Towards Soundscape Information Retrieval (SIR)

Tae Hong Park[1], Jun Hee Lee[1], Jaeseong You[1], Min-Joon Yoo, John Turner[1]

[1] Music and Audio Research Lab (MARL)
The Steinhardt School
New York University
New York, NY 10012 USA

[2] Computer Science Department
Yonsei University
Seoul, South Korea

{thp1, junheelee, jaeseongyou, minjoon.yoo, jmt508}@nyu.edu

ABSTRACT

In this paper we discuss our efforts in Soundscape Information Retrieval (SIR). Computational soundscape analysis is a key research component in the *Citygram Project* which is built on a cyber-physical system that includes a scalable robust sensor network, remote sensing devices (RSD), spatio-acoustic visualization formats, as well as software tools for composition and sonification. By combining our research in soundscape studies, which includes the capture, collection, analysis, visualization and musical applications of spatio-temporal sound, we discuss our current research efforts that aim to contribute towards the development of soundscape information retrieval (SIR). This includes discussion of soundscape descriptors, soundscape taxonomy, annotation, and data analytics. In particular, we discuss one of our focal research agendas in measuring and quantifying urban noise pollution.

1. INTRODUCTION

Some of the most complex sound environments are *soundscapes*, a term coined, and a field championed, by R. Murray Schafer. Soundscapes and acoustic ecology go hand-in-hand and many composers have engaged in soundscape composition either directly through strict, unaltered playback of field recordings; indirectly through sound synthesis interpretation; or via the creation of hybrid soundscape compositions where field recordings are processed and other "external" sound materials are introduced in works such as *Presque rien, numéro 1* (1970), *Riverrun* (1986), *Ride* (2000), and *48 13 N, 16 20 O* (2004). Soundscape studies as a computational research field, however, are still in their early stages. This is especially the case when considering it is compared to speech recognition and music information retrieval (MIR). This may be due to a number of factors including the lack of datasets for training and development, overwhelming emphasis on speech recognition [1, 2], and the complexities surrounding soundscapes – literally any sound can exist in a soundscape, making this unconstrained sound classification task extremely difficult [3]. That is not to say that research in SIR is not vibrant. As a matter of fact, research papers related to music, speech, and environmental sound tagging has increased from approximately 10 in 2003 to over 45 in 2010 [1]. Also, numerous SIR research examples exist including projects related to surveillance [4], bird species [5], traffic sounds [6], and gunshot detection [7]. Much fundamental research still has to be conducted, including topics concerning the taxonomy and vocabulary of soundscapes, dataset development, and creation of robust models that can be used to adapt to the vastly diverse soundscapes ranging from outdoor spaces such as urban environments, marshlands, tropical forests, woodlands, and Saharan deserts; to indoor spaces including offices, train stations, shopping malls, and sports arenas. Our research in soundscape currently focuses on a small subset of soundscapes: urban noise and possibilities for musical applications.

In 2011, the *Citygram Project* [8]–[10] was launched to develop dynamic non-ocular energy maps focusing on acoustic energy in its first iteration. Since the project's inception, two of its driving forces have been acoustic ecology and soundscape research from both a "technical research" perspective as well as a musical application perspective. The former has centered on source capture and identification, and the latter, on engaging in real-time spatio-acoustic music interaction. More recently, in collaboration with New York University's Center for Urban Science and Progress (CUSP) and the *Sound Project*, noise pollution has become a focal point of our soundscape research inquiry. In this paper, we present an overview of our efforts in contributing of the field we call Soundscape Information Retrieval (SIR). This includes a number of core components: (1) sound semantics, (2) sound annotation, (3) sound analysis tools, and (4) machine learning (ML).

2. SOUND ANALSYSIS TOOLBOX (SATB)

To facilitate our analysis efforts we are currently developing the Sound Analysis Toolbox (SATB) written in MATLAB. The system aims to provide a comprehensive platform for sound/semantic analysis, visualization, algorithmic development, baseline ML exploration using Weka [28], and basic audio transport features. In this section we present the basic components of SATB and detail further utilization of its features in semantic analysis and AED/AEC below.

Our current implementation includes a "quick plot" feature that allows efficient plotting of large sound files–a feature that is limited using MATLAB's default `plot` function. The quick plot function uses an efficient proprietary "min-max" envelope contour computation algorithm that allows for quick plotting, zooming, and 3D visualizations. Additionally, SATB includes a simple "plug-in" feature for adding custom feature extraction algorithms and signal processing implementations. This

is accomplished by inheriting custom MATLAB classes with system methods that are called from the SATB controller class. SATB is essentially a major revision and improvement of the EASY Toolbox [29], where SATB allows for more comprehensive exploration of all types of sounds, including soundscapes.

2.1 Freesound MATLAB API

A MATLAB Freesound API is also included in SATB. This module is used for pulling queried sound files and associate metadata and includes functions for tag querying, downloading/saving audio files with associated metadata, checking for corrupt audio files, and formatting audio channels and sampling rates. The Freesound API will serve as a model for developing our own Citygram MATLAB API.

2.2 Audio Transport

Comprehensive sound analysis software systems require synchronization between audio and visualizations. MATLAB, however, offers limited support in this area: synchronizing audio to visualizations and playing long audio files is impractical. Although MATLAB provides its Data Acquisition Toolbox, this is only available for the Windows operating system. To address these shortcomings we have adopted the open-source PsychToolbox [30] to access native audio hardware methods from MATLAB. In order to play large audio files without memory concerns and to synchronize dynamic visualization, a double audio-buffering scheme has been implemented using PsychPortAudio's playback scheduling feature. This mechanism allows for synchrony between audio output and dynamic visualizations such as waveforms, feature vector plots, and spectrograms where cursor positions are synched to the current audio output sample.

3. SOUNDSCAPE SEMANTICS

One of the key components of the Citygram Project is the exploration of acoustic ecology. As part of our urban sound classification efforts, we have begun developing a number of software tools for sound analysis and visualization; machine learning modules for acoustic event detection (AED) and acoustic event classification (AEC); development of annotated datasets; and tools for soundscape taxonomy exploration. On one hand, our research involves the investigation of acoustic ecology studies that are in resonance with the Schaferian school of thought [11] where the concepts include the identity of the sound source, the notion of *keynote* (definite background sounds), *signal* (foreground sound), *soundmarks* (culturally/symbolically important within a community), *geophony* (natural sound sources), *biophony* (non-human, non-domestic biological sources), and *anthrophony* (human-generated sounds). This is conceptually similar to what Gaver refers to as *everyday listening* opposed to *musical listening* [12]. On the other hand, we also concentrate in research that is in the realm of Big Data science where waves of spatio-acoustic data are collected to develop DSP, feature extraction, and machine learning techniques for urban soundscape analysis. Big Data is one of the "hottest" topics in data analytics today, and in a sense, the notion of *found data*[1] is quite fitting when viewed from the *found sound* and *musique concrète* perspective: a perspective where the data itself is the focus and starting point into research inquiry. In this section we discuss issues related to semantics of urban soundscapes.

3.1 Urban Soundscapes are Noisy

One of the sounds we are interested in automatically capturing is urban acoustic noise. Our recent collaborative efforts with NYU CUSP has made this focus an especially intriguing one as urban noise pollution is a major problem for city-dwellers around the world including New York City (NYC). For instance, since the creation of the NYC non-emergency 311-hotline in 2010, the largest number of complaints has been noise. The urgency in developing mechanisms and technologies to measure, map, and help mitigate noise pollution, and thus improve the living conditions of urban communities is not difficult to imagine when we consider that 68% of the global population is projected to live in so-called megacities. While issues such as *noise annoyance* [13] have to be considered in noise research, fundamental technical issues in capturing noise have to be addressed as well. Simply employing dB sound pressure level measurements is inadequate [11, 12] as both spectral and temporal acoustic dimensions have to be considered. For example, heavy rain measured at 90 dB SPL is experienced very differently to scratching a blackboard with fingernails at the same level. The first step in defining noise involves the measurement of spatial sounds from whence acoustic noise can be identified. Another step includes the development of nomenclature, an "agreed-upon" acoustic noise taxonomy that reflects soundscapes, which can then be used for automatic soundscape classification. The initial first step of capturing spatio-temporal sound is enabled by the creation of a dense sensor network, a goal that the Citygram infrastructure aims to accomplish. Various aspects of the aforementioned steps are discussed in the following sections.

3.2 Sound Semantics and Taxonomy

A key element in supervised machine learning is the requirement of large human-annotated datasets. The problem of automatic instrument classification in Classical music, for example, has a clearly defined search space: most, if not all, acoustic instruments and their associated names (classes) are known and thus easily labeled, and annotated datasets are available in abundance. For non-classical music genres such as popular music, the instrumental taxonomical space becomes less clear due to the introduction of electronic instruments and for electro-acoustic music, the ambiguity further increases and it is not uncommon to find diverse variance in sound nomenclatures for describing similar/same instruments/sounds. For urban sound taxonomy the question of what exists,

[1] http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2yJGerxRG

what we hear, and how we annotate and label is an interesting problem in itself. Furthermore, developing a "standard" urban soundscape taxonomy and vocabulary is difficult in part due to its tremendous sonic variety, the dominance of vision in information processing, and an emphasis on speech signals [16, 18]. It is easy enough to imagine the usual urban noise suspects including sirens, jackhammers, garbage trucks, music, dog barks, and car horns. But things quickly become murky once the *entire* soundscape is considered. It becomes even more ambiguous when we begin to consider what sound sources are *perceived* as "noisy."

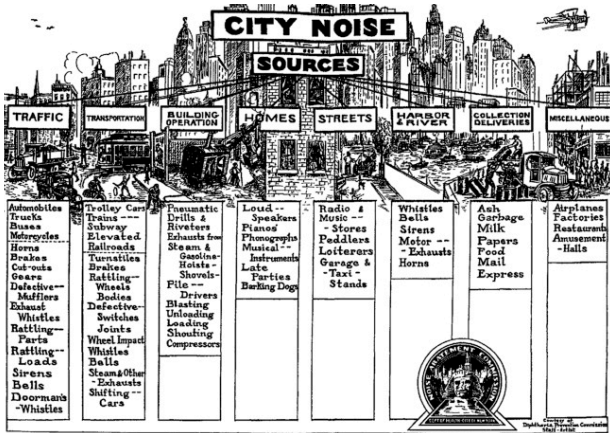

Figure 1. City noise sources from 1930s New York.

A number of soundscape taxonomies exist. One of the earliest was published in 1913 by Luigi Russolo, which is articulated in his *The Art of Noises* manifesto [20]. Other composers who have developed soundscape related taxonomies include John Cage. In *Williams Mix* (1952), Cage discusses the sound classes that are labeled as *city sounds*, *country sounds*, *wind-produced sounds*, and *electronic sounds*. Some 14 years later, Stockhausen developed his own intricate catalogue of sound class nomenclature–although not exclusively addressing soundscapes but rather *moments* – consisting of 68 labels for noise that included *whirring*, *crackling*, *rustling*, *clapping*, *clanking*, *falling*, and *thundering*. Another soundscape taxonomy example can be seen in [21], produced as part of the Noise Abatement Commission of New York. In this study, a "noise truck" logged over 500 miles and collected 10,000 measurements from over 18 locations as shown in Figure 1 [22]. More recent examples include work by Gaver [12] and Brown [23], where the former presents the idea of basic-level sound-producing events: *liquids*, *vibrating objects*, and *aerodynamic sounds* as the basis for mapping environmental sounds such as passing vehicles, motorboats, and lakes. Brown's taxonomy is more rigid and uses a tree-branch-leaf structure with clear categorical divisions where top branches are more general and bottom leaves are most specific. Although strict "standardized" taxonomies can be helpful when beginning to explore soundscapes and associated hierarchical semantic labels, they can also be biased, reflecting the opinions, priorities, and interests of the researchers devising them which may not necessarily reflect general public consensus [24]. For example, in Brown's taxonomy, a bifurcation between amplified and non-amplified urban sound sources exists, a distinction that can arguably be difficult to make.

3.2.1 Mining Collective Listening

In the field of AED and AEC, ML-based algorithms typically classify audio events using ground-truth data: after defining a limited set of semantic labels, feature vectors are used as inputs to train ML algorithms. The trained algorithms then attempt to classify new sound input to its proper class. However, soundscape-based semantic labels and tags developed by researchers do not necessarily reflect a collective consensus. Conversely, crowdsourcing the annotation process may offer an auxiliary mechanism for a more robust repository of sonic semantics, and reverses this notion of "annotation by decree." This approach is in resonance with developing soundscape semantics via open-ended labeling and surveying methodologies [18, 25]. Inviting researchers *and* a larger community to define and refine the pool of semantic concepts in relation to novel sonic inputs can potentially contribute to a more agreed upon soundscape taxonomy. Furthermore, using crowdsourcing for taxonomical development may yield more than an expanded tag-pool for labeling audio events: it can potentially reveal connectivity between sounds and everyday concepts as defined by collective consensus. As such, we are taking initial steps in using Big Data mining of audio semantics to reveal insights into transforming subjective, qualitative associations between sounds and concepts into a quantifiable and communicable format. Of utmost importance in this approach is to ensure that the collected data is sufficiently large enough to develop a robust taxonomy. In determining the feasibility of developing a collective listening taxonomy, we are currently using a custom MATLAB API to pull crowdsourced audio files and its associated annotations from Freesound[2] as further described in Section 3.1. Subsequent collective listening exploration will entail mining sono-semantics from other existing datasets such as the NYC Open Data (noise complaint records) and the World Wide Web itself using keyword search strategies.

3.3 Development of Datasets and Ground Truth

One of the issues with soundscape-based ground-truth dataset is its accessibility and availability: existing annotated datasets are difficult to find and often focus on indoor environmental sounds [26]. A free online sound repository we have found very useful is Freesound. Freesound is an incredibly rich crowdsourced sound database resource with numerous annotations and tags that accompany uploaded sounds. However, it is also limited for machine learning usage as each uploaded audio file can only be annotated by its contributor, and its tags represent the *entire* audio file regardless of duration. This is ideal for single, finely cropped sound files, but the majority of the uploaded sound files vary greatly in duration. Although Freesound is not ideal a *ground-truth* resource, it provided an excellent opportunity in our initial efforts to: (1) develop procedures for soundscape audio annota-

[2] https://www.freesound.org/

tion/labeling, (2) explore collective listening data mining strategies via crowdsourced annotations, (3) create an initial small dataset of ground truth, and (4) develop custom online interfaces and potential practices for soundscape data annotation. Our procedure (adopted in [27]) for open-ended annotation is shown in Figure 2 where sounds and tags are downloaded via our custom MATLAB-Freesound API as part of the SATB Toolbox further discussed in Section 3. Sounds are then imported into Audacity[3] as an audio track. This followed by creating "label tracks" to annotate acoustic events, which are saved as text files that, can be read into systems such as MATLAB.

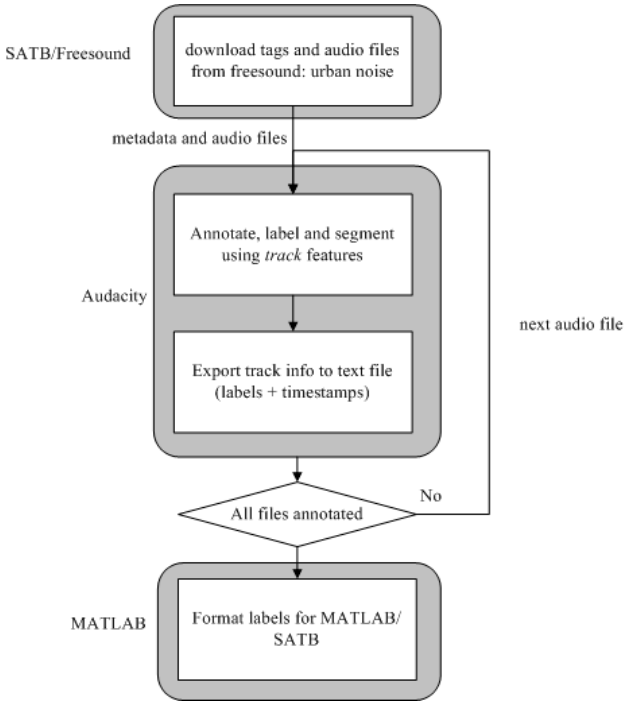

Figure 2. Annotation and labeling procedure.

We are also currently finishing up our custom online crowdsourced annotation software. This software is based on our initial studies in taxonomy and ground-truth dataset development using existing tools and soundscape recordings that reside on the Citygram server. This software will allow for multiple annotate, label, and segment audio events from a large pool of sound files and expect approximately 50 hours of multi-person annotations.

4. SOUND ANALYTICS

In this section we describe some of the ways we have used SATB for "organized" soundscape auditioning, feature space exploration, and soundscape tag analysis.

4.1 Soundscape Exploration

As we are in the beginning stages of exploring soundscape information retrieval research, gaining insights into the feature space, semantic space, and acoustic event dimensions is important. As a first step, k-means clustering was employed to automatically group acoustic events according to their low-level acoustic properties. Feature vectors currently being used include RMS, zero-crossing rate, spectral centroid, spectral flatness, spectral flux, spectral spread, and 13 MFCCs. Each sound file is segmented into acoustic events using AED techniques further described in Section 4.4. For each acoustic event, a 38-dimensional feature vector is obtained by computing the mean and standard deviation across the analyzed feature values. In addition, the mean and standard deviation of the first and second derivatives are also calculated to provide velocity and acceleration information per acoustic event resulting in a total of 114 dimensions. At this point, the grouping task for the collection of acoustic events is reduced to a typical vector quantization problem that can be effectively done via k-means.

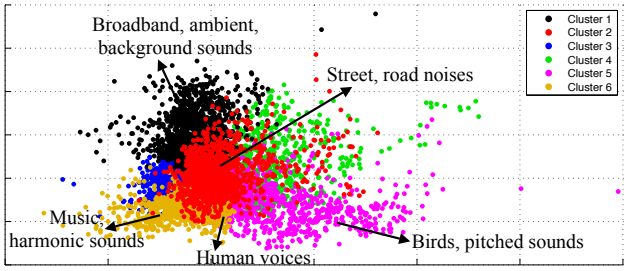

Figure 3. 2D event plot by the cluster hovering interface.

When acoustic events are laid out and grouped in clusters as a static plot, it is difficult to determine the characteristics of each "data point" – e.g., what it sounds like, what its tags are, and what its relationship is to neighboring events. Motivated by being able to interact with important information associated with acoustic events including its filename, tags, durations, and sound, we have begun developing an interactive the SATB "cluster hovering" tool as shown in Figure 3. This feature space exploration tool can be used for multimodal monitoring and interaction with each of the acoustic events. In Figure 3, 7,850 acoustic events from 1,188 Freesound soundscape recordings totaling 36 hours are organized into six clusters and plotted in a 2D feature space, where the axes are chosen for maximum separation of the events via principal component analysis (PCA). Visual and auditory monitoring is done by simply moving the mouse around the events – the event that is the closest to the mouse pointer is automatically triggered to play in real time. The cluster hovering interface also provides a feature to "denoise" the dataset. This is accomplished by simply deleting data points – acoustic events – that are considered irrelevant or clustered incorrectly. This may facilitate in quickly creating a ground-truth dataset for subsequent machine learning efforts.

4.2 Crowdsourced Tag Incorporation

Crowdsourced tags can be helpful in developing soundscape taxonomy, which effectively connects a continuous acoustic signal with semantic labels and descriptors provided by its contributor. In fact, simply interactively visualizing and monitoring acoustic events with a list of the

[3] http://audacity.sourceforge.net/

the associated tags helps in developing a sense of soundscape taxonomy. However, there is a fundamental issue with crowdsourced labels (e.g., Freesound) in that they are open-ended. There are little to no restrictions and guidelines as to how to tag each file. Hence, noise in the form of consistency, reliability, and relevancy are rendered as artifacts. To address the issue of de-noising crowdsourced annotations, we have begun implementing simple pre-processing steps: (1) tag normalization, (2) spelling correction/lemmatization (grouping of the words with a same root form), and (3) occurrence pruning. Tag normalization entails removing all non-character symbols in tags and converting all characters to lowercase, which improves consistency while decreasing redundancy. The remaining tags are collected as a set that represents a given acoustic event. Spelling correction/lemmatization is currently implemented via computation of *edit distance* [31] for each and every pair of tags; when the edit distance is less than a predefined threshold, the pair is registered on a dictionary as potentially containing the same semantics. Since the morphological distance may not match the semantic distance, manual adjustments are additionally made on the dictionary to discard the irrelevant pairs. After the tag pairs in the dictionary are lumped together, the occurrence pruning stage completes the preprocessing procedure: occurrence of each tag is counted and infrequent tags, with less than five occurrences, are removed. Using the above procedures, 1,979 tags were filtered down to 373 tags obtained from 1,188 Freesound sound files. Figure 4 shows the five most representative tags for three example clusters after the pre-processing. The representativeness score is based on the occurrence ratio in the target cluster minus the maximum occurrence ratio among the rest of the clusters.

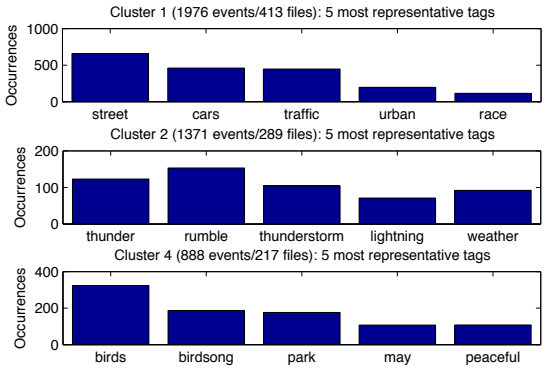

Figure 4. Five most representative tags examples.

Upon observation of the remaining tags, it was clearly noticeable that additional filtering could be implemented to inform hierarchy and taxonomical information. For example, there were a number of tags that referred to geophonies, others referred to biophonies, while others were related to human sounds. Other observations were that it would also be possible to use thesaurus APIs and tools such as WordNet[4] to further extract label hierarchies and taxonomies while reducing redundancies.

[4] http://wordnet.princeton.edu/

4.3 Tag Hierarchy Extraction

Currently, a simple statistical method is employed to derive basic hierarchical information from the user-uploaded tags. The conditional probability of the presence of tag **A** given the presence of tag **B** is calculated for each tag pair based on co-occurrence counts. When the conditional probability is close to 1 and its inverse probability is small but not insignificant, it can be inferred that tag **A** may be an antecedent of tag **B**. This forms the basis of our statistical approach, and more sophisticated methods will be devised to derive multi-level tag hierarchy.

4.4 Acoustic Event Detection (AED)

The majority of audio classification methodologies simultaneously do AEC and AED. Examples of popular AEC methods include HMM or GMM based classifiers [2], [4], [32]. Although such AECs have been proven effective, it is often required that the target audio scene is specific and the event classes are well-defined and small in number. In Citygram, we are starting to develop an approach where AED is conducted separately from AEC. That is, we first do a computationally light AED, and only when an acoustic event is detected do we run the classification module. This is due to a number of reasons including system efficiency, sensor network transmission bandwidth, and consideration of soundscape characteristics, which greatly vary depending on location and time. In the continuous context of real-time soundscape classification using a heavy AEC system that runs 24/7 is therefore wasteful.

4.4.1 AED Algorithm

The AED algorithm was developed by manually varying SNR levels to mimic the dynamicity of "background noise." The algorithm consists of four main modules: initialization, pre-processing, de-noising, and energy-thresholding modules.

Initialization

When an RSD goes online for the first time, a "coarse" AED algorithm is employed to detect acoustic events based on (1) spectral peak of the STFT envelope, (2) magnitude of the spectral peak, and (3) spectral spread defined by band edges at -10dB below the peak magnitude. The three spectral parameters are adaptively updated to initially determine acoustic event segments. As further discussed in the following section, an initialization period is required in order to roughly measure the noise profile which will be removed during the pre-processing stage.

Preprocessing: De-noising

The noise floor, ambiance, or background noise of soundscapes vary with time and is dependent on in-situ elements such as traffic noise during rush hours. To improve the performance of our AED algorithm, we employ a pre-processing module to spectrally remove background noise from the signal before applying a simple energy-based thresholding procedure. Noise is assumed to be

ergodic in the short term but is also capable of significant variation in the long term. The noise profile is adaptively measured during non-acoustic event periods and is used to compute a representative spectral noise template of a continually changing soundscape. The de-noising algorithm is based on [33], which produces an SNR matrix from a dynamical noise profile template. This matrix is used to discriminately weight the DFT frames' individual magnitude components. The IDFT of the modulated spectrum renders the de-noised signal. The applied window size is 50ms with a hop size of 55% of of the window size. Each segment is then enveloped with a hamming window. The same windowing is also applied to RMS computation. Finally, we tested other de-noising algorithms including LPC filtering and spectral subtraction. The latter two produced poor results.

RMS thresholding

The final stage in determining the acoustic event segment is achieved by first computing the RMS of the de-noised signal followed by its multiplication with the original signal's RMS vector. A moving average RMS is dynamically compressed and shifted up by the mean of the entire RMS values. This process attempts to dynamically model noise floor characteristics render an adaptive thresholding mechanism for robust AED.

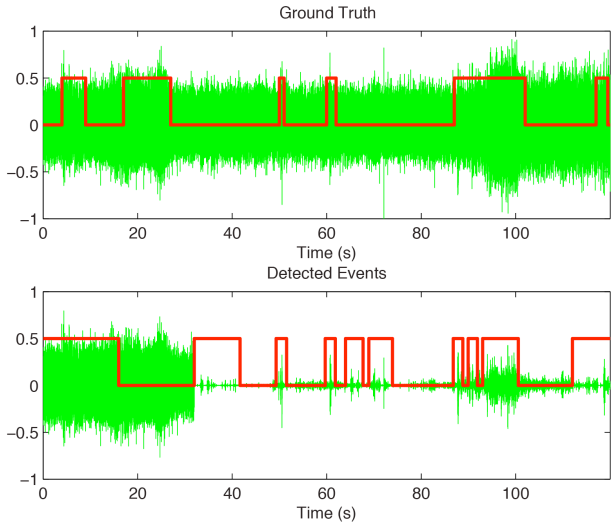

Figure 5. NYC Times Square recording: before and after de-noising (top is original).

4.4.2 AED Test Results

Three datasets were used in AED performance evaluation: (1) a dataset from Freesound (annotated by Dhruv Bhatia) (2) an in-house NYC Times Square field-recording set, and (3) a dataset with varying noise levels obtained from NYC soundscapes. AED performance was then evaluated on 80 in situ soundscape recordings consisting of 248 events. We used three standard metrics for evaluation of AED performance: *precision*, *recall*, and *AED-ACC* [3], [34]. To simulate environmental SNR change, a set of audio samples with varying SNR were produced. Acoustic events such as gunshots, crowds cheering, musical sounds, and other sounds were mixed with increasing SNR levels. The sound classes were chosen while considering diversity in spectral content, event duration, and amplitude envelope. The SNR level was modulated between 0.0 and 1. in 0.1 increments occurring at every 10-second interval as shown in Table 2.

		Freesound	Times Square	SNR mod
Audio samples	Num. of files	62	9	11
	Num events (min)	1	2	2
	Num events (max)	7	7	6
	Num events (total)	176	36	36
	Num events (mean)	2.84	4	4
	Duration total (s)	4481.0	1080	1320
Event dur. (s)	GT event dur min	0.42	1	1
	GT event dur max	59.84	34	6
	GT event dur mean	6.52	7.8	2.45
	AED event dur min	0.26	1.39	0.95
	AED event dur max	27.26	17.58	13.19
	AED event dur mean	5.27	5.26	3.70
Perf.	Precision	0.36	0.35	0.70
	Recall	0.73	0.61	1.00
	AED-Acc.	0.43	0.36	0.82

Table 1. Summary of sample stats, segmentation stats, and AED performance.

SNR	Precision	Recall	AED-ACC
0	.85	1	.92
0.1	.83	1	.91
0.2	.83	1	.91
0.3	.83	1	.91
0.4	.77	1	.87
0.5	.71	1	.83
0.6	.64	1	.78
0.7	.60	1	.75
0.8	.56	1	.72
0.9	.53	1	.70
1	.53	1	.70

Table 2. SNR modulation results.

4.4.3 Discussion

A clear observation is that *recall* consistently outperforms *precision*, which means that AED identifies additional events not labeled by the annotator. It is currently difficult to arrive on a conclusive explanation given the size of our dataset. However, upon further considering the AED results and carful listening to the audio samples where the precision errors occurred, it was surprisingly difficult to assess whether the additional AED events were actually "incorrect." This may perhaps suggest that different modes of audition – hearing vs. listening – may yield different results, much like when watching a football game: humans would likely not notice a sparrow flying above the grounds even if the bird were within one's line of sight. Another observation is that annotators tended to group sequences of short events (< 3 sec) into a single event while the AED algorithm identified short acoustic events separately. Again, upon more careful lis-

tening, it was not clear whether the AED or annotator was actually "correct." As a matter of fact, both seemed correct depending on perspective. This result suggests the possibility of including an "auditioning mode"–detailed vs. *less* detailed, *listening* vs. hearing –depending on what type of information is needed and for what purpose.

Figure 4 shows event duration distributions where we note that the majority of events (> 50%) have short durations (< 5 sec). The overall distribution of the ground-truth events and that of the AED events are similar except for the very short ones. The scarcity of very short acoustic events (< 0.5 sec) detected by our AED algorithm is the result of extending such potential events to render a longer acoustic event with multiple impulsive spikes.

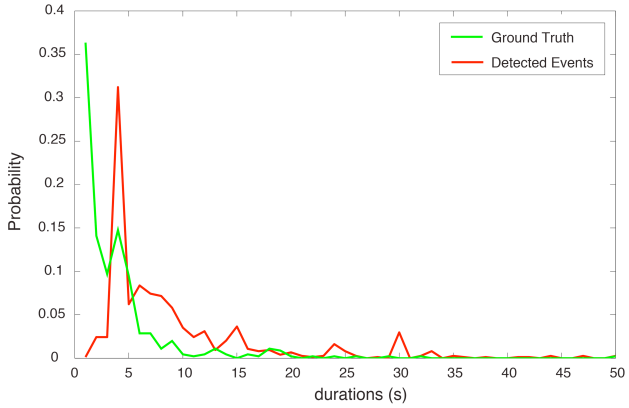

Figure 6. Distribution of events by duration.

5. FUTURE WORK

The recent developments of SATB have laid the foundation for four important areas of future work: (1) large-scale quantification of the efficacy and efficiency of various ML approaches, (2) the collection and collation of a massive database of urban sounds, (3) investigation of collective listening strategies for semantic data mining, and (4) the development of additional exploration interfaces to help gain insights into complex feature spaces. A prevalent issue in ML is the correct parameterization of a model so as to avoid "over-fitting" and also "under-fitting." Currently, we are employing the aforementioned features, as they are common in traditional MIR tasks. However, as has been demonstrated, such a reliance on canonical tools may not make sense when approaching the problems unique to soundscapes. Although our AED algorithm will need to be further refined, our aim for the immediate future is to begin focusing on the AEC component of the Citygram Project using SATB's Weka module.

As previously stated, there is a dearth of readily available datasets for soundscape research. Though field-recording is a technique that has been utilized since the inception of electro-acoustic music, efforts to collate and label these sounds for research or artistic purposes have been difficult. As such we are finishing up development of custom cloud-based annotation software and expect approximately 50 hours of labeled ground truth data.

In the longer term, Citygram's over-arching goal is the collection of a large soundscape dataset, and the proposed large-scale deployment of dense sensor networks opens up the possibility for such a repository. Much like Freesound or the Million Song Database, future work in this area seeks to produce a "Million Sound Dataset," and indeed this collection approach in concert with the wide inclusion of all sounds endemic to urban life may suggest that "Billion Sound Dataset" will be a more appropriate label. Perhaps the feature of SATB currently underway most appreciable by its (currently) small user-base is the cluster hovering feature. This planned feature allows a user to interact with a multidimensional feature space, exploring it in an intuitive and poly-sensory manner. Currently, the feature allows for a single sound to be played when it is "hovered over," future plans include an expanded pallet of interaction options for playing single or multiple sounds. For instance, the "lasso" tool familiar from many graphics editing programs could be used to select and playback sounds within a given space.

6. CONCLUSIONS

In this paper, we have discussed ongoing efforts in the measuring, archiving, and quantification of soundscapes with an emphasis on data analytics. Using the notion of a densely deployed sensor infrastructure of the Citygram project, efforts are being made to build upon soundscape research in the domains of AED/AEC, and the design and application of a robust, descriptive taxonomy for urban soundscapes. It has been demonstrated that the problem of SIR is non-trivial and that it bears important dissimilarities from the field of MIR. In order to facilitate research, exploration, and study of soundscapes we have started to develop SATB that includes extensible plugin architecture for analysis algorithm expandability, while handling interactive visualizations, proprietary AED methodology, and taxonomic exploration.

7. REFERENCES

[1] X. Valero Gonzalez and F. Alías Pujol, "Automatic classification of road vehicles considering their pass-by acoustic signature," *J. Acoust. Soc. Am.*, vol. 133, no. 5, p. 3322, 2013.

[2] G. Tur and A. Stolcke, "Unsupervised Languagemodel Adaptation for Meeting Recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007, vol. 4, pp. IV–173–IV–176.

[3] S. Duan, J. Zhang, P. Roe, and M. Towsey, "A survey of tagging techniques for music, speech and environmental sound," *Artif. Intell. Rev.*, Oct. 2012.

[4] C. Clavel, T. Ehrette, and G. Richard, "Events Detection for an Audio-Based Surveillance System," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 1306–1309.

[5] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition," in *2007 3rd International Conference on Intelligent Sensors,*

Sensor Networks and Information, 2007, pp. 293–298.

[6] R. Mogi and H. Kasai, "Noise-Robust environmental sound classification method based on combination of ICA and MP features," *Artif. Intell. Res.*, vol. 2, no. 1, p. p107, Nov. 2012.

[7] J. F. van der Merwe and J. A. Jordaan, "Comparison between general cross correlation and a template-matching scheme in the application of acoustic gunshot detection," in *2013 Africon*, 2013, pp. 1–5.

[8] T. H. Park, B. Miller, A. Shrestha, S. Lee, J. Turner, and A. Marse, "Citygram One: Visualizing Urban Acoustic Ecology," in *Digital Humanities*, 2012.

[9] T. H. Park, J. Turner, C. Jacoby, A. Marse, M. Musick, A. (California I. of the A. Kapur, and J. (California I. of the A. He, "Locative Sonification: Playing the World Through Citygram," in *International Computer Music Conference Proceedings (ICMC)*, 2013, pp. 11–17.

[10] T. H. Park, J. Turner, M. Musick, J. H. Lee, C. Jacoby, C. Mydlarz, and J. Salamon, "Sensing Urban Soundscapes," in *Workshop on Mining Urban Data*, 2014.

[11] R. M. Schafer, "The Soundscape: Our Sonic Environment and the Tuning of the World," 1977.

[12] W. W. Gaver, "What in the world do we hear? an ecological approach to auditory event perception."

[13] M. R. Ismail, "Sound preferences of the dense urban environment: Soundscape of Cairo," *Front. Archit. Res.*, vol. 3, no. 1, pp. 55–68, Mar. 2014.

[14] D. Dubois, C. Guastavino, and M. Raimbault, "A Cognitive Approach to Urban Soundscapes: Using Verbal Data to Access Everyday Life Auditory Categories." S. Hirzel Verlag.

[15] T. Houtgast, "Speech perception: Ideas and concepts initiated by Reinier Plomp," *J. Acoust. Soc. Am.*, vol. 105, no. 2, p. 1238, Feb. 1999.

[16] B. Gygi, "Factors in the identification of environmental sounds," Indiana University, 2001.

[17] "Thinking in Sound: Paperback: Stephen McAdams - Oxford University Press." [Online]. Available: http://ukcatalogue.oup.com/product/9780198522577.do. [Accessed: 16-Apr-2014].

[18] M. M. Marcell, D. Borella, M. Greene, E. Kerr, and S. Rogers, "Confrontation naming of environmental sounds.," *J. Clin. Exp. Neuropsychol.*, vol. 22, no. 6, pp. 830–64, Dec. 2000.

[19] S. Sinnett, C. Spence, and S. Soto-Faraco, "Visual dominance and attention: the Colavita effect revisited.," *Percept. Psychophys.*, vol. 69, no. 5, pp. 673–86, Jul. 2007.

[20] "The Art of Noises: Luigi Russolo: 9781576471142: Amazon.com: Books." [Online]. Available: http://www.amazon.com/The-Art-Noises-Luigi-Russolo/dp/1576471144. [Accessed: 16-Apr-2014].

[21] B. Edward, E. B. Dennis Jr., J. Henry, and G. E. Pendray, *City Noise*. New York City: Academy Press, 1930.

[22] "The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900-1933 / Edition 1 by Emily Thompson | 9780262701068 | Paperback | Barnes & Noble." [Online]. Available: http://www.barnesandnoble.com/w/soundscape-of-modernity-emily-thompson/1103854501?ean=9780262701068. [Accessed: 16-Apr-2014].

[23] A. L. Brown, J. Kang, and T. Gjestland, "Towards standardization in soundscape preference assessment," *Appl. Acoust.*, vol. 72, no. 6, pp. 387–392, May 2011.

[24] W. J. D. K. Foale, "A listener-centred approach to soundscape evaluation."

[25] C. Guastavino, B. F. Katz, J. D. Pollaack, D. J. Levitin, D. Dubois"Ecological validity of soundscape reproduction," Acta ACustica united with Acustica, 91(2, 333-34, 2005.

[26] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[27] J. Salamon, C. Jacoby, and J. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *ACM International Conference onf Multimedia*, 2014.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, 2009.

[29] T. Park, Z. Li, and W. Wu, "Easy Does It: The Electro-Acoustic Music Analysis Toolbox.," in *10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009, pp. 693–698.

[30] M. Kleiner, D. H. Brainard, and D. G. Pelli, "What's new in Psychtoolbox-3?," *Perception*, vol. 36, p. S14, 2007.

[31] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2009, p. 988.

[32] J. J. Eva Vozáriková, *Multimedia Communications, Services and Security*, vol. 149. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 191 – 197.

[33] G. Y. G. Yu, S. Mallat, and E. Bacry, "Audio Denoising by Time-Frequency Block Thresholding," *IEEE Trans. Signal Process.*, vol. 56, 2008.

[34] A. Temko, C. Nadeu, D. Macho, and R. Malkin, "Acoustic event detection and classification," *Comput.*, pp. 61–73, 2009.