class for SVM.

their findings, we set a maximum occurrence duration limit of 4 seconds, and segment longer occurrences into 4 s slices using a sliding window with a hop size of 2 s. To avoid large differences in the class distribution, we set a limit of 1000 slices per class, resulting in a total of 8732 labeled slices (8.75 hours). The distribution of slices per class in UrbanSound8K with a breakdown into salience is provided in Figure 2(b).

4. SOUND CLASSIFICATION

In order to learn about the characteristics and challenges presented by this new dataset, we run a set of classification experiments using a baseline approach. Note that we are not searching for an optimal combination of feature/classifier parameters to maximize accuracy, but rather are interested in learning about the characteristics of the dataset itself.

4.1 Feature extraction

In all of the following experiments, we extract Mel-Frequency Cepstral Coefficients (MFCC) from the audio slices using the Essentia audio analysis library [1]. MFCCs are commonly used in environmental sound analysis (including recent work [4]) and frequently used as a competitive baseline to benchmark novel techniques [7]. In all experiments we extract the features on a per-frame basis using a window size of 23.2 ms and 50% frame overlap. We compute 40 Mel bands between 0 and 22050 Hz and keep the first 25 MFCC coefficients (we do not apply any pre-emphasis nor liftering). The per-frame values for each coefficient are summarized across time using the following summary statistics: minimum, maximum, median, mean, variance, skewness, kurtosis and the mean and variance of the first and second derivatives, resulting in a feature vector of dimension 225 per slice.

4.2 Experimental setup

To experiment with different classification algorithms we use the Weka data mining software [9]. Every experiment is run using 10-fold cross validation. Within each fold we perform correlation-based attribute selection to avoid overfitting the training data. As it is not our goal to find an optimal parametrization, all classification algorithms are used with their default parameter settings. For each experiment we report the average accuracy across all 10 folds.

Important care must be taken when creating the folds – since there are multiple slices coming from the same original recording, if we generate the folds completely randomly, we may end up with slices from the same recording used both for training and testing, which can lead to artificially high classification accuracies. To avoid this, we designed a random allocation process of slices into folds such that all slices originating from the same Freesound recording go into the same fold, whilst trying to balance the number of slices-per fold for each sound class. The UrbanSound8K subset available online provides the audio slices grouped into 10 folds generated using this methodology. In this way, researchers interested in comparing their results to our baseline are guaranteed unbiased and comparable results.

4.3 Results

In Section 3 we motivated our choice of 4 s as the maximum slice duration for UrbanSound8K. In this first experiment we examine how the choice of this threshold affects the performance of the baseline approach. To this end, we generated 10 copies of UrbanSound8K, each time varying the maximum slice duration from 10 s down to 1 s. To ensure the observed changes in accuracy are not an artifact of a specific classification algorithm, we compare 5 different algorithms: decision tree (J48), k-NN ($k = 5$), random forest (500 trees), support vector machine (radial basis function kernel), and a baseline majority vote classifier (ZeroR).

The results are presented in Figure 3(a). The difference in performance between all classifiers is statistically significant (paired t-test with $p < 0.001$) except for the top two (SVM and random forest). More importantly, we observe consistent behavior across all classifiers – performance remains stable from 10 to 6 s, after which it starts decreasing gradually. However, if we consider the top performing classifier (SVM), there is no statistically significant difference between performance using 6s slices and 4s slices (whereas below 4s the difference becomes significant), in accordance with [5] and supporting the choice of 4 s slices for UrbanSound8K. Further insight can be gained by observing the per-class accuracies for the SVM, provided in Figure 3(b). We see that different sound classes are affected differently by the slice duration: classes such as gun shot and siren have fast events that are clearly identifiable at short temporal scales and are thus mostly unaffected by duration; while classes such as street music and children playing drop almost monotonically, showing the importance of analyzing them at longer temporal scales and suggesting multi-scale analysis could be a relevant path for urban sound research.

To understand the relative difference in performance between classes, we examined the confusion matrix for the SVM classifier on UrbanSound8K. We found that the clas-

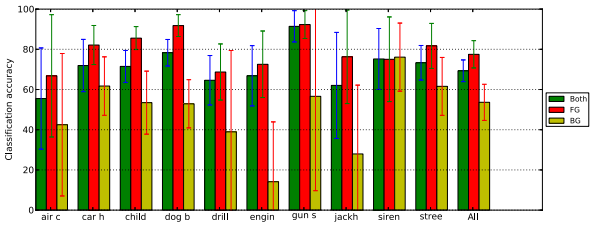

Figure 4: Accuracy as a function of salience.

sifier mostly confuses three pairs of classes: air conditioners and idling engines, jackhammers and drills, children and street music. This makes sense, as the timbre of each pair is quite similar (for the last pair the presence of complex harmonic tones is a possible cause). We see that the baseline bag-of-frames approach based on MFCCs fails especially on wide-band noise-like continuous sounds, and we intend to investigate approaches that better model the temporal dynamics of energy and timbre as part of future work.

However, the relative similarity between sound classes is only part of the story. As explained in Section 3, every sound occurrence in the dataset is also labeled with a (subjective) salience label – foreground (FG) and background (BG, also used to label occurrences where there are other equally-salient sources). Intuitively, one would expect the baseline algorithm to do better on slices where there is less background interference, especially since MFCCs are known to be sensitive to noise [6]. In this final experiment, we compare the performance of the SVM classifier for each class as a function of salience, displayed in Figure 4. As expected, we see a considerable difference in performance between sounds labeled FG and BG, the exception being sirens, possibly because their frequency content does not overlap with most other sources (by design). Whilst we cannot quantify the effect of interference from this experiment due to the subjectivity of the labeling, the results point to an important challenge presented by the dataset – identifying sound sources in the presence of (real) background noise. This problem is an active area of research (e.g. [6]), and we believe this real-world dataset will further empower the research community in coming up with novel solutions to this problem.

5. SUMMARY

Automatic urban sound classification can benefit a variety of multimedia applications. In this paper we identified two main barriers to research in this area – the lack of a common taxonomy and the scarceness of large, real-world, annotated data. To address the first issue we presented the Urban Sound Taxonomy, based on previous soundscape research with a focus on sound classes from real noise-complaint data. To address the second issue we presented UrbanSound, a dataset containing 27 hours of audio with 18.5 hours of manually labelled sound occurrences. We also presented UrbanSound8K, a subset of the dataset designed for training sound classification algorithms. Through a series of classification experiments we studied the challenges presented by the dataset, and identified avenues for future research: sensitivity to temporal scale in the analysis, confusion due to timbre similarity (especially for noise-like continuous sounds), and sensitivity to background interference. We believe the dataset will open the path to new an exciting research in sound and multimedia applications with a focus on urban environments and urban informatics.

6. ACKNOWLEDGMENTS

This work was supported by a seed grant from New York University's Center for Urban Science and Progress (CUSP).

7. REFERENCES

[1] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra. ESSENTIA: an open-source library for sound and music analysis. In *21st ACM Int. Conf. on Multimedia*, pages 855–858, 2013.

[2] A. L. Brown, J. Kang, and T. Gjestland. Towards standardization in soundscape preference assessment. *Applied Acoustics*, 72(6):387–392, 2011.

[3] L.-H. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai. A flexible framework for key audio effects detection and auditory context inference. *IEEE TASLP*, 14(3):1026–1039, 2006.

[4] S. Chaudhuri and B. Raj. Unsupervised hierarchical structure induction for deeper semantic analysis of audio. In *IEEE ICASSP*, pages 833–837, 2013.

[5] S. Chu, S. Narayanan, and C.-C. Kuo. Environmental sound recognition with time-frequency audio features. *IEEE TASLP*, 17(6):1142–1158, 2009.

[6] C. V. Cotton and D. P. W. Ellis. Spectral vs. spectro-temporal features for acoustic event detection. In *IEEE WASPAA'11*, pages 69–72, 2011.

[7] D. P. W. Ellis, X. Zeng, and J. H. McDermott. Classifying soundtracks with audio texture features. In *IEEE ICASSP*, pages 5880–5883, 2011.

[8] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley. A database and challenge for acoustic scene classification and event detection. In *21st EUSIPCO*, 2013.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[10] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen. Audio context recognition using audio event histograms. In *18th EUSIPCO*, pages 1272–1276, 2010.

[11] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen. Context-dependent sound event detection. *EURASIP JASMP*, 2013(1), 2013.

[12] S. R. Payne, W. J. Davies, and M. D. Adams. Research into the practical and policy applications of soundscape concepts and techniques in urban areas. DEFRA, HMSO, London, UK, 2009.

[13] R. Radhakrishnan, A. Divakaran, and P. Smaragdis. Audio analysis for surveillance applications. In *IEEE WASPAA'05*, pages 158–161, 2005.

[14] R. M. Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books, 1993.

[15] D. Steele, J. D. Krijnders, and C. Guastavino. The sensor city initiative: cognitive sensors for soundscape transformations. In *GIS Ostrava*, pages 1–8, 2013.

[16] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo. Audio keywords generation for sports video analysis. *ACM TOMCCAP*, 4(2):1–23, 2008.