Audio Engineering Society

Convention e-Brief 500

Presented at the 146th Convention
2019 March 20–23, Dublin, Ireland

CityTones: a Repository of Crowdsourced Annotated Soundfield Soundscapes

Agnieszka Roginska[1], Hyunkook Lee[2], Ana Elisa Mendez Mendez[1], Scott Murakami[1], Andrea Genovese[1]

[1] *New York University, 35 West 4th St. New York NY USA*

[2] *University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, UK*

Correspondence should be addressed to Agnieszka Roginska (roginska@nyu.edu) and Hyunkook Lee (H.Lee@hud.ac.uk)

ABSTRACT

Immersive environmental soundscape capture and annotation is a growing area of audio engineering and research, with applications in the reproduction of immersive sound experiences in AR and VR, sound classification, and environmental sound archiving. This Engineering Brief introduces CityTones as a crowdsourced repository of soundscapes captured using immersive sound capture methods, that the audio community can contribute to. The database will include descriptors containing information about the technical details of the recording, physical information, subjective quality attributes, as well as sound content information.

1 Introduction

The capture and analysis of soundscapes and environmental sounds has been an area of increased interest in the recent past. Emerging research projects utilizing small- and large-scale microphone arrays have contributed to our understanding of the behavior and patterns of sounds in urban, and non-urban environments, and have been utilized in applications of immersive experiences. Of particular interest has been the capture of not only content, but also the spatial characteristic of environments and sources contained within them, and their immersive qualities. These have been enabled by the greater accessibility of sound capture technologies and equipment, and a more widely available network of sensors and distributed microphone arrays.

CityTones is a crowdsourced project of immersive soundscapes captured by audio engineers. The database will include descriptors about the recordings and the contents of the recordings. The technical information will describe the sound capture method, microphones, and placement; the physical information describes the location of the recording, environmental factors, and level of the recording; the subjective quality attributes contain information about perceptual factors as reported by the audio engineer; and the sound annotations will contain information about specific sounds and the content of the recordings.

Over the past two decades, there has been an increase in research related to the investigation of the capture, analysis and visualization of urban and non-urban soundscapes, and more recently into capturing soundscapes using Ambisonic methods. IHearNY3D [1] and IHearBangalore3D [2] used ambisonic recordings of various Manhattan and Bangalore locations and allowed listeners to navigate around them. At the time of the recording, SPL measurements were taken - information which could be used to calibrate to the same SPL level during playback. SoNYC [3] is a large-scale sound monitoring project in New York City that uses a distributed array of sensors, machine learning technology and big data analysis in order to monitor, analyze and, ultimately, mitigate noise pollution. The Sounds in the City project [4] aims to capture and research the urban soundscape in Montreal,

Canada. The Urban Soundscapes of the World project [5] compiles a database of high-quality audio-visual recordings of a selected group of locations of urban environments based on the perceptual properties of the soundscape.

2 Applications

The CityTones project has multiple practical and research applications. As recordings submitted will be publicly available for users to download it can be used for simulation of environments and sound design and also in research areas such as audio engineering, human computer interaction and machine listening. More specifically, in audio engineering we can use the data collected to understand the techniques used by those who submitted recordings, learn new approaches and the reasons behind their selections. Gaining insight on this area can also help develop new techniques. From the human computer interaction perspective, we can use crowdsourcing techniques to attract users in both the recording collection stage and the annotation/labelling stage. It is important to establish a clear motivation that will attract users to the system and develop a task that will keep users engaged. Finally, with annotations collected from the crowd we can build a labelled dataset that can later be used in machine listening research to train models that can identify sound sources. These can go from something broad as differentiating music from urban sounds, or as specific as identifying musical instruments.

3 Recording technique

3.1 Microphone system

The microphone system to be used for audio recording must be compatible for 360° audio rendering in the 1st order B-format. Any of the currently available First Order Ambisonics (FOA) microphone systems are accepted (e.g., Sennheiser Ambeo, Core Sound TetraMic, Rode NTSF-1, Zoom H3-VR, etc.). Raw recordings made using certain Ambisonic microphones are in the form of A-format. They should be converted into B-format (W, X, Y and Z), which is the submission format, using a dedicated software plugin provided by the microphone manufacturer. If a Higher Order Ambisonics (HOA) system (e.g. mh Acoustics Eigenmic, Zylia ZM-1, Core Sound Octomic, etc.) is used, the recording must be converted into the 1st order B-format for submission.

A multichannel spaced microphone array designed for 360° audio capture can also be used (e.g., ESMA-3D [6], Schoeps ORTF-3D, etc.). Signals captured by such an array must be encoded in the 1st order B-format in such a way that they are discretely mapped to their corresponding loudspeaker positions of the original reproduction format. For example, the four main channel signals of an ESMA-3D should be mapped to ±45° and ±135° azimuths at 0° elevation, with the height channel signals to ±45° and ±135° at 30° to 45° elevation. A number of Ambisonic encoder software plugins are freely available (e.g., IEM plugin suite, SPARTA, The Ambisonic Toolkit, etc.). Recent research by Lee et al. [7] found that, in binaural headphone reproduction, multichannel spaced microphone array recordings that were rendered in FOA had only a little degradation from the original recording in terms of both spatial and timbral fidelity, depending on the decoder used and the type of sound source.

3.2 Visual Media

A spherical visual recording or at least a panoramic picture must accompany the audio recording in order to provide all-around visual information about the recording location. Using a 360° video camera with a minimum 4K (3840 x 1920) resolution is recommended (e.g., GoPro Fusion, Ricoh Theta V, Insta360 One X, Samsung Gear360, etc.). The frame rate (fps) can vary depending on the region, but the highest possible fps is recommended (e.g., 50 fps or 60 fps).

3.3 Microphone/camera placement

For the three-dimensional (3D) reproduction of an FOA recording using the popular ‘cube’ loudspeaker format [8], the microphone should ideally be placed at a typical ear height in order to realistically render the perceived elevations of sound sources from the listener’s perspective. If the microphone is higher than the sound source that is desired to be perceived

at the ear height, the perceived image of the source will be lower than the ear level theoretically. In practice, however, the perceptual resolution of image elevation in FOA is not high as evident in [9]. For example, the encoding of a negative elevated source at -22.5° would cause the image to be perceived at a position that is not significantly different from 0° elevation in binaural headphone reproduction using anechoic head-related impulse responses. It would sometimes be necessary to raise the microphone to about 2m to 2.5m in order to avoid physical contact from people, especially when recording in a crowded place (e.g., busy street). In such a case, it is recommended that the distance between the microphone to the target source is kept at least about 4m so that the perceived negative elevation of the target ear-height source is minimised.

On the other hand, 360° spaced microphone arrays such as ESMA-3D or ORTF-3D are designed to be raised higher than target sound sources. The main microphones (cardioids or supercardioids) of such arrays should be tilted downwards at -10° to -20° (depending on the microphone-to-source distance) to have an on-axis response for a target sound source, whereas their height microphones (cardioids or supercardioids), which are vertically coincident to the main microphones based on [10], should be angled upwards to achieve sufficient level difference to the main microphone signals (i.e., at least -7 dB [11] for optimal vertical localisation. It is recommended to raise the array up to about 2.5m or higher from the ground level if possible. This will ensure the main microphones tilted at -10° to -15° to have an on-axis response for a sound source that is about 1.6m high and about 3m to 4m away from the microphone array. This will also help avoid the array to be visible in 360° video.

The video camera should be placed at a typical eye height (e.g., 160 – 180 cm) in order to provide a realistic visual perspective.

3.4 Recording practice

The audio should be recorded digitally in the PCM wave format at a sampling rate of 48 kHz with a bit depth of 24 bits.

Recording level should be carefully set according to the average and peak loudness of the location in order to prevent clipping during recording. In an environment where the dynamic range of sound pressure level (SPL) is large, it is recommended to set up gains so that the peak level does not exceed -10 dBFS. In an environment where the SPL is relatively constant, it is recommended to record signals at around -18 dBFS.

The microphone system and video camera must be mounted on a microphone stand rather than being hold by a hand. It is essential to use both windjammer and windshield in a windy weather condition. It is recommended to use the windshield regardless of the weather condition for the protection of the microphone.

It is recommended to use a battery-powered mobile recorder and place it as close to the microphone system as possible (e.g., at the bottom of the microphone stand). It is not ideal to use long microphone cables to control the recording system from a distance, especially in crowded urban environments. During recording, the recordist should stay away from the microphone system and video camera to avoid unwanted noise and visual disturbance.

If taking a video, slating is necessary in order to easily sync the audio and video in post processing. Slating can be done via a clapboard or clapping three times while in clear view of the camera as well as near the microphones for a clear signal for syncing. Slating at the beginning of the recording is encouraged, although tail slating is acceptable as well.

3.5 Post-Processing

If the audio recording contains too much low frequency energy below around 100 Hz, a low-pass filter could be applied for correction. Parametric equalization should be applied only for a tonal balance correction purpose. Excessive amount of

noise in recording (e.g., wind, rumble, hum, etc.) could also be reduced using an appropriate noise reduction tool. Dynamic range limiting or compression should not be applied. The recordist

should ensure that digital clipping is avoided in the recording stage.

Level normalisation should be applied to the B-format waveforms as a group so that the peak level of the W component becomes -6 dBFS. Finally, it is recommended to apply a short fade in and out to the final waveforms with an equal duration for all B-format components. The file format for submission must be wave at 48 kHz/24 bits.

Video recordings made using a spherical camera should be stitched and rendered in 2:1 aspect ratio using appropriate software. The final file for submission must be created in MP4 using the H.264 codec.

4 Descriptors

Descriptors of the recordings are an essential part of this project, as they will help categorize the recordings in a way that is useful for individuals using these in their research. Descriptors will be collected in two separate stages: when collecting the recordings, and during crowdsourcing in order to obtain more detailed information.

When collecting recordings, users will fill out a form where they will be asked to enter a series of descriptors that correspond to information accessible at the time of the recording, including:

- Technical details
- Physical information
- Subjective quality attributes

These recordings can be useful for building an annotated dataset which can be used to train machine learning models. The annotations of the recordings will be crowdsourced by asking crowd workers to identify categories that are considered useful when training models. The physical and subjective attributes are presented in Table 1.

5 Submission Methods

The audio/video recordings for CityTones will be temporarily housed in a repository through the NYU Immersive Audio Group website, until a dedicated server and website can be established. The Immersive Audio Group website is as follows: https://wp.nyu.edu/immersiveaudiogroup/citytones/

Physical Attributes	
Date and Time	Date and time when the recording was made
Location	Precise latitude and longitude where the recording was made
Location description	Type of location where the recording was made (e.g. urban, suburban, woods)
Inside/Outside	Confined space or outside
Weather	Temperature, humidity, conditions (e.g. sunny, rainy)
Microphone technique	e.g. FOA, HOA, ESMA-3D, ORTF-3D, etc.
Sound pressure level (SPL)	Long term average, A-weighted (LAeq)
Subjective quality attributes	
Eventfulness	Acoustic density
Spatial impression	Source/environmental width, depth and height
Annoyance	Pleasant to annoying scale
Energetic	Level of energy
Loudness	Perceptual impression: quiet to loud scale
Proximity	Proximity of sources
Sound categories	
Human	e.g. crowd, talking, footsteps
Sports	e.g. baseball, soccer, football
Animals	e.g. Birds, dogs, cats, horse
Nature	Wind, fire, water, leaves
Music	Radio, live concert, music in vehicle
Engines & vehicles	Small engine (power tools, lawn mower, chainsaw) Medium engine (cars, motorcycle) Large engine (bus, train, plane, truck, diesel generator)
Alert signals	Sirens, car horn, ambulance, police car siren
Construction & impact sounds	Jackhammer, metal, backhoe, dumping

Table 1. Descriptors of recordings.

The submission portal will also be via the Immersive Audio Group website. The submission process involves a survey and submission of recording(s) via the CityTones section of the Immersive Audio Group website. The survey will be a Google form in which submitters will provide the necessary information for their submission, including technical details, physical and subjective quality attributes.

All submitted recordings and pertinent information shall be submitted as a single zip file using the following naming convention. The submission will be evaluated and certified before the recording is added to the database.

Country_City_DDMMYYYY.zip, where the Country and City represent the location where the recording was captured. For example, a recording taken in New York City, USA on May 9, 2016 would be stored as USA_NewYork_12052016.zip. In the zip file there should be four audio files representing the B-format audio components of W, X, Y and Z, and the movie or accompanying picture contained in files named as follows:

Country_City_DDMMYYYY_W.wav
Country_City_DDMMYYYY_X.wav
Country_City_DDMMYYYY_Y.wav
Country_City_DDMMYYYY_Z.wav
Country_City_DDMMYYYY.mp4 or jpeg (if picture)

The duration of the recordings must be a minimum of 3 minutes in length, with no specified maximum time limit. A length of about 5 minutes for each recording is recommended.

Submissions will be reviewed by a panel of selected CityTones reviewers, and if accepted will be added to the database for use. CityTones recording engineers will be certified through this process. Future plans include a dedicated, public and open source repository for which CityTones can be housed.

6 References

[1] B. Boren, A. Andreopoulou, M. Musick, H. Mohanraj, and A. Roginska, "I Hear NY3D: Ambisonic Capture and Reproduction of an Urban Sound Environment," in Proceedings of the 135th Audio Engineering Society Convention, New York, NY, 2013.

[2] S. Aswathanarayana, A. Roginska. "I Hear Bangalore3D: Capture and Reproduction of urban sounds of Bangalore using an Ambisonic Microphone", Proceedings of the 20th International Conference on Auditory Display, New York, NY, 2014

[3] J.P. Bello, C. Silva, O. Nov, R.L. DuBois, A. Arora, J. Salamon, C. Mydlarz, H. Doraiswamy. "SONYC: A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution", CoRR, abs/1805.00889, 2018.

[4] https://www.sounds-in-the-city.org/ accessed Jan. 17, 2019.

[5] B. De Coensel, K. Sun, D. Botteldooren "Urban soundscapes of the world: selection and reproduction of urban acoustic environments with soundscape in mind." *Proceedings of the 46th International Congress and Exposition on Noise Control Engineering*. Inter-Noise, 2017.

[6] H, Lee, "Capturing 360° Audio Using an Equal Segment Microphone Array (ESMA)," *J. Audio Eng. Soc., to be published in Jan/Feb 2019 issue,* vol. 67, (2019).

[7] H. Lee, M. Frank and F. Zotter, "Spatial and Timbral Fidelities of Binaural Ambisonics Decoders," Presented in AES Conference on Immersive and Interactive Audio, 2019.

[8] https://www.york.ac.uk/sadie-project/ambidec.html, accessed 17 Jan 2019.

[9] C. Millns, M. Mironovs and H. Lee, "Vertical Localisation Accuracy of Binauralised First Order Ambisonics across Multiple Horizontal Positions," Presented in 146th AES Convention (2019).

[10] H. Lee and C. Gribben, "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array," *J. Audio Eng. Soc., vol. 62*, pp. 870–884 (2014).

[11] R. Wallis and H. Lee, "The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localization Thresholds for Natural Sound Sources," *Appl. Sci.*, vol. 7, p. 278 (2017).