
CityNet - Deep Learning Tools for Urban Ecoacoustic Assessment

A. J. Fairbrass\textsuperscript{1,2,3,\textdagger,\ast}, M. Firman\textsuperscript{4, \textdagger,\ast}, C. Williams\textsuperscript{3}, G. J. Brostow\textsuperscript{4}, H. Titheridge\textsuperscript{1}, and K. E. Jones\textsuperscript{2, 5, \ast}

\textsuperscript{1}Centre for Urban Sustainability and Resilience, Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT, United Kingdom.

\textsuperscript{2}Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, United Kingdom.

\textsuperscript{3}Bat Conservation Trust, 5th floor, Quadrant House, 250 Kennington Lane, London, SE11 5RD, United Kingdom.

\textsuperscript{4}Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, United Kingdom.

\textsuperscript{5}Institute of Zoology, Zoological Society of London, Regent’s Park, London, NW1 4RY, United Kingdom.

\textdagger Corresponding authors: alison.fairbrass.10@ucl.ac.uk, michael.firman.10@ucl.ac.uk, and kate.e.jones@ucl.ac.uk (Tel: +44 (0)20 31084230)

\textdagger Denotes joint first authorship

Running title: Deep Learning Urban Ecoacoustic Tools

Word Count (Items): Total 6923, Summary 346, Main text 4381, References 1659 (58),

Tables 151 (2), Figures 386 (4).
SUMMARY

1. Cities support unique and valuable ecological communities, but understanding urban wildlife is limited due to the difficulties of assessing biodiversity. Ecoacoustic surveying is a useful way of assessing habitats, where biotic sound measured from audio recordings is used as a proxy for biodiversity. However, existing algorithms for measuring biotic sound have been shown to be biased by non-biotic sounds in recordings, typical of urban environments.

2. We develop CityNet, a deep learning system using convolutional neural networks (CNNs), to measure audible biotic (CityBioNet) and anthropogenic (CityAnthroNet) acoustic activity in cities. The CNNs were trained on a large dataset of annotated audio recordings collected across Greater London, UK. Using a held-out test dataset, we compare the precision and recall of CityBioNet and CityAnthroNet separately to the best available alternative algorithms: four acoustic indices (AIs): Acoustic Complexity Index, Acoustic Diversity Index, Bioacoustic Index, and Normalised Difference Soundscape Index, and a state-of-the-art bird call detection CNN (bulbul). We also compare the effect of non-biotic sounds on the predictions of CityBioNet and bulbul. Finally we apply CityNet to describe acoustic patterns of the urban soundscape in two sites along an urbanisation gradient.

3. CityBioNet was the best performing algorithm for measuring biotic activity in terms of precision and recall, followed by bulbul, while the AIs performed worst. CityAnthroNet outperformed the Normalised Difference Soundscape Index, but by a smaller margin than CityBioNet achieved against the competing algorithms. The CityBioNet predictions were impacted by mechanical sounds, whereas air traffic and wind sounds influenced the bulbul predictions. Across an urbanisation gradient, we
show that CityNet produced realistic daily patterns of biotic and anthropogenic
acoustic activity from real-world urban audio data.

4. Using CityNet, it is possible to automatically measure biotic and anthropogenic
acoustic activity in cities from audio recordings. If embedded within an autonomous
sensing system, CityNet could produce environmental data for cites at large-scales
and facilitate investigation of the impacts of anthropogenic activities on wildlife. The
algorithms, code and pre-trained models are made freely available in combination
with two expert-annotated urban audio datasets to facilitate automated environmental
surveillance in cities.

Keywords: Acoustic Indices, Anthropogenic, Biodiversity Assessment, Convolutional Neural
Networks, Deep Learning, Ecoacoustics, London, Machine Learning, Soundscapes, Urban
Ecology.
INTRODUCTION

Over half of the world’s human population now live in cities (UN-DESA 2016) and urban biodiversity can provide people with a multitude of health and well-being benefits including improved physical and psychological health (Natural England 2016; Crouse et al. 2017). Cities can support high biodiversity including native endemic species (Aronson et al. 2014), and act as refuges for biodiversity that can no longer persist in intensely managed agricultural landscapes surrounding cities (Hall et al. 2016). However, our understanding of urban biodiversity remains limited (Faeth, Bang & Saari 2011; Beninde, Veith & Hochkirch 2015). One reason for this is the difficulties associated with biodiversity assessment, such as gaining repeated access to survey sites and the resource intensity of traditional methods (Farinha-Marques et al. 2011). This inhibits our ability to conduct the large-scale assessment that is necessary for understanding urban ecosystems.

Ecoacoustic surveying has emerged as a useful method of large-scale quantification of ecological communities and their habitats (Sueur & Farina 2015). Passive acoustic recording equipment facilitates the collection of audio data over long time periods and large spatial scales with fewer resources than traditional survey methods (Digby et al. 2013). A number of automated methods have been developed to measure biotic sound in the large volumes of acoustic data that are typically produced by ecoacoustic surveying (Sueur & Farina 2015). For example, Acoustic Indices (AIs) use the spectral and temporal characteristics of acoustic energy in sound recordings to produce whole community measures of biotic and anthropogenic sound (Sueur et al. 2014). However, several commonly used AIs have been shown to be biased by non-biotic sounds (Towsey et al. 2014; Fuller et al. 2015; Gasc et al. 2015a), and are not suitable for use in the urban environment without the prior removal of certain non-biotic sounds from recordings (Fairbrass et al. 2017).
Machine learning (ML) is being increasingly applied to biodiversity assessment and monitoring because it facilitates the detection and classification of ecoacoustic signals in audio data (Acevedo et al. 2009; Walters et al. 2012; Stowell & Plumbley 2014). Using annotated audio datasets of soniferous species, a ML model can be trained to recognise biotic sounds based on multiple acoustic characteristics, or features, and to associate these features with taxonomic classifications, and can then assign a probabilistic classification to sounds within recordings. AIs only use a limited number of acoustic features in their calculations, such as spectral entropy within defined frequency bands (Boelman et al. 2007; Villanueva-Rivera et al. 2011; Kasten et al. 2012) or entropy changes over time (Pieretti, Farina & Morri 2011). Additionally, the relationship between the features and the algorithm outputs are chosen by a human, rather than learned automatically from an annotated dataset. In contrast, ML algorithms can utilise many more features in their calculations, and the relationship between inputs and outputs is determined automatically based on the annotated training data provided. Convolutional Neural Networks, CNNs (or Deep learning) (LeCun, Bengio & Hinton 2015) can even choose, based on the annotations in the training dataset, the features that best discriminate different classes in datasets without being specified a priori, and can take advantage of large quantities of training data where their ability to outperform human defined algorithms increases as more labelled data become available.

Species-specific ML algorithms have been developed to automatically identify the sounds emitted by a range of soniferous organisms including birds (Stowell & Plumbley 2014), bats (Walters et al. 2012; Zamora-Gutierrez et al. 2016), amphibians (Acevedo et al. 2009) and invertebrates (Chesmore & Ohya 2004). However, these algorithms are focussed on a small number of species limiting their usefulness for broad classification tasks across communities. More recently, algorithms that detect whole taxonomic groups are being developed, for example bird sounds in audio recordings from the UK and the Chernobyl Exclusion Zone.

(Grill & Schlüter 2017), but these algorithms remain untested on noisy audio data from urban environments. There are currently no algorithms that produce whole community measures of biotic sound that are known to be suitable for use in acoustically complex urban environments.

Here, we develop the CityNet acoustic analysis system, which uses two CNNs for measuring audible (0-12 kHz) biotic (CityBioNet) and anthropogenic (CityAnthroNet) acoustic activity in audio recordings from urban environments. We use this frequency range as it contains the majority of sounds emitted by audible soniferous species in the urban environment (Fairbrass et al. 2017). The CNNs were trained using CitySounds2017, an expert-annotated dataset of urban sounds collected across Greater London, UK that we develop here. We compared the performance of CityNet using a held-out dataset by comparing the algorithms’ precision and recall to four commonly used AIs: Acoustic Complexity Index (ACI) (Pieretti, Farina & Morri 2011), Acoustic Diversity Index (ADI) (Villanueva-Rivera et al. 2011), Bioacoustic Index (BI) (Boelman et al. 2007), Normalised Difference Soundscape Index (NDSI) (Kasten et al. 2012), and to bulbul, a state-of-the-art algorithm for detecting bird sounds in order to summarise avian acoustic activity (Grill & Schlüter 2017). As the main focus of the study was the development of algorithms for ecoacoustic assessment of biodiversity in cities, we conducted further analysis on the two best performing algorithms for measuring biotic sound, CityBioNet and bulbul, by investigating the effect of non-biotic sounds on the accuracy of the algorithms. Finally, we applied CityNet to investigate daily patterns of biotic and anthropogenic sound in the urban soundscape.
MATERIALS AND METHODS

We developed two CNN models, CityBioNet and CityAnthroNet within the CityNet system to generate measures of biotic and anthropogenic sound, respectively. The CityNet pipeline (Figure 1) consisted of 7 main steps as follows:

1. **Record audio**: Audible frequency (0-12 kHz) .wav audio recordings were made using a passive acoustic recorder.

2. **Audio conversion to Mel spectrogram**: Each audio file was automatically converted to a Mel spectrogram representation with 32 frequency bins, represented as rows in the spectrogram, using a temporal resolution of 21 columns per second of raw audio. Before use in the classifier, each spectrogram $S$ was converted to a log-scale representation, using the formula $\log(A + B \times S)$. For biotic sound detection the parameters $A = 0.001$ and $B = 10.0$ were used, while for anthropogenic sound detection the parameters $A = 0.025$ and $B = 2.0$ were used.

3. **Extract window from spectrogram**: A single input to the CNN comprised a short spectrogram chunk $Ws$, 21 columns in width, representing 1 second of audio.

4. **Apply different normalisation strategies**: There are many different methods for pre-processing spectrograms before they are used in ML; for example whitening (Lee et al. 2009) and subtraction of mean values along each frequency bin (Aide et al. 2013). CNNs are able to accept inputs with multiple channels of data, for example the red, green and blue channels of a colour image. We exploited the multiple input channel capability of our CNN by providing as input four spectrograms each pre-processed using a different normalisation strategy (see Supplementary Methods), which gave considerable improvements to network accuracy above any single normalisation scheme in isolation. After applying different normalisation strategies, the input to the network consisted of a $32 \times 21 \times 4$ tensor.
(5) Apply CNN classifier: As described above, classification was performed with a CNN, whose parameters were learnt from training data. The CNN comprised a series of layers, each of which modified its input data with parameterised mathematical operations which were optimised to improve classification performance during training (see Supplementary Methods for details). The final layer produced the prediction of presence or absence of biotic or anthropogenic sound.

(6) Make prediction for each moment in time: At test time, steps (3)-(5) were repeated every 1 second throughout the audio file, to give a measure of biotic or anthropogenic activity throughout time. Predictions for each chunk of audio were made independently.

(7) Summarise: Where appropriate, the chunk-level predictions were summarised to gain insights into trends over time and space. For example, predicted activity levels for each half-hour window could be averaged to inspect the level of biotic and anthropogenic activity at different times of day.

The ML pipeline was written in Python v.2.7.12 (Python Software Foundation 2016) using Theano v.0.9.0 (The Theano Development Team et al. 2016) and Lasagne v.0.2 (Dieleman et al. 2015) for ML and librosa v.0.4.2 (McFee et al. 2015) for audio processing.

Acoustic Dataset

We selected 63 green infrastructure (GI) sites in and around Greater London, UK to collect audio data to train and test the CityNet algorithms. These sites represent a range of GI in and around Greater London in terms of GI type, size and urban intensity. Each site was sampled for 7 consecutive days systematically across the months of May to October between 2013 and 2015 (Figure 2, Table S1). At each location, a Song Meter SM2+ digital audio field sensor (Wildlife Acoustics, Inc., Concord, Massachusetts, USA) was deployed, recording sound between 0 and 12 kHz at a 24 kHz sample rate. The sensor was equipped with a single...
omnidirectional microphone (frequency response: \(-35\pm4\) dB) oriented horizontally at a height of 1m. Files were saved in .wav format onto a SD card. Audio was recorded in computationally manageable chunks of 29 minutes of every 30 mins (23.2 hours of recording per day), which were divided into 1-minute audio files using Slice Audio File Splitter (NCH Software Inc. 2014), leading to a total of 613,872 discrete minutes of audio recording (9,744 minutes for each of the 63 sites). This constituted the CitySounds2017 dataset.

Acoustic Training Dataset

To create our training dataset (CitySounds2017\textsubscript{train}) we randomly selected twenty five 1-minute recordings from 70\% of the study sites (44 sites, 1100 recordings). A.F. manually annotated the spectrograms of each recording, computed as the log magnitude of a discrete Fourier transform (non-overlapping Hamming window size=720 samples=10 ms), using AudioTagger (available at https://github.com/groakat/AudioTagger). Spectrograms were annotated by localising the time and frequency bands of discrete sounds by drawing bounding boxes as tightly as visually possible within spectrograms displayed on a Dell UltraSharp 61cm LED monitor. Types of sound, such as “invertebrate”, “rain”, and “road traffic”, were identified by looking for typical patterns in spectrograms (Figure S1), and by listening to the audio samples represented in the annotated parts of the spectrogram. Categories of sounds were then grouped into biotic, anthropogenic and geophonic classes following Pijanowski \textit{et al.} (2011), where we define biotic as sounds generated by non-human biotic organisms, anthropogenic as sounds associated with human activities, and geophonic as non-biological ambient sounds e.g. wind and rain.

Acoustic Testing Dataset and Evaluation

To evaluate the performance of the CityNet algorithms, we created a testing dataset (CitySounds2017\textsubscript{test}) by strategically selecting 40 recordings from CitySounds2017 from the
remaining 30% of sites (19 sites) that contained a range of both biotic and anthropogenic acoustic activity. CitySounds2017\textsubscript{test} was sampled from different recording sites to CitySounds2017\textsubscript{train} to demonstrate that the CityNet algorithms generalise to sounds recorded at new site locations (Figure 2, Table S1). To optimise the quality of the annotations in CitySounds2017\textsubscript{test}, we selected five human labellers to separately annotate the sounds within the audio recordings (using the same methods as above) to create a single annotated test dataset. Conflicts were resolved using a majority rule, and in cases where there was no majority, we used our own judgement on the most suitable classification. Our CitySounds2017 annotated training and testing datasets are available at \url{https://figshare.com/s/adab62c0591afaeafedd}.

Using the CitySounds2017\textsubscript{test} dataset, we separately assessed the performance of the two CityNet algorithms, CityBioNet and CityAnthroNet, using two measures: precision and recall. The CityBioNet and CityAnthroNet algorithms give a probabilistic estimate of the level of biotic or anthropogenic acoustic activity for each 1-second audio chunk as a number between 0 and 1. Different thresholds could be used to convert these probabilities into sound category assignments (e.g. ‘sound present’ or ‘sound absent’). At each threshold, a value of precision and recall was computed, where precision was the fraction of 1-second chunks correctly identified as containing the sound according to the annotations in CitySounds2017\textsubscript{test}, and recall was the fraction of 1-second chunks labelled as containing the sound which was retrieved by the algorithm under that threshold. As the threshold was swept between 0 and 1, the resulting values of precision and recall were plotted as a precision-recall curve. Summary statistics were computed for the average precision under all the threshold values and the recall when the threshold chosen gave a precision of 0.95. Using a threshold of 0.5 on the predictions, confusion matrices were calculated showing how each moment of time was classified relative to the annotations. These analyses were conducted in Python v.2.7.12.

(227) Python Software Foundation (2016) using Scikit-learn v.0.18.1 (Pedregosa et al. 2011) and Matplotlib v.1.5.1 (Hunter 2007).

Competing Algorithms

We also compared the precision and recall of the CityNet algorithms to acoustic measures produced by four AIs: Acoustic Complexity Index (ACI) (Pieretti, Farina & Morri 2011), Acoustic Diversity Index (ADI) (Villanueva-Rivera et al. 2011), Bioacoustic Index (BI) (Boelman et al. 2007), and Normalised Difference Soundscape Index (NDSI) (Kasten et al. 2012). The NDSI generates a measure of anthropogenic disturbance according to the formula

\[
NDSI = \frac{NDSI_{bio} - NDSI_{anthro}}{NDSI_{bio} + NDSI_{anthro}} \tag{1}
\]

where NDSI\textsubscript{bio} and NDSI\textsubscript{anthro} are the total biotic and anthropogenic acoustic activity in each recording, respectively. Rather than compare CityNet to the NDSI, we compared the biotic (NDSI\textsubscript{bio}) and anthropogenic (NDSI\textsubscript{anthro}) elements of the NDSI to the measures produced by CityBioNet and CityAnthroNet, respectively, as these were more comparable. As the AIs are all designed to give a summary of acoustic activity for an entire file, they were analysed on the CitySounds2017\textsubscript{test} dataset by treating each 1-second chunk of audio as a separate sound file to enable direct comparisons to CityNet. The AI measures do not have a natural threshold for classification into biotic/non-biotic sound, meaning we could not calculate confusion matrices. However, a threshold between their lowest value and their highest value was used in combination with the range of precision and recall values to form precision-recall curves. All AIs were calculated in R v.3.4.1 (R Core Team 2017) using the ‘seewave’ v.1.7.6 (Sueur, Aubin & Simonis 2008) and ‘soundecology’ v.1.2 (Villanueva-Rivera & Pijanowski 2014) packages.
The precision and recall of CityBioNet was also compared to bulbul (Grill & Schlüter 2017), an algorithm for detecting bird sounds in entire audio recordings in order to summarise avian acoustic activity which was the winning entry in the 2016-7 Bird Audio Detection challenge (Stowell et al. 2016). Like CityNet, bulbul is a CNN-based classifier which uses spectrograms as input. However, it does not use the same normalisation strategies as CityNet, and it was not trained on data from noisy, urban environments. Bulbul was applied to each second of audio data in CitySounds2017_{test}, using the pre-trained model provided by the authors together with their code.

Impact of Non-Biotic Sounds

We conducted additional analysis on the non-biotic sounds that affect the predictions of CityBioNet and bulbul, as these were found to be the best performing algorithms for measuring biotic sound. To do this, we created subsets of the CitySounds2017_{test} dataset comprising all the seconds that contained a range of non-biotic sounds, e.g. a road traffic data subset containing all of the seconds in CitySounds2017_{test} where the sound of road traffic was present. We then used a Chi-squared test to identify significant differences in the proportion of seconds in which the presence/absence of biotic sound at threshold 0.5 was correctly predicted in the full and subset datasets by each algorithm, and the Cramer’s V statistic was used to assess the effect size of differences (Cohen 1992). These analyses were conducted in R v.3.4.1 (R Core Team 2017).

Ecological Application

We used CityNet to generate daily average patterns of biotic and anthropogenic acoustic activity for two study sites across an urbanisation gradient (sites E29RR and IG62XL with high and low urbanisation respectively, Table S1). To control for the date of recording; both sites were surveyed between May and June 2015. CityNet was run over the entire 7 days of
recordings from each site to predict the presence/absence of biotic and anthropogenic sound for every 1-second audio chunk using a 0.5 probability threshold. Measures of biotic and anthropogenic activity were created for each half hour window between midnight and midnight by averaging the predicted number of seconds containing biotic or anthropogenic sound within that window over the entire week.

RESULTS

Acoustic Performance

CityBioNet had an average precision of 0.934 and recall of 0.710 at 0.95 precision, while CityAnthroNet had an average precision of 0.977 and recall of 0.858 at 0.95 precision (Table 1, Figure 3). In comparison the ACI, ADI, BI and NDSI.bio had a lower average precision (0.663, 0.439, 0.516, and 0.503, respectively) and lower recall at 0.95 (all less than 0.01). CityBioNet also outperformed bulbul which had an average precision of 0.872 and recall at 0.95 of 0.398 (Table 1). In comparison to CityAnthroNet, the NDSI.anthro had a lower average precision (0.975) and lower recall at 0.95 precision (0.815). When biotic sound was present in recordings, CityBioNet correctly predicted the presence of biotic sound (True Positives) in a greater proportion of audio data than bulbul (33.2% in comparison with 18.5%, for CityBioNet and bulbul respectively) (Figure 4). However, CityBioNet failed to correctly predict the presence of biotic sound (False Negatives) in 1.7% of recordings in comparison with 1.0% incorrect predictions by bulbul. When biotic sound was absent from recordings, CityBioNet correctly predicted the absence of biotic sound (True Negatives) in 51.6% of the audio data in comparison with 52.6% for bulbul, and CityBioNet failed to correctly predict the absence of biotic sound (False Positives) in 13.5% of audio data in comparison with 20.0% incorrect predictions by bulbul (Figure 4).
Impacts of Non-Biotic Sounds

CityBioNet was strongly (Cramer’s V effect size >0.5) negatively affected by mechanical sound (the presence/absence of biotic sound was correctly predicted in 28.60% less of the data when mechanical sounds were also present) (Table 2). Bulbul was moderately (Cramer’s V effect size 0.1-0.5) negatively affected by the sound of air traffic and wind (the presence/absence of biotic sound was correctly predicted in 5.34% and 6.93% less of the data when air traffic and wind sounds were also present in recordings, respectively).

Ecological Application

CityNet produced realistic patterns of biotic and anthropogenic acoustic activity in the urban soundscape at two study sites of low and high urban intensity (Figure 2B and C). At both sites, biotic acoustic activity peaked just after sunrise and declined rapidly after sunset. A second peak of biotic acoustic activity was recorded at sunset at the low urban intensity site but not at the high urban intensity site. At both sites anthropogenic acoustic activity rose sharply after sunrise, remained constant throughout the day and declined after sunset.

DISCUSSION

Both CityBioNet and CityAnthroNet outperformed the competing algorithms on the CitySound2017\textsubscript{test} dataset. CityBioNet performed better than bulbul on noisy recordings from the urban environment; it was robust to more non-biotic sounds, including road traffic, air traffic and rain. Being robust to the sound of road traffic supports the suitability of CityBioNet for use in cities, as the urban soundscape is dominated by the sound of road traffic (Fairbrass \textit{et al.} 2017) which has been shown to bias several of the AIs tested here (Fuller \textit{et al.} 2015; Fairbrass \textit{et al.} 2017). The sound of rain has also been shown to bias several AIs (Depraetere \textit{et al.} 2012; Gasc \textit{et al.} 2015b; Fairbrass \textit{et al.} 2017) and the development of a method that is robust to this sound is a considerable contribution to the field.
of ecoacoustics. The urban biotic soundscape is dominated by the sounds emitted by birds (Fairbrass et al. 2017), and the good performance of bulbul, an algorithm for measuring exclusively bird sounds, on the CitySounds2017 test dataset, confirms this. Birds are used as indicator species in existing urban biodiversity monitoring schemes (Kohsaka et al. 2013) using data collected from traditional forms of biodiversity survey. The algorithms developed here could be used to support such existing schemes by making it easier to collect data on these indicator taxa.

CityNet is the only method currently available for measuring both biotic and anthropogenic acoustic activity using a single system in noisy audio data from urban environments. There is increasing evidence that anthropogenic noise affects wildlife in a variety of ways including altering communication behaviour (Gil & Brumm 2014) and habitat use (Deichmann et al. 2017). However, these investigations are limited in scale by the use of resource intensive methods of measuring biotic and anthropogenic sound in the environment or from audio data. Others rely on AIs (Pieretti & Farina 2013) which have been shown to be unreliable in acoustically disturbed environments (Fairbrass et al. 2017). CityNet could facilitate the investigation of the impacts of anthropogenic activities on wildlife populations at scales not currently possible with traditional acoustic analysis methods.

CityBioNet clearly outperformed all the AIs tested, but the difference in performance between CityAnthroNet and the competing algorithm for measuring anthropogenic acoustic activity (NDSIanthro) was much less marked. These results suggest that the measurement of biotic sound in noisy audio data from urban environments requires more sophisticated algorithms than the measurement of anthropogenic sound. Possibly anthropogenic sounds are more easily separable from other sounds in frequency space, a theory which is the basis of a number of AIs (Boelman et al. 2007; Kasten et al. 2012), facilitating the use of human defined algorithms such as NDSIanthro. Whereas, because biotic sounds occur in a frequency
space shared with anthropogenic and geophonic sounds (Fairbrass et al. 2017), algorithms
such as AIs which only use a small number of features to discriminate sounds are not
sufficient for use in cities. Therefore, ML algorithms which are able to utilise larger numbers
of features to discriminate sounds, such as the CNNs implemented in the CityNet system, are
better able to detect biotic sounds in recordings that also contain non-biotic sounds. A recent
unsupervised method developed by Lin, Fang and Tsao (2017) to separate biological sounds
from long recordings could be used as a pre-processing step to further improve CityNet’s
performance.

Low cost acoustic sensors and algorithms for the automatic measurement of biotic sound in
audio data is facilitating the assessment and monitoring of biodiversity at large temporal and
spatial scales (Sueur & Farina 2015), but to date this technology has only been deployed in
non-urban environments (e.g. Aide et al. 2013). In cities, the availability of mains power and
Wifi connections is supporting the development of the urban Internet of Things (IoT) using
sensors integrated into existing infrastructure to monitor environmental factors including air
pollution, noise levels, and energy use (Zanella et al. 2014). The CityNet system could be
integrated into an IoT sensing network to facilitate large-scale urban environmental
assessment. Large-scale deployment of algorithms such as CityNet requires low power usage
and fast running times. One way to help to achieve this aim would be to combine the two
networks (CityBioNet and CityAnthroNet) into one CNN which predicts both biotic and
anthropogenic acoustic activity simultaneously.

An expansion of CityNet to ultrasonic frequencies would increase the generality of the tool as
it could be used to monitor species in cities that emit sounds at frequencies higher than 12
kHz such as bats and some invertebrates. Bats are frequently used as ecological indicators
because they are sensitive to environmental changes (Walters et al. 2013). Acoustic methods
are commonly used to monitor bat populations using passive ultrasonic recorders meaning bat
researchers and conservationists are faced with the challenge of extracting meaningful information from large volumes of audio data. The development of automated methods for measuring bat calls in ultrasonic data has focused to date on the identification of bat species calls and many algorithms are proprietary (e.g., Szewczak 2010; Wildlife Acoustics 2017). The development of an open-source algorithm that produces community-level measures of bats would be a valuable addition to the toolbox of bat researchers and conservationists.

Retraining CityNet with labelled audio data from other cities would make it possible to use the system to monitor urban biotic and anthropogenic acoustic activity more widely. However, as London is a large and heterogeneous city, CityNet has been trained using a dataset containing sounds that characterise a wide range of urban environments. Our data collection was restricted to a single week at each study site, which limits our ability to assess the ability of CityNet system to detect environmental changes. Future work should focus on the collection of longitudinal acoustic data to assess the sensitivity of the algorithms to detect environmental changes. Our use of human labellers would have introduced subjectivity and bias into our dataset. The task of annotating large audio datasets from acoustically complex urban environments is highly resource intensive, a problem which has been recently tackled with citizen scientists to create the UrbanSounds and UrbanSound8k datasets using audio data from New York city, USA (Salamon, Jacoby & Bello 2014). These comprise short snippets of 10 different urban sounds such as jackhammers, engines idling and gunshots. These datasets do not fully represent the characteristics of urban soundscapes for three reasons. Firstly, they assume only one class of sound is present at each time, while in fact multiple sound types can be present at one time (consider a bird singing while an aeroplane flies overhead). Secondly, they only include anthropogenic sounds, while CityNet measures both anthropogenic and biotic sounds. Finally, each file in these datasets has a sound present, while urban soundscapes contain many periods of silence or geophonic sounds, two
important states which are not present in UrbanSounds and UrbanSounds8k. Due to these factors, these datasets are unsuitable for the purpose of this research project, although recent work has overcome a few of these shortcomings using synthesised soundscape data (Salamon et al. 2017). This highlights the need for an internationally coordinated effort to create a consistently labelled audio dataset from cities to support the development of automated urban environmental assessment systems with international application.

Conclusions

The CityNet system for measuring biotic and anthropogenic acoustic activity in noisy urban audio data outperformed the state-of-the-art algorithms for measuring biotic and anthropogenic sound in entire audio recordings. Integrated into an IoT network for recording and analysing audio data in cities it could facilitate urban environmental assessment at greater scales than has been possible to date using traditional methods of biodiversity assessment.

We make our system available open source in combination with two expertly annotated urban soundscape datasets to facilitate future research development in this field.

AUTHOR CONTRIBUTION STATEMENT

AF, MF, HT and KJ conceived ideas and designed methodology; AF collected the data; AF and MF analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGMENTS

We thank multiple site owners and managers for supporting the study by providing access to recording sites, and multiple acoustic annotators and a transport expert for help creating the CitySounds2017 dataset. We were financially supported by a BHP Billiton Sustainable Resources for Sustainable Cities Catalyst Grant and by the Engineering and Physical

Sciences Research Council (EPSRC) through a doctoral training grant (EP/G037698/1) to H.T., and EPSRC grant (EP/K015664/1) to K.E.J, G.B. and M.F.

DATA ACCESSIBILITY

All recordings and annotations in the CitySounds2017 dataset and all Python code underlying the CityNet algorithms are available on Figshare (https://figshare.com/s/adab62c0591afaeafedd).

REFERENCES


Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., De Fauw, J., Heilman, M., de Almeida, D.M., McFee, B., Weideman, H., Takács, G., de Rivaz, P., Crall, J., Sanders, G., Rasul, K.,


Springer, Netherlands.


Istanbul, Turkey.


Accessed: 24/08/2017


### Table 1. Average precision and recall results for CityNet and competing algorithms for each 1-second audio chunk in the CitySounds2017\textsubscript{test} dataset. Recall results are presented at 0.95 precision. Higher values are better for both metrics. The highest values in each section are shown in bold. ACI represents Acoustic Complexity Index, ADI Acoustic Diversity Index, BI Bioacoustic Index, and NDSI\textsubscript{biotic} and NDSI\textsubscript{anthro} biotic and anthropogenic Normalised Difference Soundscape Index, respectively.

<table>
<thead>
<tr>
<th>Acoustic Measures</th>
<th>Recall at 0.95 precision</th>
<th>Average precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biotic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CityBioNet</td>
<td>0.710</td>
<td>0.934</td>
</tr>
<tr>
<td>Bulbul</td>
<td>0.398</td>
<td>0.872</td>
</tr>
<tr>
<td>ACI</td>
<td>0.000</td>
<td>0.663</td>
</tr>
<tr>
<td>ADI</td>
<td>0.001</td>
<td>0.439</td>
</tr>
<tr>
<td>BI</td>
<td>0.002</td>
<td>0.516</td>
</tr>
<tr>
<td>NDSI\textsubscript{biotic}</td>
<td>0.000</td>
<td>0.503</td>
</tr>
<tr>
<td><strong>Anthropogenic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CityAnthroNet</td>
<td>0.858</td>
<td>0.977</td>
</tr>
<tr>
<td>NDSI\textsubscript{anthro}</td>
<td>0.815</td>
<td>0.975</td>
</tr>
</tbody>
</table>
Table 2. Impact of non-biotic sounds on the CityBioNet and bulbul predictions. Values represent differences in the proportion of 1-second audio chunks in the full CitySound2017 test dataset (40 minutes) and the subset datasets (size in time indicated in left-hand column) in which the presence/absence of biotic sound was correctly predicted by both algorithms, (chi-squared test statistic for difference in proportions of successes in each dataset, and Cramer’s V effect size measure). Effect sizes indicated as <0.1 (*), 0.1-0.3 (**) and >0.5 (***)..

<table>
<thead>
<tr>
<th>Sound Type</th>
<th>CityBioNet</th>
<th>Bulbul</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anthropogenic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Air traffic (9m 4s)</td>
<td>-2.11 (30.35, 0.05)*</td>
<td>-5.34 (162.73, 0.12)**</td>
</tr>
<tr>
<td>Mechanical (11s)</td>
<td>-28.60 (134.38, 0.77)**</td>
<td>0.02 (0.01, 0.01)*</td>
</tr>
<tr>
<td>Road traffic (29m 15s)</td>
<td>0.79 (10.15, 0.02)*</td>
<td>1.41 (27.67, 0.03)*</td>
</tr>
<tr>
<td>Siren (1m 21s)</td>
<td>2.28 (5.73, 0.06)*</td>
<td>3.70 (12.95, 0.09)*</td>
</tr>
<tr>
<td><strong>Geophonic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rain (2m 44s)</td>
<td>-0.77 (1.29, 0.02)*</td>
<td>-1.51 (4.17, 0.04)*</td>
</tr>
<tr>
<td>Wind (53s)</td>
<td>0.76 (0.47, 0.02)*</td>
<td>-6.93 (33.11, 0.17)**</td>
</tr>
</tbody>
</table>
Figure 1. The CityNet analysis pipeline for measuring biotic and anthropogenic acoustic activity. Raw audio (1), recorded in the field, is converted to a spectrogram representation.
(2). A sliding window is run across the time dimension, and a window of the spectrogram extracted at each step (3). This spectrogram window is pre-processed with four different normalisation strategies, and the results concatenated. This stack of spectrograms is passed through a CNN (5), which was trained on CitySounds2017\textsubscript{train}. The CNN gives, at each 1-second time step, a prediction of the presence/absence of biotic or anthropogenic acoustic activity (6). Finally, these per-time-step measures can be aggregated to give summaries over time or space (7).

Figure 2. Location of study sites and average daily acoustic patterns at two sites along an urbanisation gradient. Points in (A) represent locations used for the training dataset, CitySounds2017\textsubscript{train} (black) and testing dataset, CitySounds2017\textsubscript{test} (red). Here CityNet was run across the entire 7 days of recording at two sites of high (B) and low (C) urban intensity to predict the presence/absence of biotic and anthropogenic sound at each second of the week.
using a 0.5 probability threshold. The predicted number of seconds containing biotic and
anthropogenic sound for each half-hour period was averaged over the week to produce
average daily patterns of acoustic activity. Greater London boundary indicated with bold line.

Figure 3. Precision-recall curves for CityNet and competing algorithms predicting A) biotic and B) anthropogenic acoustic activity for each 1-second audio chunk in the CitySounds2017 test dataset. Dots indicate the precision and recall values at a threshold value of 0.5. ACI represents Acoustic Complexity Index, ADI Acoustic Diversity Index, BI Bioacoustic Index, and NDSI\textsubscript{bio} and NDSI\textsubscript{anthro} bio and anthropogenic Normalised Difference Soundscape Index, respectively.
Figure 4. Confusion matrices comparing the predicted acoustic activity of A) CityBioNet, B), bulbul, and C) CityAnthroNet for each 1-second audio chunk in the CitySounds2017 test dataset. Numbers in each cell report the number of 1-second audio clips in the CitySounds2017 test dataset predicted either correctly (True Positives and True Negatives) or incorrectly (False Positives and False Negatives) as containing biotic (A and B) or anthropogenic (C) sound. To create the confusion matrices, the probabilistic predictions from the classifiers are converted to binary classifications using a threshold that gives a precision of 0.95.
SUPPORTING INFORMATION

Section S1: Supplementary Methods

Normalisation Methods

The four normalisation methods used are as follows:

1. The entire spectrogram $S$ was subtracted from each row in $W_S$. This helped to act as a noise-reducing normalisation strategy.
2. Each row of $W_S$ was whitened to have zero mean and unit variance.
3. Each value in $W_S$ was whitened to have zero mean and unit variance.
4. Each value in $W_S$ was divided by the maximum value in $W_S$.

Prediction Process

Both CityBioNet and CityAnthroNet have a convolutional layer with 32 filters, followed by a max pooling layer, then another 32-filter convolutional layer and finally two dense layers (with 128 units) before a binary class output - see Figure 1 for an overview of the network architecture. For nonlinearities very leaky rectifiers were used (Maas, Hannun & Ng 2013), and Dropout (Srivastava et al. 2014) was used to help to regularise the network and batch normalisation (Ioffe & Szegedy 2015) to increase the speed of convergence during training. The network was trained for 30 epochs using the Adam (Kingma & Ba 2015) update scheme with a learning rate of 0.0005. An ensemble of five such networks was trained using the same architecture and training data, but with different random initialisations. The final predictions are made by averaging together the predictions of each member in the ensemble.

Table S1. Details of acoustic recording sites across Greater London, UK. Sites separated into two groups illustrating whether recordings from sites were included in the CitySounds2017\textsubscript{train} or CitySounds2017\textsubscript{test} datasets. Urban intensity categories defined based on the predominant land cover surrounding sites within a 500m radius: (i) high (contiguous multi-storey buildings); (ii) medium (detached and semi-detached housing); and (iii) low (fields and/or woodland). DD denotes decimal degrees. In terms of site type, C denotes church or churchyard, CG denoted community garden, GR denotes green roof, GW denotes green wall, and NR denotes nature reserve.

<table>
<thead>
<tr>
<th>Site Code</th>
<th>Site Type</th>
<th>Survey Start Date</th>
<th>Survey End Date</th>
<th>Latitude (DD)</th>
<th>Longitude (DD)</th>
<th>Urban Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM14 3YB</td>
<td>C</td>
<td>11/06/2013</td>
<td>19/06/2013</td>
<td>51.55121</td>
<td>0.266853</td>
<td>Low</td>
</tr>
<tr>
<td>W8 4LA</td>
<td>C</td>
<td>21/06/2013</td>
<td>28/06/2013</td>
<td>51.50223</td>
<td>-0.19147</td>
<td>High</td>
</tr>
<tr>
<td>SW15 4LA</td>
<td>C</td>
<td>02/07/2013</td>
<td>07/07/2013</td>
<td>51.44914</td>
<td>-0.23697</td>
<td>Medium</td>
</tr>
<tr>
<td>NW1</td>
<td>C</td>
<td>24/06/2013</td>
<td>01/07/2013</td>
<td>51.5105</td>
<td>-0.20574</td>
<td>High</td>
</tr>
<tr>
<td>SW11 2PN</td>
<td>C</td>
<td>16/08/2013</td>
<td>23/08/2013</td>
<td>51.47057</td>
<td>-0.16973</td>
<td>High</td>
</tr>
<tr>
<td>E4 7EN</td>
<td>C</td>
<td>06/10/2013</td>
<td>13/10/2013</td>
<td>51.63101</td>
<td>0.001266</td>
<td>High</td>
</tr>
<tr>
<td>SE1 2RT 7</td>
<td>GR</td>
<td>19/05/2014</td>
<td>27/05/2014</td>
<td>51.30.16N</td>
<td>0.4.53W</td>
<td>High</td>
</tr>
<tr>
<td>SE1 2RT 10</td>
<td>GR</td>
<td>19/05/2014</td>
<td>27/05/2014</td>
<td>51.30.16N</td>
<td>0.4.50W</td>
<td>High</td>
</tr>
<tr>
<td>SW1W 0QP</td>
<td>GW</td>
<td>30/05/2014</td>
<td>06/06/2014</td>
<td>51.49627</td>
<td>-0.14489</td>
<td>High</td>
</tr>
<tr>
<td>SW1E 6BN</td>
<td>GR</td>
<td>30/05/2014</td>
<td>06/06/2014</td>
<td>51.4981</td>
<td>-0.14138</td>
<td>High</td>
</tr>
<tr>
<td>SE11 6DN</td>
<td>GR</td>
<td>11/06/2014</td>
<td>20/06/2014</td>
<td>51.49313</td>
<td>-0.11199</td>
<td>High</td>
</tr>
<tr>
<td>SE4 1SA</td>
<td>GR</td>
<td>20/06/2014</td>
<td>30/06/2014</td>
<td>51.45817</td>
<td>-0.02751</td>
<td>Medium</td>
</tr>
<tr>
<td>WC2N 6RH</td>
<td>GR</td>
<td>01/07/2014</td>
<td>10/07/2014</td>
<td>51.50706</td>
<td>-0.12388</td>
<td>High</td>
</tr>
<tr>
<td>Location</td>
<td>Type</td>
<td>Start Date</td>
<td>End Date</td>
<td>Latitude</td>
<td>Longitude</td>
<td>Level</td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>-----------</td>
<td>-------</td>
</tr>
<tr>
<td>CR0 1SG</td>
<td>C</td>
<td>02/07/2014</td>
<td>09/07/2014</td>
<td>51.3722</td>
<td>-0.10604</td>
<td>High</td>
</tr>
<tr>
<td>CR0</td>
<td>C</td>
<td>02/07/2014</td>
<td>09/07/2014</td>
<td>51.33934</td>
<td>-0.01266</td>
<td>Medium</td>
</tr>
<tr>
<td>RM2 5EL</td>
<td>C</td>
<td>10/07/2014</td>
<td>17/07/2014</td>
<td>51.58773</td>
<td>0.201817</td>
<td>Medium</td>
</tr>
<tr>
<td>RM4 1LD</td>
<td>C</td>
<td>10/07/2014</td>
<td>17/07/2014</td>
<td>51.62349</td>
<td>0.223904</td>
<td>Low</td>
</tr>
<tr>
<td>SE22 0SD</td>
<td>GR</td>
<td>28/07/2014</td>
<td>04/08/2014</td>
<td>51.45332</td>
<td>-0.05583</td>
<td>Medium</td>
</tr>
<tr>
<td>TW7 6BE</td>
<td>C</td>
<td>30/07/2014</td>
<td>06/08/2014</td>
<td>51.4719</td>
<td>-0.31981</td>
<td>Medium</td>
</tr>
<tr>
<td>W4 2PH</td>
<td>C</td>
<td>30/07/2014</td>
<td>06/08/2014</td>
<td>51.48308</td>
<td>-0.25326</td>
<td>Medium</td>
</tr>
<tr>
<td>SE6</td>
<td>C</td>
<td>19/08/2014</td>
<td>26/08/2014</td>
<td>51.42804</td>
<td>-0.01095</td>
<td>Medium</td>
</tr>
<tr>
<td>SE8 4EA</td>
<td>C</td>
<td>19/08/2014</td>
<td>27/08/2014</td>
<td>51.46841</td>
<td>-0.02344</td>
<td>Medium</td>
</tr>
<tr>
<td>IG11 0FJ</td>
<td>GR</td>
<td>21/08/2014</td>
<td>01/09/2014</td>
<td>51.52069</td>
<td>0.109187</td>
<td>Medium</td>
</tr>
<tr>
<td>W5 5EQ</td>
<td>GR</td>
<td>28/08/2014</td>
<td>05/09/2014</td>
<td>51.50975</td>
<td>-0.30812</td>
<td>Medium</td>
</tr>
<tr>
<td>E14 0EY</td>
<td>C</td>
<td>02/09/2014</td>
<td>10/09/2014</td>
<td>51.51072</td>
<td>-0.01192</td>
<td>High</td>
</tr>
<tr>
<td>E1 0NR</td>
<td>C</td>
<td>03/09/2014</td>
<td>11/09/2014</td>
<td>51.51676</td>
<td>-0.04122</td>
<td>Medium</td>
</tr>
<tr>
<td>SE10 9EY</td>
<td>GR</td>
<td>05/09/2014</td>
<td>12/09/2014</td>
<td>51.4849</td>
<td>0.006003</td>
<td>Medium</td>
</tr>
<tr>
<td>N2 9BX</td>
<td>GR</td>
<td>15/09/2014</td>
<td>22/09/2014</td>
<td>51.59274</td>
<td>-0.16569</td>
<td>Medium</td>
</tr>
<tr>
<td>SW6 6DU</td>
<td>GR</td>
<td>16/09/2014</td>
<td>23/09/2014</td>
<td>51.47369</td>
<td>-0.21695</td>
<td>Medium</td>
</tr>
<tr>
<td>SE6 4PL</td>
<td>CG</td>
<td>24/05/2015</td>
<td>01/06/2015</td>
<td>51.43821</td>
<td>-0.02711</td>
<td>Medium</td>
</tr>
<tr>
<td>W1T 4BQ</td>
<td>GR</td>
<td>22/06/2015</td>
<td>30/06/2015</td>
<td>51.52143</td>
<td>-0.13836</td>
<td>High</td>
</tr>
<tr>
<td>N4 1ES</td>
<td>NR</td>
<td>23/06/2015</td>
<td>02/07/2015</td>
<td>51.57656</td>
<td>-0.1017</td>
<td>Medium</td>
</tr>
<tr>
<td>TN14 7QB</td>
<td>NR</td>
<td>25/06/2015</td>
<td>03/07/2015</td>
<td>51.31364</td>
<td>0.067323</td>
<td>Low</td>
</tr>
<tr>
<td>NW3 3RY</td>
<td>NR</td>
<td>14/07/2015</td>
<td>22/07/2015</td>
<td>51.54357</td>
<td>-0.16054</td>
<td>High</td>
</tr>
<tr>
<td>N8 8JD</td>
<td>CG</td>
<td>11/07/2015</td>
<td>19/07/2015</td>
<td>51.58333</td>
<td>-0.13292</td>
<td>Medium</td>
</tr>
<tr>
<td>KT18 6AP</td>
<td>NR</td>
<td>27/07/2015</td>
<td>05/08/2015</td>
<td>51.29036</td>
<td>-0.26158</td>
<td>Low</td>
</tr>
<tr>
<td>NW2 3SH</td>
<td>NR</td>
<td>11/08/2015</td>
<td>18/08/2015</td>
<td>51.55287</td>
<td>-0.20628</td>
<td>Medium</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Code</th>
<th>City</th>
<th>Date</th>
<th>Start Date</th>
<th>Start Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>N17</td>
<td>CG</td>
<td>17/08/2015</td>
<td>27/08/2015</td>
<td>51.59105</td>
</tr>
<tr>
<td>RM4 1PL</td>
<td>C</td>
<td>27/08/2015</td>
<td>04/09/2015</td>
<td>51.61588</td>
</tr>
<tr>
<td>SE23 2NZ</td>
<td>NR</td>
<td>16/09/2015</td>
<td>23/09/2015</td>
<td>51.43224</td>
</tr>
<tr>
<td>NW3 2BZ</td>
<td>NR</td>
<td>17/09/2015</td>
<td>25/09/2015</td>
<td>51.55181</td>
</tr>
<tr>
<td>NW1 0TA</td>
<td>NR</td>
<td>15/10/2015</td>
<td>22/10/2015</td>
<td>51.54073</td>
</tr>
<tr>
<td>SE15 4EE</td>
<td>CG</td>
<td>13/10/2015</td>
<td>20/10/2015</td>
<td>51.46301</td>
</tr>
<tr>
<td>RM15 4HX</td>
<td>NR</td>
<td>20/10/2015</td>
<td>28/10/2015</td>
<td>51.51749</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CitySounds2017</th>
<th>Test</th>
<th>Date</th>
<th>Start Date</th>
<th>Start Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>W11 2NN</td>
<td>C</td>
<td>08/07/2013</td>
<td>16/07/2013</td>
<td>51.53452</td>
</tr>
<tr>
<td>WC2H 8LG</td>
<td>C</td>
<td>08/07/2013</td>
<td>14/07/2013</td>
<td>51.51521</td>
</tr>
<tr>
<td>HA8 6RB</td>
<td>C</td>
<td>23/07/2013</td>
<td>30/07/2013</td>
<td>51.60862</td>
</tr>
<tr>
<td>HA5 3AA</td>
<td>C</td>
<td>23/07/2013</td>
<td>30/07/2013</td>
<td>51.59478</td>
</tr>
<tr>
<td>SE23</td>
<td>C</td>
<td>06/09/2013</td>
<td>13/09/2013</td>
<td>51.45047</td>
</tr>
<tr>
<td>SE3</td>
<td>C</td>
<td>06/09/2013</td>
<td>13/09/2013</td>
<td>51.46261</td>
</tr>
<tr>
<td>CR8</td>
<td>C</td>
<td>15/09/2013</td>
<td>22/09/2013</td>
<td>51.3305</td>
</tr>
<tr>
<td>CR0 5EF</td>
<td>C</td>
<td>15/09/2013</td>
<td>22/09/2013</td>
<td>51.37199</td>
</tr>
<tr>
<td>E10 5JP</td>
<td>C</td>
<td>06/10/2013</td>
<td>13/10/2013</td>
<td>51.56386</td>
</tr>
<tr>
<td>SW15 4JY</td>
<td>GR</td>
<td>27/08/2014</td>
<td>03/09/2014</td>
<td>51.45012</td>
</tr>
<tr>
<td>IG6 2XL</td>
<td>CG</td>
<td>08/05/2015</td>
<td>15/05/2015</td>
<td>51.60046</td>
</tr>
<tr>
<td>E2 9RR</td>
<td>NR</td>
<td>25/05/2015</td>
<td>02/06/2015</td>
<td>51.5295</td>
</tr>
<tr>
<td>TW7 6ER</td>
<td>C</td>
<td>23/06/2015</td>
<td>30/06/2015</td>
<td>51.46711</td>
</tr>
<tr>
<td>BR2 0EG</td>
<td>C</td>
<td>17/07/2015</td>
<td>26/07/2015</td>
<td>51.4047</td>
</tr>
<tr>
<td>BR2 8LB</td>
<td>C</td>
<td>31/07/2015</td>
<td>07/08/2015</td>
<td>51.38029</td>
</tr>
<tr>
<td>BR6 7US</td>
<td>C</td>
<td>31/07/2015</td>
<td>07/08/2015</td>
<td>51.33605</td>
</tr>
</tbody>
</table>

*The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.*
<table>
<thead>
<tr>
<th>Site</th>
<th>Code</th>
<th>Date</th>
<th>Date</th>
<th>Latitude</th>
<th>Temperature</th>
<th>Noise Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR4</td>
<td>C</td>
<td>18/08/2015</td>
<td>25/08/2015</td>
<td>51.38261</td>
<td>-0.00868</td>
<td>Medium</td>
</tr>
<tr>
<td>DA5</td>
<td>NR</td>
<td>24/08/2015</td>
<td>01/09/2015</td>
<td>51.42268</td>
<td>0.156502</td>
<td>Medium</td>
</tr>
<tr>
<td>CM16 7NP</td>
<td>NR</td>
<td>08/09/2015</td>
<td>15/09/2015</td>
<td>51.65396</td>
<td>0.101227</td>
<td>Low</td>
</tr>
<tr>
<td>Biotic</td>
<td>Animal</td>
<td>Bird</td>
<td>Invertebrate</td>
<td>Vegetation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>-----------------</td>
<td>--------------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><img src="image1" alt="Bird" /></td>
<td><img src="image2" alt="Bird" /></td>
<td><img src="image3" alt="Invertebrate" /></td>
<td><img src="image4" alt="Vegetation" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bird wing beats</td>
<td><img src="image5" alt="Bird Wing Beats" /></td>
<td><img src="image6" alt="Bird Wing Beats" /></td>
<td><img src="image7" alt="Invertebrate Wing Beats" /></td>
<td><img src="image8" alt="Vegetation Wing Beats" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anthropogenic</th>
<th>Air traffic</th>
<th>Braking vehicle</th>
<th>Electrical buzz</th>
<th>Human speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image9" alt="Air Traffic" /></td>
<td><img src="image10" alt="Braking Vehicle" /></td>
<td><img src="image11" alt="Electrical Buzz" /></td>
<td><img src="image12" alt="Human Speech" /></td>
</tr>
<tr>
<td>Mechanical</td>
<td><img src="image13" alt="Metal Crash" /></td>
<td><img src="image14" alt="Road Traffic" /></td>
<td><img src="image15" alt="Siren" /></td>
<td><img src="image16" alt="Vehicle Alarm" /></td>
</tr>
<tr>
<td>Vehicle Alarm</td>
<td><img src="image17" alt="Vehicle Horn" /></td>
<td><img src="image18" alt="Vehicle Horn" /></td>
<td><img src="image19" alt="Vehicle Horn" /></td>
<td><img src="image20" alt="Vehicle Horn" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geophonic</th>
<th>Rain</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image21" alt="Rain" /></td>
<td><img src="image22" alt="Wind" /></td>
</tr>
</tbody>
</table>

**Figure S1.** Examples of all sound types present in CitySounds2017. ‘Animal’ denotes biotic sounds that could not be taxonomically identified. Unidentified sounds not shown due to wide range of sound types within this group. Data is represented in spectrograms (FFT non-overlapping Hamming window size=1024) where blue to yellow corresponds to sound amplitude (dB). Frequency (kHz) and time (s) are represented on the y- and x-axes, respectively. Spectrograms represent biotic (sounds generated by non-human biotic
organisms), anthropogenic (sounds associated with human activities including human speech) and geophonic sounds.

REFERENCES


