Sensing Urban Soundscapes

Tae Hong Park[1], Johnathan Turner[1], Michael Musick[1], Jun Hee Lee[1],
Christopher Jacoby[1], Charlie Mydlarz[1,2], Justin Salamon[1,2]

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

[2]Center for Urban Science and Progress (CUSP)
1 MetroTech Center, 19th floor
New York University
New York, NY 11201 USA

{thp1, jmt508, musick, junheelee, cbj238, cmydlarz, justin.salamon}@nyu.edu

ABSTRACT

Noise pollution is one of the most serious quality-of-life issues in urban environments. In New York City (NYC), for example, more than 80% of complaints[1] registered with NYC's 311 phone line[2] are noise complaints. Noise is not just a nuisance to city dwellers as its negative implications go far beyond the issue of quality-of-life; it contributes to cardiovascular disease, cognitive impairment, sleep disturbance, and tinnitus[3], while also interfering with learning activities [21]. One of the greatest issues in measuring noise lies in two of the core characteristics of acoustic noise itself — transiency and structural multidimensionality. Common noise measurement practices based on average noise levels are severely inadequate in capturing the essence of noise and sound characteristics in general. Noise changes throughout the day, throughout the week, throughout the month, throughout the year, and changes with respect to its frequency characteristics, energy levels, and the context in which it is heard. This paper outlines a collaborative project that addresses critical components for understanding spatio-temporal acoustics: measuring, streaming, archiving, analyzing, and visualizing urban soundscapes [28] with a focus on noise rendered through a cyber-physical sensor network system built on Citygram [23, 24].

General Terms

Algorithms, Measurement, Design, Reliability, Experimentation, Security, Human Factors, Standardization, Theory.

[1]"Improving Our Quality of Life: Operation Silent Night," http://www.nyc.gov

[2]http://www.citymayors.com/environment/nyc_noise.html

[3]http://www.euro.who.int/__data/assets/pdf_file/0008/136466/e94888.pdf

Keywords

Cyber-physical sensor network, data mining, data streaming, big data, acoustics, noise, mobile/distributed computing, mobile data management, soundscapes, machine learning.

1. INTRODUCTION

In 1800, a mere 1.7% of the global population lived in cities of 100,000 or more; at the beginning of the 1950s, that number rose to 13% [17], and as of 2013, there are 24 *megacities*[4], each inhabited by more than 20 million people. By 2050, the projection is that 68% of the global population will dwell in urban areas, which will include more than 37 megacities [31]. The growth of cities has been incredible and when megacities began to emerge in the 1970s (there were only two: NYC and Tokyo [31]), interest in the quality-of-life of city dwellers began to garner the attention of researchers. In the United States, for example, the shift from considering noise as a mere nuisance and an artifact of city-life began to change in the 1970s resulting in Congress putting forth The Noise Pollution and Abatement Act[5]. Research on the impact of noise on urban inhabitants by environmental psychologists also began during this period [6, 5, 10]. Today in NYC, urban noise comprises approximately 80% of all 311 complaints where approximately 40% of the noise complaints are related to loud music, 18% construction noise, and 13% loud talking[6]. With the annual doubling of the world's population, increases in urban noise complaints have the potential to reach alarming levels. However, measuring noise is not trivial. One of the most comprehensive noise codes in the world is The Portland Urban Noise Code[7], which is based on spatio-temporal metrics of sound. That is, noise is described in terms of location, time, and acoustic energy measurements and is supervised by a control officer with codes enforced by the police department. Noise, however, cannot simply be defined by measuring the decibel (dB) levels at a specific time and place. Furthermore, manually and continuously measuring urban soundscapes is impractical and

[4]Cities that have a population in excess of roughly 20 million people.

[5]42 USC § 4901, 1972

[6]https://nycopendata.socrata.com

[7]Herman, Paul. Portland, Ord. No. 139931. 1975, amend. 2001. http://www.nonoise.org/lawlib/cities/portland_or

for all intents and purposes, unfeasible. Our project aims to contribute in developing a comprehensive cyber-physical system to automatically measure, stream, archive, analyze, explore, and visualize acoustic soundscapes with a focus on noise. We begin our paper with a survey of related works, an introduction to the Citygram cyber-physical system and its various modules, and a summary and outline of future work.

1.1 Related Work

Quite a few "sound mapping" examples exist including The BBC's *Save Our Sounds*[8], *NoiseTube* [18], and *Locustream SoundMap* [16]. Most of the soundmaps are, however, non-real-time and are based on "click and play audio snapshot" interfaces. An example is Save Our Sounds, which puts forth the idea of archiving "endangered sounds". NoiseTube and WideNoise are two other examples that use crowd-sourcing concepts while utilizing cell-phones' microphones to measure and share geolocalized dB(A) levels. The Locustream SoundMap project is one of the few sound maps that stream real-time audio using an "open mic" concept. In Locustream, participants (known as "streamers") install the developer-provided custom boxes in their apartments and share "non-spectacular or non-event-based quality of the streams." Other examples include *da_sense* project [29], which provides a platform for data acquisition, processing, and visualization of urban sensor data (sound, temperature, brightness, and humidity). Their publicly available *NoiseMap* Android application crowd-sources acoustic energy data from participants' smartphones, which is presented on an online mapping interface with an accompanying public API for data sharing with a daily update rate. Their online platform can also accept data from static sensor networks and has collected over 40,000 data points to date. Commercially available noise monitoring sensors have also been utilized in spatio-temporal urban noise assessments [29]. The discontinued *Tmote Invent* noise-monitoring sensor was used to sample sound levels and transmit them wirelessly at regular intervals. The study focused on the power consumption of these remote sensors, identifying significant power savings when the data transmission strategy is modified, at the cost of increased system latency. A final example is a project called *Sensor City*[9], which aims to deploy hundreds of static sensing units equipped with acoustic monitoring equipment around a small city in the Netherlands. This solution utilizes its own dedicated fiber-optic network and high-end calibrated audio recording devices. The project is taking a soundscape analysis approach and is looking to investigate human perception and evaluation of acoustic environments within urban settings. The project aims to qualify soundscapes through the development of machine learning algorithms that will analyze incoming data from the sensor network.

2. THE CYBER-PHYSICAL SYSTEM

Our interest in this project began in 2011 when observing that current topological mapping paradigms were typically static and focused on visualizing city layouts characterized by slowly changing landmarks such as buildings, roads, parking lots, lakes, and other fixed visual objects. Three-dimensional physical shapes, however, do not only define urban environments; they are also defined by "invisible energies" including acoustic energy. Noticing the underrepresentation of sound in modern mapping systems, we began to explore ways to capture spatio-acoustic dimensions and map them on conventional online mapping interfaces. The Citygram project currently involves collaborators from New York University's (NYU) Steinhardt School, NYU's Center for Urban Science and Progress (CUSP), and the California Institute of the Arts (CalArts) with support from Google. The project was launched in 2011 to develop methodologies, concepts, and technologies to facilitate the capture and visualization of urban non-ocular energies. The first iteration of Citygram is specifically focused on acoustic data — with such applications as creating dynamic soundmap overlays for online systems, such as Google Maps. Many modern mapping systems have no need to address the issue of temporality. Roads, buildings, and parks do not change on a regular basis; the image update rate for Google Earth, for example, is typically between 1 to 3 years[10]. Such slow update rates, however, are grossly inadequate for mapping sound due to its inherent temporality. The project's main goal began as an effort to contribute to existing geospatial research by embracing the idea of time-variant, poly-sensory cartography via multi-layered and multi-format data-driven maps based on continuous spatial energies captured by terrestrially deployed remote sensor devices (RSDs). The project's goals have revolved around creating dynamic, spatio-acoustic maps to help us better understand urban soundscapes and develop technologies to automatically capture man-made, environmental, and machine-made sounds.

NYU CUSP was formed in fall 2013 to utilize "New York City as its laboratory and classroom to help cities around the world become more productive, livable, equitable, and resilient." CUSP aims to observe, analyze, and model cities "...to optimize outcomes, prototype new solutions, formalize new tools and processes, and develop new expertise/ experts."[11] One of the projects that CUSP has started to focus on is noise in NYC. This has brought together NYU Steinhardt's Citygram project and CUSP's initiatives in noise research.

2.1 Citygram: System Overview

The Citygram system, which includes three main modules, is shown in Figure 1. The three modules include the RSDs on the left, the server in the middle, and users on the right-hand side. In the following subsections we summarize each module starting with the sensor network and remote sensing devices.

2.2 Sensor Network and RSDs

The RSDs form the sensor network which capture, compute, and stream spatio-acoustic data and metadata to the server. The server collects, archives, manages, and analyzes data received from RSDs. We employ two main RSD deployment strategies to create our sensor network as shown in Figure 2: (1) fixed RSDs and (2) crowd-sourced RSDs. Fixed RSDs are installed at fixed locations, which allows

[8] http://www.bbc.co.uk/worldservice/specialreports/saveoursoundsintro.shtml
[9] http://www.sensorcity.nl
[10] http://sites.google.com/site/earthhowdoi/Home/ageandclarityofimagery
[11] http://cusp.nyu.edu/about/

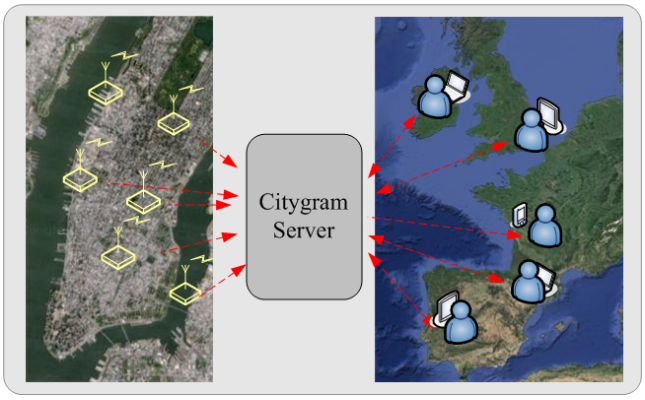

Figure 1: Cyber-physical system.

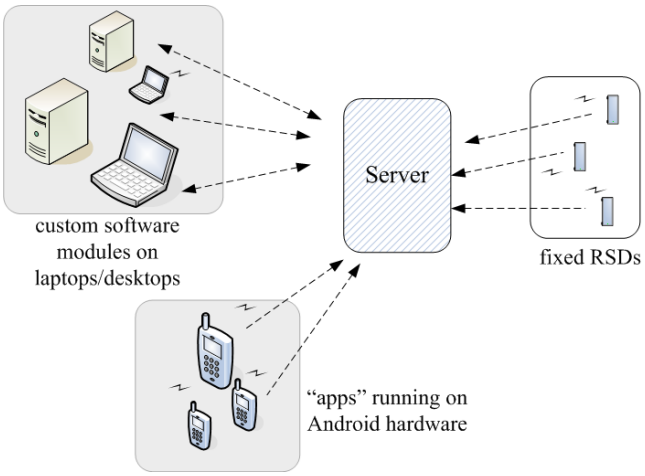

Figure 2: Sensor network showing different RSD types.

for continuous, consistent, and reliable data steams via sensor devices, further detailed below. Difficulties in creating a large-scale, fixed sensor network include selection of appropriate RSDs, as further detailed below, as well as issues concerning deployment in public spaces as discussed in the Future Work section. To create a robust and growing sensor network in lieu of the fixed RSD deployment strategy, we also employ crowd-sourced RSDs to address the issue of coverage expansion, spatial granularity, and engagement of community and citizen scientists. The crowd-sourced RSDs are further divided into mobile apps (e.g. smartphones) and custom software modules for existing personal computer applications that run on software platforms including Max[12] and SuperCollider[13]. The mobile app-based RSDs currently run on the Android OS. Personal computing RSDs only require a microphone and Internet connectivity to stream acoustic data from the user's device to our server. This crowd-sourced RSD strategy allows for additional measurements that can be especially useful in capturing data that are beyond existing fixed RSD network boundaries.

[12]`http://cycling74.com`
[13]`http://supercollider.sourceforge.net`

2.2.1 Selection of RSDs

A number of hardware platforms have been considered in selecting possible candidates for our fixed RSDs including the Alix system, Raspberry Pi, Arduino, smartphones, and others. Our approach in selecting an appropriate RSD for geospatial acoustic monitoring followed nine criteria: (1) audio capture capability, (2) processing power and RAM, (3) power consumption, (4) flexible OS, (5) onboard storage, (6) wireless connectivity, (7) I/O options/expandability, (8) robustness/sturdiness, and (9) cost. Rather than opting to use custom boards such as the Raspberry Pi, the solution decided upon was Android-based hardware devices including Android mini-PCs. A sub $50 Android mini-PC, for example, contains the following specifications: quad-core CPU, 2GB RAM, onboard Wi-Fi, Bluetooth, built-in microphone and camera, an HDMI port, and multiple USB ports. A selection of smartphones and mini-PC's was chosen and run through a preliminary set of objective and subjective tests to determine their suitability for acoustic monitoring in terms of their frequency response and recording quality. The devices are shown in Figure 3, from top left to bottom right: (a) HTC Evo 4G, (b) Samsung Star, (c) KD812, (d) Cozyswan MK812, and (e) Measy U2C. One of the goals of our project is the automatic classification of sound events in urban spaces as further discussed in Section 2.6. To avoid the loss of information, potentially important for the machine learning stage, it is imperative to use appropriate microphones, sampling rates, and bit resolution. During the process of short-listing RSD candidates, we also considered audio capture specifications specifically used in automatic sound classification. The literature seems to suggest varying sampling frequencies. For example, work in classification of soundscape-based sound events in [9] used a 12 kHz sampling rate and in [32] a 16 kHz and 16 bit resolution system was used; in other classification and sound detection work where the focus is on musical signals, filter-banks between 100 Hz–10 kHz [26] and 27.7 Hz–16 kHz were used [3]. Although it is difficult to draw conclusions on the minimal requirements of sample rate and bit resolution parameters, literature in the field of machine learning pertinent to sound seems to suggest that it is not always necessary to have a granularity of 44.1 kHz/16 bit. As part of determining appropriate sound capture parameters, we plan to further investigate the influence of time/frequency resolution on our machine learning algorithms.

To test the potential RSD candidates, environmental recordings were made using all devices simultaneously from a third floor apartment window overlooking a busy urban intersection in Brooklyn, New York. Both of the smartphones sounded clear and individual sound sources were easily identifiable. The frequency bands around the voice range, however, seemed to be accentuated, possibly a result of the devices' optimization for speech handling. The KD812 and the Cozyswan mini-PCs clearly showed artifacts of dynamic compression. The Measy U2C produced far clearer recordings than the other mini-PCs, with a better low frequency response. Reference sine-sweeps (20 Hz to 20 kHz) were recorded using an Earthworks M30 measurement microphone mounted on-axis placed one meter from a Genelec 8250a loudspeaker in an acoustically soundproofed lab. These first measurements allowed compensation of loudspeaker and room frequency response coloring. The measurement microphone was then replaced by each device with microphone ports on-

Figure 3: Potential RSDs: (a) HTC Evo 4G, (b) Samsung Star, (c) KD812, (d) Cozyswan MK812, and (e) Measy U2C.

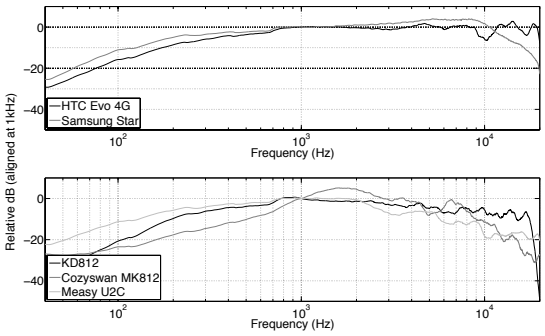

Figure 4: Frequency response of Android devices.

axis to the loudspeaker. The same sine-sweeps were then repeated for each device. Impulse responses were extracted from each sweep using deconvolution techniques, generating device frequency responses as shown in Figure 4. Each device response was level-normalized at 1 kHz.

All devices show a relatively poor response at lower frequencies. This is especially the case for Cozyswan MK812 mini-PC. The large peak in the device's frequency response at 1–3 kHz highlights the perceived coloration of the environmental recording. The enhanced high frequency response from the HTC and Samsung smartphone devices can also be observed. Between 2–10 kHz the response varies between device types with the most amount of high frequency drop-off occurring on the mini-PC's. All of the microphones on the mini-PC's were mounted on the internal printed circuit board facing up, explaining the filtering effects caused by the small microphone port coupled to the relatively large internal cavity of the device. Based on our tests, the smartphones seemed to provide the preferred solution. However, to match the processing power of the mini-PC's, a high-end smartphone would be required. Furthermore, audio recording quality of the Measy U2C's built-in microphone is comparable to the tested smartphones with the added benefit of increased processing power and expandability. This expandability allows for the connection of external USB microphones, providing the potential for consistent and enhanced audio quality as well as flexibility in microphone placement. The external USB microphone allows the device itself to be kept modular and separate from the mini-PC, allowing flexibility in its placement and replacement. We are currently in the process of investigating the use of external USB audio peripherals.

2.3 Signal Processing and Feature Extraction

The RSDs autonomously capture audio in real-time and employ distributed computing strategies to address efficiency and scalability issues by running various computational tasks including feature extraction. This strategy alleviates stress on the server in the context of sensor network scalability and efficiency. Blurred acoustic data and non-invertible low-level acoustic features are streamed to the server.

Our current custom Android software, which runs both on our fixed sensor network as well as on conventional Android-based smartphones, uses OpenSL ES[14] to record audio blocks using circular buffering techniques. As shown in Figure 5, feature vectors are transmitted as JSON[15] objects via HTTP posts. The raw audio signal is compressed and streamed to the server where users can then monitor soundscapes in real-time. Before transmission to our server, the raw audio is passed through a speech-blurring algorithm as further detailed below. For development of machine learning algorithms we also have the option to stream raw audio data using libsndfile[16] with FLAC codec[17], which is separately archived on the server and inaccessible by the public.

To enable users to hear the "texture" and characteristics of spaces without compromising the privacy of people in the monitored areas, we employ a modified granular synthesis technique [27] to blur the spectrum of the voice. This is achieved prior to streaming the raw audio data to our server where users can access the blurred audio streams. To accomplish these conflicting tasks — blurring the audio while retaining the soundscape's texture — a multi-band signal processing approach was devised. The audio signal is first divided into three sub-bands using overlapping filter windows with cascaded third-order Butterworth filters. The middle band represents the voice band. The isolated voice band is then segmented into 32 windows, or "grains", with 50% window overlap. Blurring is achieved by randomizing the temporal order of the windows. The randomized grains are then used to construct the blurred voice band signal via overlap-and-add [4, 25]. The final audio signal sent to the server is a sum of the unaltered lower/upper bands and the modulated voice band. Not shown in Figure 5 is a separate audio thread and audio streaming block that transmits raw audio data for machine learning research purposes. This module will eventually be removed from our system once our algorithms are fully developed.

2.4 Server and Database

The basic structure of the Citygram system is comprised of users, RSDs, and the Citygram server. A user may reg-

[14] `http://www.khronos.org/opensles`
[15] `http://www.json.org`
[16] `http://www.mega-nerd.com/libsndfile`
[17] `http://xiph.org/flac`

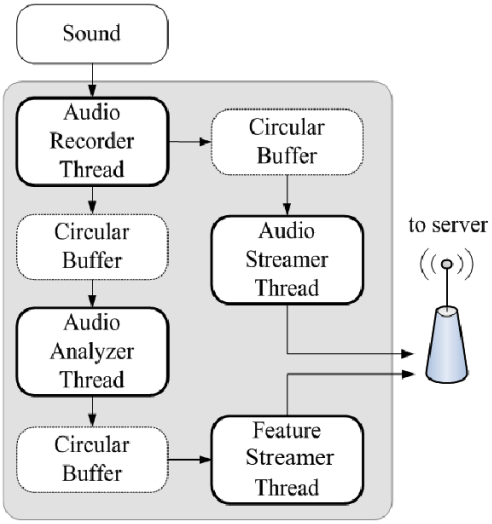

Figure 5: Block diagram of the Android software.

ister multiple RSDs enabling simultaneous streaming from multiple locations. The server currently uses MySQL to store all data and metadata. The information submitted by each RSD is stored in our database where feature values and references are stored in tables. Feature vectors include time- and frequency-domain data, a compressed representation of the audio's discrete Fourier transform, and UTC[18] timestamps. For timestamp synchronization across platforms, each RSD uses OS-specific timing libraries that provide microsecond granularity; a server-side cron job regularly updates and synchronizes its internal clock to network UTC time. Metadata information for all the RSDs is also stored in tables and includes RSD latitude and longitude drawn from the Google Maps Geolocation API[19], and the device's activity status. Additionally, the Citygram website holds a separate database of all user data, including number of RSDs and platform information. Feature pushing is accomplished through simple HTTP POST requests using the cURL library[20] and pulling is accomplished using a GET request. All requests are formatted using JSON. This relatively simple mechanism allows for consistency across all platforms.

2.5 User Interaction and Visualization

The entry point for becoming a streamer is the Citygram website[21]. The website is the central hub for management and communication between devices and users. A user registers through the site and creates a username and password. This registration information, along with details involving location and platform type, are stored in the database before being linked to specific RSDs. Within this portal, a user may edit the location of their RSDs as well as view an interactive visualization of the currently streaming devices. Beyond administrative roles, two types of users exist on the Citygram site: default user and researcher. A default user

[18] `http://www.time.gov`
[19] `https://developers.google.com/maps/documentation/business/geolocation`
[20] `http://curl.haxx.se`
[21] `http://citygram.smusic.nyu.edu`

Figure 6: Dynamic soundmap snapshot.

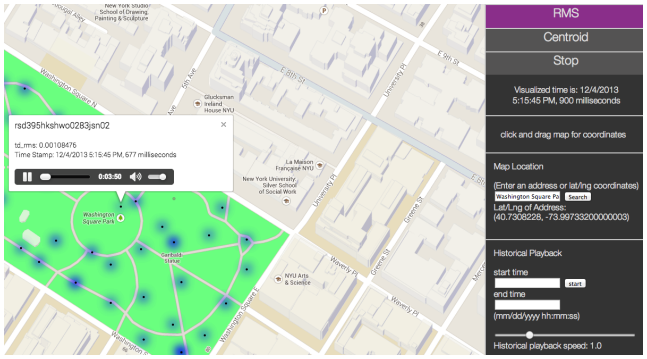

Figure 7: Screenshot of interface.

may register up to 50 RSDs and is restricted in their access to historical data. In contrast, a "research user" has no cap on the amount of devices that can be registered and may access all historical data. Additionally, researchers may also register other users.

In order for Citygram to attract potential streamers, users, urban scientists, the general public, and artists, we have developed a number of interfaces and software applications. This includes tools for commonly used software environments such as MATLAB, Max, Processing, and SuperCollider. Once registered through the Citygram website, a user may easily enter their username and device ID during the initialization of a streaming session. This will allow a number of different interaction modes with our server including pulling data from select RSDs.

The Citygram system currently provides quasi-real-time (subject to network and buffering latency) visualizations via standard web browsers such as Google Chrome. The interface dynamically visualizes RSD-streamed audio features and also provides the ability to visualize historical data stored in the database.

The web interface is designed to function as a spatio-acoustic exploration portal. By default, data from all RSDs on the map are visualized. However, it is also possible for a user to select a specific area on the map to visualize only a subset of RSDs. Our maps are built on the Google Maps API. Each RSD is represented by a marker, which is user-selectable for additional information and extended function-

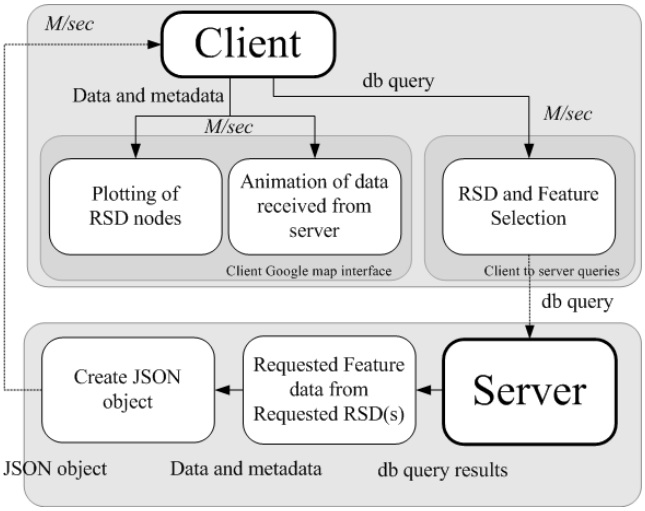

Figure 8: Data animation in Citygram.

ality. Clicking on the marker brings up a window populated by the latitude/longitude coordinates and address, real-time feature vector values, photo snapshot from Google Street View, the precise feature timestamp, and an audio monitoring button that allows the user to hear RSD-specific real-time audio streams. Our current visualizations utilize heatmaps to dynamically display low-level acoustic feature data streamed by RSDs as shown in Figure 6.

Additional features include controls that can be used to enable/disable feature visualizations. Once a user selects features for visualization, the client computer queries the server for its current timestamp and synchronizes its animation to this timecode. This timecode is used to query the list of available RSDs. A two second buffer is created, containing data from RSDs that are active for the specified time duration. This list is returned to the client, which currently animates the 2-seconds worth of samples at a rate of 10 frames per second (cf. Figure 6). These parameters are adjustable. If an RSD streams at a lower rate than the visualizer's frame rate, frame compensation takes place: a sample-and-hold technique is employed to previous samples, which are held across animation frames. Additionally, the control menu has an optional section in which to enter a start and end time for historical animation. The playback speed of historical data is controllable by a slider facilitating browsing of past data sets.

2.6 Analytics

One of the key goals of the project is to automatically analyze and classify audio data captured by our sensor network. This includes identifying urban sounds such as sirens, car horns, street music, children playing, dogs barking, wind blowing, and automobiles humming and idling. In the context of measuring and identifying sounds that can be considered noise, the initial stages of our research is framed within soundscape analysis and auditory scene analysis [7, 14, 28, 34]. This allows us to go beyond the somewhat simplistic "more decibels equal more noise" paradigm. Soundscape research typically begins with source identification [8], and as such, we believe automatic source identification to be a key research component for our project. Automatic source identification can be further divided into acoustic event detection (AED) and acoustic event classification (AEC) where the former refers to providing a semantic label [1, 13, 12] and the latter assigning temporal level segmentation that can then be used for AEC [1, 9, 12, 19].

In order to develop supervised learning systems, we require large annotated datasets. Annotated datasets for music, speech, and birds are readily available. For example, for music there exists the Million Song Dataset [2], CAL500 (500 songs with multiple annotations) [33], Last.fm (960,000 tags) [15], McGill Billboard dataset[22]; bird sound examples include HU-ASA [11], Cornell-Macaulay Library of Natural Sounds, Peterson Field Guides, Common Bird Songs, and Mirtovic's database. For general acoustic events (especially outdoor spaces) databases are currently scarce. Only a few exist and include the CLEAR[23] and C4DM D-CASE dataset [12]. Both datasets, however, only contain annotated sound samples primarily recorded within office spaces. We are, therefore, in the process of collecting and annotating data. One of the data sources we are using in the meantime is Freesound[24] which provides an API for accessing crowd-sourced and annotated audio samples that are keyword searchable (e.g. city + noise). There are a number of issues with Freesound as a source for ground-truth data including singular annotations, varying audio quality, and potentially erroneous or irrelevant labeling. In parallel to collecting and creating annotated datasets, we are also working on the development of an urban sound taxonomy to determine classes of sound sources we are interested in having our AED/AEC system identify. The taxonomy efforts will also serve in contributing towards creating a coherent framework for dealing with sounds both within the project and when relating to the existing literature on soundscape research.

3. SUMMARY

NYC is one of the largest, busiest, and most complex cities in the world. In the past two years, the city has on average received 227 noise complaint calls per day, made by city dwellers utilizing NYC's 311 hotline. In the years to come, these statistics are projected to get worse as city population growth is expected to increase significantly on a global scale. The current state of noise measurement and control infrastructures based on average dB levels are ineffective in capturing the spectral, temporal, and spatial characteristics of noise. Our early efforts and interest in cyber-physical systems to address this issue began with research and development of dynamic cartographic mapping systems to explore non-ocular urban energies. This resulted in creation of the Citygram project in 2011. In its first iteration, Citygram's focus was aimed towards acoustic energy — i.e. soundscapes that included all types of sound. More recently, in collaboration with NYU CUSP, however, we have narrowed our scope of research to spatio-acoustic noise. In this paper, we have outlined our cyber-physical system that addresses the acquisition, archival, analysis, and visualization of sound captured from urban spaces. The various components discussed in our paper included cross-platform remote sensing devices (RSDs), data acquisition and crowd-sourcing strategies for

[22] `http://ddmal.music.mcgill.ca/billboard`
[23] `http://www.clear-evaluation.org`
[24] `http://www.freesound.org`

data streaming, sensor network designs, summarizing our database architecture, our Internet exploration portal, and work concerning analytics. We expect that our research and development outcomes will contribute towards creating multimodal interactive digital maps based on poly-sensory RSDs to quantitatively measure and represent soundscapes and acoustic noise in real-time; provide interactive spatio-acoustic software tools for researchers, educators, the general public, and citizen scientists; and also contribute to urban planning, noise city code development, and improving the quality-of-life of city inhabitants.

4. FUTURE WORK

To address the complexities related to sensor network deployment and scalability, the next stage of the project will be focused on small-scale RSD deployment. Two NYC parks have been chosen for deployment of 40 RSDs. These two sites will allow the project infrastructure to be rigorously tested under outdoor weather conditions, which will help us gain valuable insights into Wi-Fi connectivity, equipment malfunction/damage, system performance under various weather conditions, and power supply issues. We also plan to set up another sensor network of 30 RSDs on the CalArts campus near Los Angeles. Our RSDs will stream in-situ raw indoor and outdoor soundscape data to our server in an effort to create a large ground truth dataset: this dataset will be annotated and labeled.

Permanently fixed mounting locations are also presently being considered. Ongoing efforts include working with cities to deploy RSDs through existing infrastructures that already include communication and power supply solutions. One such infrastructure is the urban payphone system, which will become, or already is, functionally irrelevant. NYC, for example, is planning to repurpose the payphone system: of the currently active 12,360 public payphones, 10 have recently been converted to include free Wi-Fi, with more planned for the future. This infrastructure is ideal for our project, as it would quickly lead to the availability of a large number of nodes throughout the city employing what we call "mount and play" strategy — mounting inexpensive, highly sophisticated, and robust RSDs onto existing urban infrastructures such as payphones, electronic parking meters, and traffic control systems.

The acoustic data obtained through our sensor network will also be used to investigate connections between spatio-acoustic characteristics and existing geolocated datasets, such as crime statistics, weather patterns, school attainment metrics, municipal/census data and public social network feeds, and real-estate statistics which can provide rich quantitative and contextual location-based information. Another type of information that we are interested in is environmental emotion/mood. Although this interdisciplinary research is still in its nascent stages, automatic mood detection has found increasing interest in the field of music, speech analysis, face-recognition, and natural language processing [22, 20, 30, 35]. Much of the emotion/mood detection research for sound has been in the realm of music. However, there is strong potential that algorithms and methodologies used in music will translate to urban acoustic signals as: (1) music is omnipresent in urban spaces and (2) many of the low-level feature vectors are timbral rather than musical and reflect acoustic dimensions of sound. Voice blurring will be another area for further research and development. To improve sound monitoring quality without compromising private speech that might be captured by the RSDs, voice activity detection (VAD) techniques will be explored. Currently, voice blurring occurs regardless of the absence of speech in the soundscape. With the inclusion of VAD, blurring will only occur when speech is detected by an RSD.

Another key goal for the project is developing effective and informative visualizations by embracing the idea that "a [moving] picture is worth ~~a thousand~~ [trillions of] words." In order to reach our visualization goals, we will: (1) develop robust and accurate sound classification algorithms, (2) design interactive visualizations to effectively present an ocean of data that is continuously in flux, and (3) integrate other spatial data and provide unified multi-data visualizations.

Today's megacities we inhabit are very complex. Future megacities and their soundscapes will become even more complex and will likely dwarf the noise and loudness levels of today's cities. To improve the quality-of-life of current and future city-dwellers, we need to better understand urban spaces. However, "you can't fix what you can't measure." Therefore, our hope is to contribute towards developing a cyber-physical system to dynamically and quantitatively measure and "sense" urban soundscapes and ultimately help improve the quality-of-life of city communities.

5. ACKNOWLEDGMENTS

Our thanks to Google and NYU CUSP for their support.

6. REFERENCES

[1] J.-J. Aucouturier, B. Defreville, and F. Pachet. The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, 122(2):881–91, Aug. 2007.

[2] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference, October 24-28, 2011, Miami, Florida*, pages 591–596. University of Miami, 2011.

[3] S. Böck, F. Krebs, and M. Schedl. Evaluating the Online Capabilities of Onset Detection Methods. In *ISMIR*, pages 49–54, 2012.

[4] M. Bradshaw and I. Xenakis. Formalized Music: Thought and Mathematics in Composition. *Music Educators Journal*, 59(8):85, Apr. 1973.

[5] A. L. Bronzaft. The effect of a noise abatement program on reading ability. *Journal of environmental psychology*, 1(3):215–222, 1981.

[6] a. L. Bronzaft and D. P. McCarthy. The Effect of Elevated Train Noise On Reading Ability. *Environment and Behavior*, 7(4):517–528, 1975.

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

[8] L. D. Brown, H. Hua, and C. Gao. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pages 1–10. ACM, 2003.

[9] C. V. Cotton and D. P. W. Ellis. Spectral vs. spectro-temporal features for acoustic event detection.

In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 69–72. IEEE, 2011.

[10] G. W. Evans and S. J. Lepore. Nonauditory effects of noise on children: A critical review. *Children's environments*, pages 31–51, 1993.

[11] K.-H. Frommolt, R. Bardeli, F. Kurth, and M. Clausen. *The animal sound archive at the Humboldt-University of Berlin: Current activities in conservation and improving access for bioacoustic research*. Slovenska akademija znanosti in umetnosti, 2006.

[12] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley. IEEE AASP challenge: Detection and classification of acoustic scenes and events. Technical report, Technical Report, Queen Mary University of London, 2013.

[13] 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. *submitted to Proc. EUSIPCO*, 2013.

[14] C. Guastavino. Categorization of environmental sounds. *Canadian journal of experimental psychology = Revue canadienne de psychologie expérimentale*, 61(1):54–63, Mar. 2007.

[15] V. Henning and J. Reichelt. Mendeley-A Last. fm For Research? In *eScience, 2008. eScience'08. IEEE Fourth International Conference on*, pages 327–328. IEEE, 2008.

[16] J. Joy and P. Sinclair. Networked music & soundart timeline (NMSAT): a panoramic view of practices and techniques related to sound transmission and distance listening. *Contemporary Music Review*, 28(4-5):351–361, 2009.

[17] J. J. Macionis and V. N. Parrillo. *Cities and urban life*. Pearson Education, 2004.

[18] N. Maisonneuve, M. Stevens, and M. E. Niessen. NoiseTube: Measuring and mapping noise pollution with mobile phones. *Environmental Engineering*, (May), 2009.

[19] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen. Acoustic event detection in real life recordings. In *18th European Signal Processing Conference*, pages 1267–1271, 2010.

[20] O. C. Meyers. *A mood-based music classification and exploration system*. PhD thesis, Massachusetts Institute of Technology, 2007.

[21] A. Nadakavukaren. *Our global environment: A health perspective*. Waveland Press Prospect Heights, IL, 2000.

[22] R. Nagpal, P. Nagpal, and S. Kaur. Hybrid Technique for Human Face Emotion Detection. *International Journal on Advances in Soft Computing and Its Applications (IJASCA)*, 1(6):87–90, 2010.

[23] T. H. Park, B. Miller, A. Shrestha, S. Lee, and J. Turner. Citygram One: Visualizing Urban Acoustic Ecology. *Digital Humanities 2012*, 13(6):313, 2012.

[24] T. H. Park, J. Turner, C. Jacoby, A. Marse, M. Musick, A. Kapur, and J. He. Locative Sonification: Playing the World Through Citygram. ICMC, 2013.

[25] L. R. Rabiner and B. Gold. Theory and application of digital signal processing. *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.*, 1, 1975.

[26] J. Ricard. An implementation of multi-band onset detection. *integration*, 1(2):10, 2005.

[27] C. Roads. Introduction to granular synthesis. *Computer Music Journal*, 12(2):11–13, 1988.

[28] R. M. Schafer. *The tuning of the world*. Knopf, 1977.

[29] I. Schweizer, R. Bärtl, A. Schulz, F. Probst, and M. Mühläuser. NoiseMap-real-time participatory noise maps. In *Proc. 2nd IntâĂŹl Workshop on Sensing Applications on Mobile Phones (PhoneSenseâĂŹ11)*, pages 1–5, 2011.

[30] I. Shafran and M. Mohri. A comparison of classifiers for detecting emotion from speech. In *Proc. ICASSP*, 2005.

[31] M. Sivak and S. Bao. Road safety in New York and Los Angeles: US megacities compared with the nation. 2012.

[32] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. *Cough*, 65(48):5, 2006.

[33] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 439–446. ACM, 2007.

[34] D. Wang, G. J. Brown, and Others. *Computational auditory scene analysis: Principles, algorithms, and applications*, volume 147. Wiley interscience, 2006.

[35] Y. Xia, L. Wang, and K.-F. Wong. Sentiment vector space model for lyric-based song sentiment classification. *International Journal of Computer Processing Of Languages*, 21(04):309–330, 2008.