Buenos Aires – 5 to 9 September, 2016
Acoustics for the 21st Century…

PROCEEDINGS of the 22nd International Congress on Acoustics

Smart City Sound Monitoring: Paper ICA2016-793

Data mining on urban sound sensor networks

Dick Botteldooren[(a)], Talis Vertriest[(a)], Michiel Boes[(a)], Bert De Coensel[(a)], Pierre Aumond[(b)], Arnaud Can[(c)], Carlos Ribeiro[(d)], Catherine Lavandier[(b)]

[(a)] Ghent University, Belgium, dick.botteldooren@ugent.be
[(b)] Université de Cergy-Pontoise, France, catherine.lavandier@u-cergy.fr
[(c)] IFFSTAR, France, arnaud.can@ifsttar.fr
[(d)] Bruitparif, France, Carlos.Ribeiro@bruitparif.fr

Abstract

Urban sound sensor networks deliver megabytes of data on a daily basis so the question on how to extract useful knowledge from this overwhelming dataset is eminent. This paper presents and compares two extremely different approaches. The first approach uses as much as possible expert knowledge on how people perceive the sonic environment, the second approach simply considers the spectra obtained every time step as meaningless numbers yet tries to structure them in a meaningful way. The approach based on expert knowledge starts by extracting features that a human listener might use to detect salient sounds and to recognize these sounds. These features are then fed to a recurrent neural network that learns in an unsupervised way to structure and group these features based on co-occurrence and typical sequences. The network is constructed to mimic human auditory processing and includes inhibition and adaptation processes. The outcome of this network is the activation of a set of several hundred neurons. The second approach collects a sequence of one minute of sound spectra (1/8 second time step) and summarizes it using Gaussian mixture models in the frequency-amplitude space. Mean and standard deviation of the set of Gaussians are used for further analysis. In both cases, the outcome is clustered to analyze similarities over space and time as well as to detect outliers. Both approaches are applied on a dataset obtained from 25 measurement nodes during approximately one and a half year in Paris, France. Although the approach based on human listening models is expected to be much more precise when it comes to analyzing and clustering soundscapes, it is also much slower than the blind data analysis.

Keywords: sensor networks, smart cities, urban sound

Data mining on urban sound sensor networks.

1 Introduction

The urban sound environment contains a lot of information about a neighbourhood or even a whole city [1]. It reflects its liveliness, its identity or reveals its tranquil restorative ambiance. Sound can delineate or form borders between places in the city fabric [2]. The urban sound environment also allows to identify events that need intervention or action [3].

New technologies including reliable consumer microphones [4] and affordable connectivity allow to deploy dense networks of sound observatories [5][6] often as part of smart city initiatives. In contrast to other sensor that often give low bandwidth information, the sound sensors result in huge amounts of potential useful data. Even if limited to the auditory frequency range sampling frequencies of over 40 kHz are useful for collecting the data. If soundscape analysis or categorisation is at stake, data compression taking into account human hearing capabilities is possible. The popular MFCC (Mel Frequency Cepstral Coefficients) have been used for this purpose outside their original scope of speech recognition [7]. Still the data continuously collected by the sound sensor network can be labelled as big data.

In this paper, some possibilities for extracting useful information and knowledge from these big data are explored. Both the search for categorising the usual and detecting the unexpected are considered. Categorisation of the usual is grounded in urban soundscape and our understanding of how soundscape emerges from perception of the sonic environment [8]. Outlier detection requires a fast and general algorithm as it are not only the sounds that are noticed by a human listener that might indicate a need for intervention.

The methodologies presented in Section 3 are tested on a sound sensor network deployed in Paris France, that will be discussed in Section 2. Results will be discussed in Section 4.

2 The Paris sensor network and dataset

During 2014-2015 a sound sensor network was deployed in the XIIIth district of Paris, France within the context of the GRAFIC project. This sensor network consisted of 24 sensor nodes constructed at Ghent University that were placed on the façade of private buildings by Bruitparif.

Based on previous analysis, it was decided to collect 1/3 octave band spectra 8 times per second, as this spectro-temporal resolution allows to identify short sound events such as bird cries or voices.

This dataset could qualify as big data as $24x60x6x8x30 = 2\ 10^7$ numbers are collected for each node every day. The total dataset, even when stored in an efficient format, takes 540 Gbytes of disk space. Thus, getting useful information and knowledge out of such a dataset is a clear challenge.

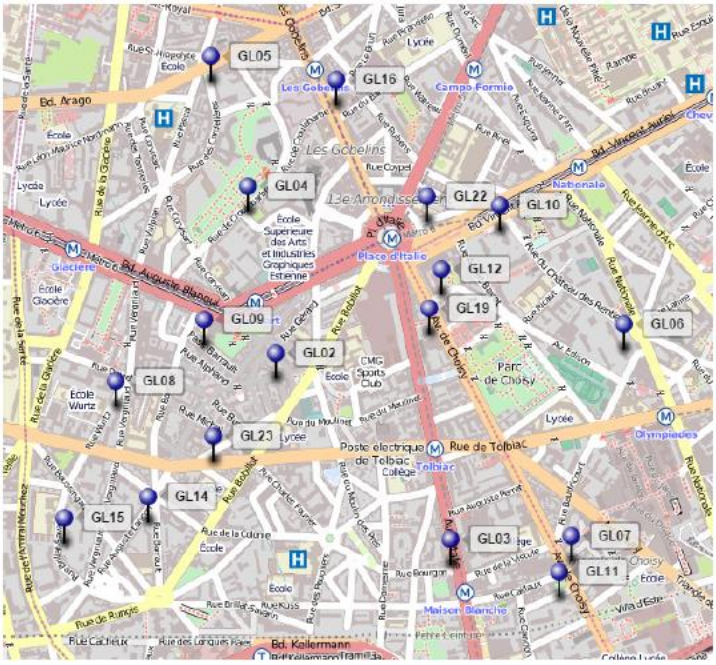

Figure 1: Location of the sensor nodes in Paris, France

3 Analysis methods

3.1 Human mimicking soundscape analysis

The first approach is strongly inspired by the soundscape concept. Soundscape defined as the sonic environment as perceived and understood by people or society within its context [9]. It has been shown that this perception and in particular also the appraisal of the sonic environment strongly depends on the types of sounds that the users of the space notice [10]. In general, persons walking in the street will not focus their attention on sound in particular. Listening in search of particular sounds is not the purpose of being at this location [11]. Thus a model evaluating soundscape should account for the attention mechanisms underlying noticing of a sound: saliency, attention focussing and gating, inhibition of return, etc. In **Error! Reference source not found.** we proposed such a model. Since then, the model has been refined to implicitly include most of the above mechanisms [12]. The recurrent artificial neural network (R-ANN) that has been used in the current work starts from a feature vector that includes 4 intensity, 6 spectral contrast and 4 temporal contrast features all based on the 125 ms temporal resolution and 1/3 octave band spectral resolution raw data. As these features correspond to the features used in [] for deriving saliency of sounds, saliency driven attention mechanisms are automatically included. Thus, in the neural network, high saliency sound input will cause strong excitations in the input layer, and consequentially in the whole network, thus effectively introducing a bias of the network towards turning attention to salient sounds. These features form the 768 neurons in the input layer of the R-ANN. The weights connecting the three layers of the ANN and the recurrent paths are trained by unsupervised training on co-occurrence of features. The recurrent paths cause additional neural excitation for neurons corresponding to

the sounds that are currently being attributed attention to by the network, thus effectively implementing top-down attention. Competitive selection is implemented by a biologically inspired intra-layer excitation-inhibition mechanism in order to make a selection amongst the neurons within each layer. Finally, the mechanism of inhibition-of-return is also included, as a neural excitation reducing mechanism as a consequence of continuous stimulation of the neuron, mimicking the gradual depletion of neurotransmitters in real neurons.

After training on several months of measured data, the 400 output nodes of the R-ANN now converge to representing sounds that occur frequently and are salient. Where sounds are defined as combinations of spectro-temporal features. Adding meaning – or in other words labelling – requires a supervised learning step that is not required for the application envisaged. Indeed, grouping and clustering of soundscapes or even detecting uncommon or outlying sounds does not require the system to know which meaning a human would give to these sounds. For interpretation one may identify the response to sounds that are expected to significantly influence the soundscape.

At any point in time (125 ms resolution) the R-ANN yields an excitation pattern of the 400 output nodes. These are now grouped over observation periods: day, evening, and night, giving a collection of ‘sounds’ that will be paid attention to over this time interval. Finally these sets of sounds can now clustered into soundscapes. Note that we use the term soundscape here to refer to a collection of sounds that will be noticed by an average observer at this location thus ignoring inter-individual differences and context dependence in the soundscape perception and understanding. Clustering performed using Ward's minimum variance clustering.

3.2 Big data analysis approach

The second approach ignores any detailed consideration on human perception and treats the data purely as numbers. That is, the interpretation of spectral shape and temporal fluctuation over one minute epochs are still considered. The basic datasets fed to the machine learning algorithm look like the one represented in Figure 2. A Gaussian mixture model (GMM) is now used to summarize these one-minute grouped spectra in a small number of features: n Gaussians with their central frequency and amplitude and their 3 parameters describing width and orientation. In human language the Gaussian components could be interpreted as for example: a high frequency sloping spectrum, a low frequency broad peak, or a strongly varying high frequency component.

In order to detect outliers or to cluster between locations and time, the 5n features extracted per minute over a whole year and a number of different measurement locations can now be collected. This huge amount of data is again clustered using a GMM model. The choice of GMM for this clustering is inspired by watching the distribution of features in the 5 dimensional space, which shows a number of distinct clear groups of data. The GMM combined with the AIC criterion will assign a number of clusters, N_c. Each cluster represents a prototypical one-minute feature vector that is often encountered in coding one-minute grouped spectra using GMM. During analysis, a minute where all Gaussian components strongly belong to one of the clusters can be called typical, a minute where at least one of the Gaussian components does not belong to any of these N_c clusters is an outlier or strange sound event.

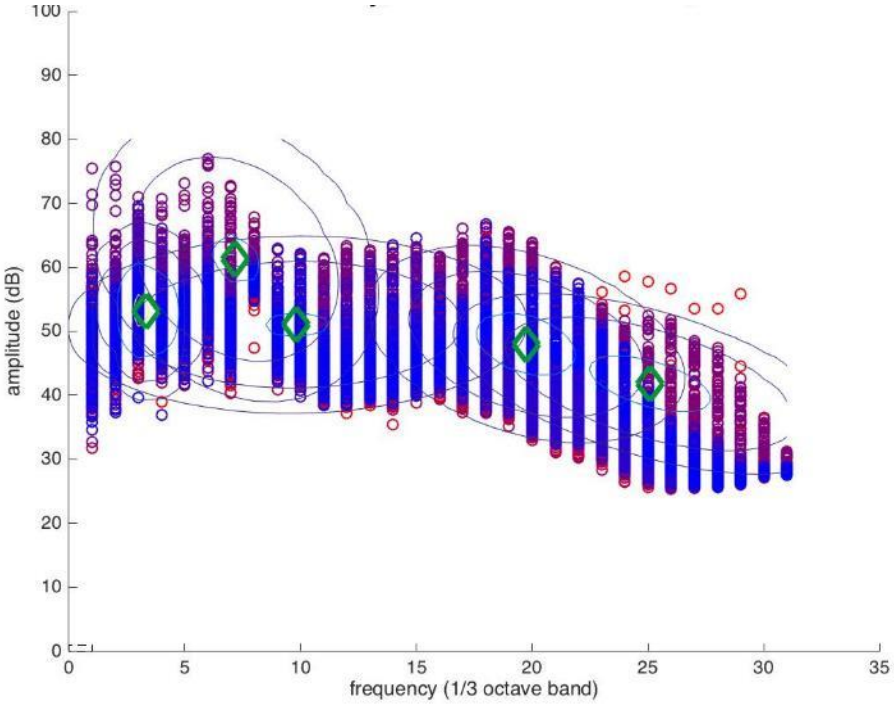

Figure 2: Example of one-minute grouped spectra together with Gaussian approximations

4 Results

4.1 Paris soundscapes

Using the human mimicking soundscape analysis the soundscapes at different locations, different times of the day, and different seasons was categorized. As this analysis is very cpu-time consuming only short time intervals could be analysed in a reasonable time frame. In particular for the winter period Saturday Feb 14th until Friday Feb 20th 2015 was used, for spring Thursday May 7th until Wednesday May 13th 2015. Locations 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 19, 22 and 23 were used (Figure 1).

Figure 3 shows the dendrogram resulting from Ward's minimum variance clustering on the output of the R-ANN on these data. A distance of 35 was used to group the soundscapes.

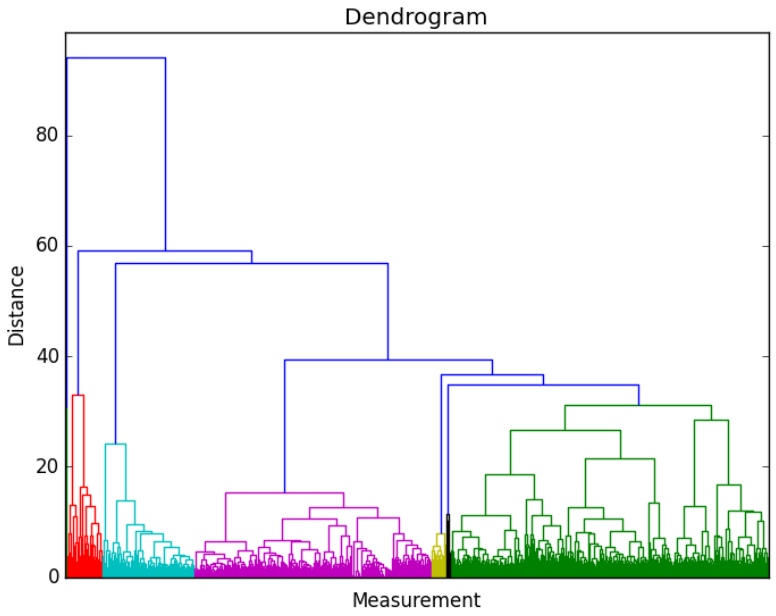

Figure 3: Dendrogram resulting from the clustering algorithm with main classes of soundscapes shown in colour

These clusters are now further analysed on the context in which they appear. Figure 4 shows where these soundscapes can be found during the evening, the day and the night. In addition we analysed the clusters also with respect to season, day of the week, and listened to specific fragments (not shown in this paper). From this analysis we conclude that the ‘red’ soundscape is strongly influenced by human voices; the ‘yellow’ soundscape contains a lot of restaurant sounds such as forks and knifes hitting a plate, glasses; the ‘cyan’ soundscape is dominated by a rather continuous traffic sound; the ‘green’ and ‘purple’ soundscapes contain a mixture of low intensity traffic sound and sounds of people but the ‘purple’ variant seems more related to the night.

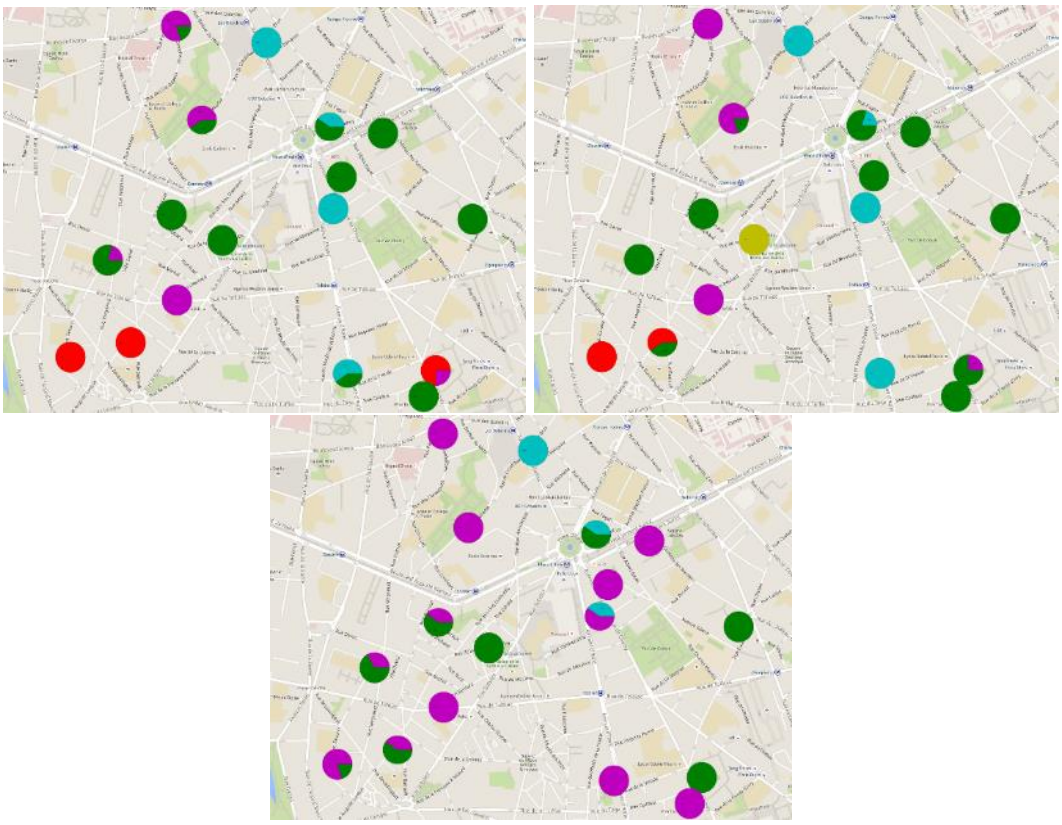

Figure 4: Occurrence analysis of the soundscape clusters identified by colours; the circles fractions refer to the percentage of the analysed time interval that this location matches a particular soundscape cluster; upper left=day, upper right=evening, lower=night.

4.2 Outliers and anomalies

An urban sound sensor network can be used to identify the typical (in sound levels or soundscape) but in terms of permanent deployment of such a sensor network, the possibility for detecting outliers or rare situations may be more appealing. The R-ANN for sound identification may be used for this purpose yet it has a few disadvantages. Firstly, as the method is trained on detecting and identifying sounds that often occur, it is by definition not going to respond appropriately when a new sound occurs. Secondly, the method is rather slow. Finally, it focusses only on sounds that people are likely to notice and not all sounds that might indicate a situation that requires intervention. Thus we mainly rely on the blind big data approach for this purpose.

As an example, outliers detected at the location shown in Figure 5 are explored.

Figure 5: Google streetview image of the location for which outlier sound minutes are detected as an example; the microphone is the small black dot in the open window in the middle of the picture on the first floor.

The GMM clustering the Gaussians representing one minute data between October 2014 and November 2015 results in N_c=42 in this case. If one of the Gaussians needed to represent the one-minute data does not fall in any of these N_c clusters, the minute is marked as containing an unexpected sound event. The number of outliers thus detected depends on the threshold set on the membership to the clusters. In the example, 1257 minutes were detected. Some examples are shown in Figure 6. A couple of human experts were asked to review the spectra and indicate whether there was something in them that could make them worthy of attention in their opinion. About 70% of the detected minutes were confirmed to be unexpected spectra for an urban environment by the human evaluators (true positives).

5 Conclusions

Data mining on urban sound sensor networks allows to extract useful knowledge not only on the sound environment and how it is perceived by users of the urban space, but also on the function of the city and on situations that might require intervention. During the GRAFIC project such a sound sensor network was deployed in the city of Paris, France. Two innovative techniques that have been used to analyse the big datasets are presented in this paper. Firstly, a human mimicking model for identifying noticed sounds is used to categorize soundscapes. This model relies on a recursive neural network that is trained in an unsupervised way to identify sounds that often occur near one of the sensors. Distinct soundscapes for different areas but also for different times of the day and seasons are identified based on this model.

Outlier detection on the other hand seems to benefit from a faster more generic method. A key factor in the process is to summarise spectrograms in a small number of features. Here Gaussian Mixture Models are used for this purpose. Each Gaussion is defined in a 5

dimensional space. Five Gaussians are usually sufficient to describe the combined spectrogram over one minute. Long term statistics on the GMM components allow to identify whether a minute does not fit the usual and thus qualifies as an unexpected sound that needs further investigating. This rather simple and relatively fast method is quite sufficient as a large majority of detected outliers would qualify as such for a human expert.

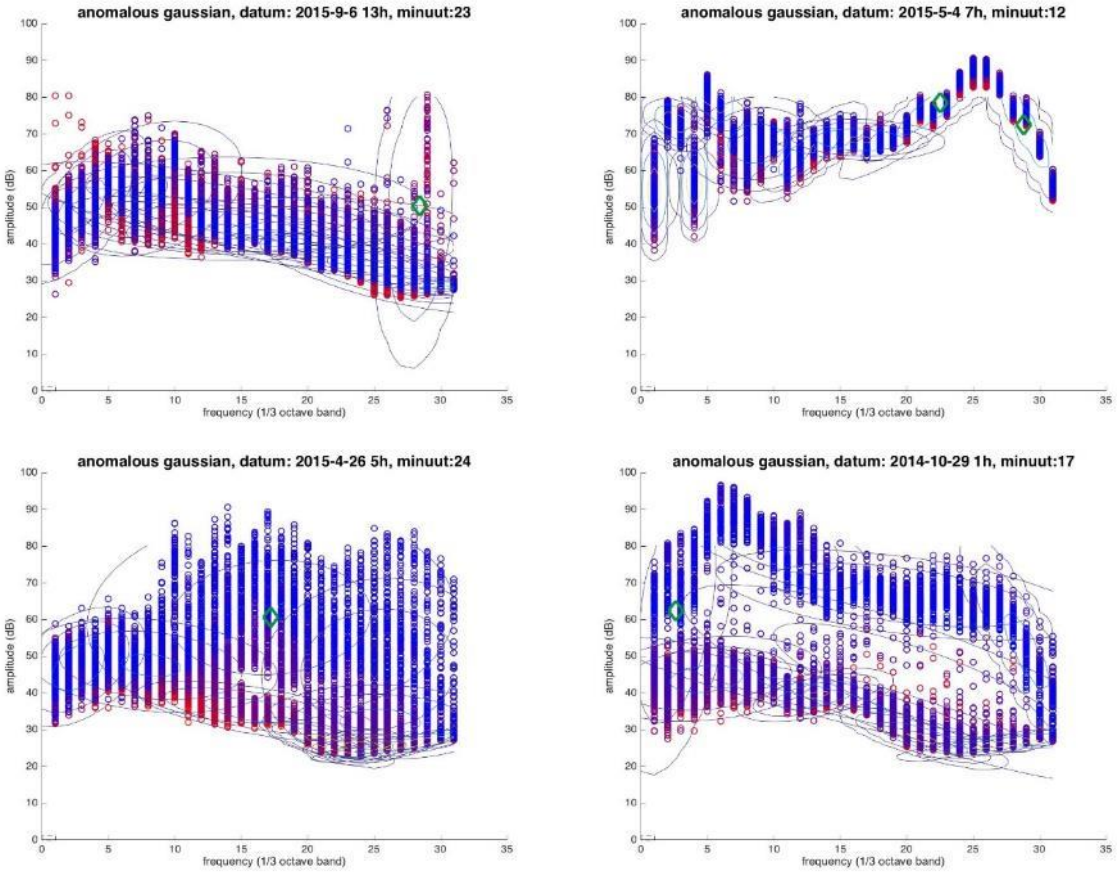

Figure 6: Four typical example spectra out of the 265 outlying sound minutes detected during the one year period at the location shown in Figure 4; x-axis=1/3 octave band, y-axis=unweighted level

Acknowledgments

Part of this work has been carried out in the framework of the GRAFIC project, supported by the French Environment and Energy Management Agency (ADEME) under contract No. 1317C0028.

M. Boes acknowledges the fund for scientific research (FWO) for a grant.

References

[1] Lankford EM. Urban Soundscapes as Indicators of Urban Health. Environment, Space, Place. 2009 Oct 1;1(2):27-50.

[2] Atkinson R. Ecology of sound: the sonic order of urban space. Urban studies. 2007 Sep 1;44(10):1905-17.

[3] Andersson M, Ntalampiras S, Ganchev T, Rydell J, Ahlberg J, Fakotakis N. Fusion of acoustic and optical sensor data for automatic fight detection in urban environments. InInformation Fusion (FUSION), 2010 13th Conference on 2010 Jul 26 (pp. 1-8). IEEE.

[4] Van Renterghem T, Thomas P, Dominguez F, Dauwe S, Touhafi A, Dhoedt B, Botteldooren D. On the ability of consumer electronics microphones for environmental noise monitoring. Journal of Environmental Monitoring. 2011;13(3):544-52.

[5] Botteldooren D, Oldoni D, Samuel D, Dekoninck L, Thomas P, Wei W, Boes M, De Coensel B, De Baets B, Dhoedt B. The internet of sound observatories. InProceedings of Meetings on Acoustics 2013 Jun 2 (Vol. 19, No. 1, p. 040140). Acoustical Society of America.

[6] Park TH, Turner J, Musick M, Lee JH, Jacoby C, Mydlarz C, Salamon J. Sensing Urban Soundscapes. InEDBT/ICDT Workshops 2014 (pp. 375-382).

[7] R. Radhakrishnan, A. Divakaran, P. Smaragdis. (2005). Audio Analysis for Surveillance Applications. Cambridge: Mitsubishi Electric Research Labs.

[8] Botteldooren D, Andringa T, Aspuru I, Brown AL, Dubois D, Guastavino C, Kang J, Lavandier C, Nilsson M, Preis A, Schulte-Fortkamp11 B. From Sonic Environment to Soundscape. Soundscape and the Built Environment. 2015 Dec 2:17.

[9] International Organization for Standardization. *Acoustics – Soundscape – Part 1: Definition and conceptual framework*, 2014.

[10] Axelsson Ö, Nilsson ME, Berglund B. A principal components model of soundscape perception. *The Journal of the Acoustical Society of America*. 2010 Nov 1;128(5):2836-46.

[11] Botteldooren D, Andringa T, Aspuru I, Brown AL, Dubois D, Guastavino C, Kang J, Lavandier C, Nilsson M, Preis A, Schulte-Fortkamp11 B. From Sonic Environment to Soundscape. Soundscape and the Built Environment. 2015 Dec 2:17.

[12] Oldoni D, De Coensel B, Bockstael A, Boes M, De Baets B, Botteldooren D. The acoustic summary as a tool for representing urban sound environments. *Landscape and Urban Planning*. 2015 Dec 31;144:34-48.

[13] Boes M, Oldoni D, De Coensel B, Botteldooren D. A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention. In *Neural Networks (IJCNN), The 2013 International Joint Conference on* 2013 Aug 4 (pp. 1-8). IEEE.