i-ETC: ISEL Academic Journal of Electronics,
Telecommunications and Computers
Vol. 6, n. 1 (2020) ID-2

A Low-Cost Sound Event Detection and Identification System for Urban Environments

J. Alves[a], P. Guerreiro[a], G. Marques[a], J. Paulo[ab]
[a]ISEL, Electronics, Telecommunications and Computer Department, Lisbon, Portugal
[b]A2L, Audio and Acoustics Laboratory of ISEL, Lisbon, Portugal
35189@alunos.isel.ipl.pt A39094@alunos.isel.pt gmarques@deetc.isel.pt jpaulo@deetc.isel.pt

***Abstract*—The ability to automatically detect and classify complex and dynamic urban sounds is an important tool for urban planning such as building efficient noise monitoring, traffic management, surveillance, and urban soundscape mapping. To monitor and ultimately understand a city's sonic environments requires long term measurements and analysis of data collected around the city. In this context, we present a proof of concept for a smart, low-cost, acoustic sensor to be deployed in urban environments. We conducted several preliminary experiments on the two major functioning parts of our system: sound event localization and classification. So far, we are able to detect the direction of arrival of two simultaneous sound sources, and classify audio clips into several predefined classes of urban sounds, but other blocks need to be implemented in order to have a fully functioning smart audio sensor. In this paper, we describe in detail the device's design in terms of its processing blocks, the experiments we performed and their key results, as well as directions of future work.**

Keywords: Urban sound classification, Direction of Arrival, multichannel microphones, machine learning, signal processing.

I. Introduction

A. Sound in Smart Cities

A Smart City relies information and communication technologies, machine learning, and the Internet of Things (IoT) to optimize city functions and operational efficiency, drive economic growth and sustainability, and share information and foster public engagement. Today, 54% of the world population lives in cities, and by 2050, according to the United Nations, it is estimated that this number will reach 68%, with cities in China, Southeast Asia, and Latin America growing the fastest. In the light of these new challenges brought by demographics and worldwide environmental and economic changes, Smart Cities have spiked the interest of policy makers, governments, city planners, researchers and common citizens in general. Smart City technologies is a valuable tool to reduce costs and resource consumption while enhancing efficiency, participation and overall quality of living. Smart Cities rely on wireless sensor networks [1], [2] to gather information and build a live picture of city operations, urban infrastructures, services, and governance. A city is a complex and dynamic entity full of movement, interactions and flows, and it is undeniably linked to the physical phenomenon of sound. For Smart Cities technologies, sound is a rich and always present source of information, with the potential to drive many applications such as audio surveillance [3], [4], traffic management [5], soundscape mapping [6], [7], and noise monitoring [8], [9], [10]. Noise pollution is a serious public health risk linked to distress and other disorders [11] and therefore a great reason of concern for city and health authorities, as demonstrated by the Environmental Noise Directive 2002/49/EC published by the European Commission. However, systematically monitoring city noise has been a difficult task due to the lack of manpower and resources. Typically, enforcement of city laws regarding noise relies on inspectors dispatched to the location of the complaints and the process can be slow and frustrating. Furthermore, and until recently, commercial acoustic devices designed to reliably monitor noise levels were proprietary and expensive, hence not configurable or re-programmable and not easily deployable due to security concerns. The advent of low-cost, versatile computing platforms such as Arduino [12], BeagleBone [13], or Raspberry Pi [14] along with the new emerging reality of IoT which seamlessly connects these devices to the cloud has opened the possibility to uninterruptedly acquire city sounds and process them in real time. Furthermore, these devices also have sufficient computing power to run complex signal processing and machine learning algorithms, making them "intelligent sensors" capable of recognizing/classifying sonic events and characterizing their sonic surroundings. Note that the process of continuously acquiring and processing sound is a demanding task due to the enormous amount of information it generates, making it difficult to deal with the raw data due to space, memory, and other limitations. This technology has open new ways to record and process city sounds because of three main reasons. First, device processing capabilities can greatly reduce the amount of information stored/transmitted. Sound detections techniques based on signal power or other features can be used to select only the relevant parts of the audio stream, and time-frequency representations of the audio can further decrease the amount of information needed. Second, the low-cost makes the devices scalable in a sense that they can be easily deployed in great numbers thus providing a precise and thorough sonic picture of the city. Third, the IoT reality allows an ubiquitous connection to the Internet making it possible to centrally archive, manage, and monitor the transmitted data from the network of sensors. Academic research groups and consortia have already put forward some initiatives to characterize city sounds via a network of low-cost sensors [9], [10], [15].

B. System's Overview

In this work, we introduce a proof of concept for one of such devices: an intelligent, low-cost, urban acoustic sensor for continuous sound acquisition and processing. The sensor is composed of an array of microphones, equipped with signal conditioning and analog to digital converters, and a Raspberry Pi computer which enables the system to run feature extraction, digital signal processing, and machine learning algorithms on the acquired audio signals. Furthermore, this type of devices are cheap and can easily be connected to the cloud, which is an important issue when building a large network of these sensors, following the concept of IoT for Smart Cities. The main contributions presented in this paper are:

- The use of a 3-dimensional sound capturing system, the Mictetrapus. The system is based on a specific configuration of an array of microphones placed in particular geometries which allow for the location and separation of multiple sound sources. In realistic situations, different sound events can coincide in time, and this can hinder performances of classifier and recognition systems. Being able to determine and separate simultaneous sources is an important step to build effective urban audio monitoring devices, since it allows to process the audio from the various microphones, and input to the classifiers single-event audio segments. Furthermore, knowing the location of the sonic events can be an important contribution to video surveillance systems, for instance, by selecting the optimal viewpoint of a 360° camera.
- The use of a Raspberry Pi computing platform allows running signal processing and machine learning algorithms, and the low-cost of the core components provides a versatile and scalable audio sensing system to be deployed in Smart Cities.
- A framework for classifying and automatically tagging complex urban sounds. This is an important aspect of many emerging applications such as noise monitoring, soundscape mapping, and traffic management, and a valuable source of information for city planners.

The remainder of this paper is organized as follows: in Section II we present the proposed framework in terms of hardware setup and in terms of signal processing and machine learning capabilities needed for urban acoustic monitoring tasks, with particular emphasis on sound event detection/localization and sound event classification, which are two core functions of the system. In Section III, we demonstrate the effectiveness of the proposed system through a series of tests on realistic urban sounds; the techniques used for sound localization are tested with real-world recordings, while the performance of the classification methods used are extensively tested on the *UrbanSound8k* dataset [16] in multi-class and multi-label classification scenarios. A discussion on future work and systems improvements is presented in Section V, along with possible ways to enhance urban living by using urban sound processing and classification techniques. Section IV concludes this work.

II. Proposed Framework

We present a proof of concept for a low-cost, intelligent sensor for automatic urban sound classification and tagging. The design took in consideration the final intent to operate in a IoT, Smart City scenario where several of these devices could be deployed around a city to collect urban acoustic information. The systems hardware components consist solely of the Mictetrapus multichannel sound capturing device and the Raspberry Pi Model B 1GB computing platform responsible for processing the audio information. This is accomplished by four main functioning blocks of the processing part of the system (see Figure 1):

- *Event detection and segmentation.* This block detects in the audio streaming the beginning and end of each sound event. Only these audio segments will be analyzed in subsequent blocks;
- *Direction of arrival (DOA) estimation.* DOA estimation is a well studied subject with a significant development of new algorithms in the past three decades [17]. This block is responsible for detecting, in a 3-dimensional space, the direction of the sound sources.
- *Source separation.* This part separates the sound sources based on the captured audio signals. When multiple sound sources occur simultaneously, each microphone captures a different mixture of the sources. Based on independent component analysis (ICA) [18], the clean sources can be recovered from the audio mixtures. The sources are then processed by the classification block.
- *Classification.* This block uses supervised classification algorithms to process and convert the acoustic signal into a symbolic description of sound events.

This work describes the progress made so far, with special emphasis in the DOA and classification blocks of the system.

A. Event Detection and Segmentation

Audio segmentation refers to the class of theories and algorithms designed to automatically reveal semantically meaningful temporal segments in an audio signal, also referred to as auditory scenes [19]. These scenes can be seen as equivalents of paragraphs in text, and can serve as input into audio categorization processes, either supervised (audio classification) or unsupervised (audio clustering). As a first approach, the acoustic events are detected by defining a set of thresholds based on energy. Therefore, an event is set if the sound levels overcome a predefined value for more than some predefined time duration and the event ends if the sound levels drop below the threshold. This simple method has proven effective enough in some experiments we have conducted, but is too rudimentary for the diversity of sonic environments and noise levels found in a city. Further tests have to be carried out to make the detection and segmentation block more robust and versatile.

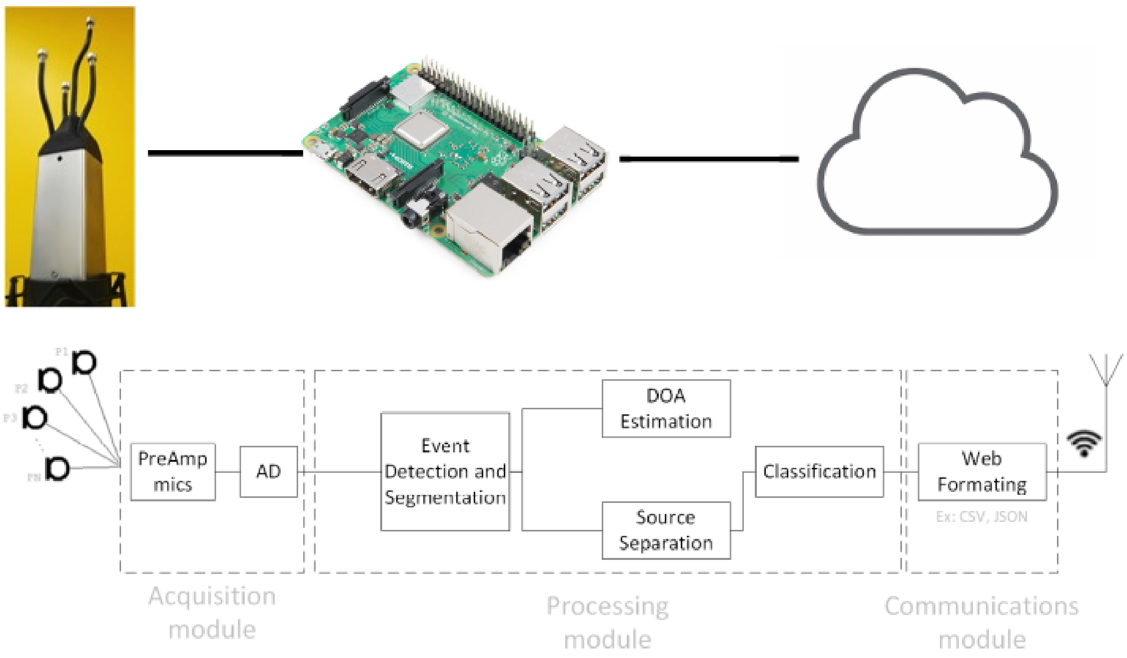


Figure 1. General block diagram of the proposed system of monitoring sound events in an urban environment.

B. Direction of Arrival Estimation

Direction of arrival (DOA) estimation is the task of identifying the relative position of the sound sources with respect to the microphone. DOA estimation is a fundamental operation in microphone array processing and forms an integral part of speech enhancement [20], distant automatic speech recognition [21], spatial audio coding [22], and sound source separation [23]. Popular approaches to DOA estimation are based on time-delay-of-arrival (TDOA) [24], the steered-response-power (SRP) [6], or on subspace methods such as multiple signal classification (MUSIC) [25] and the estimation of signal parameters via rotational invariance technique (ESPRIT) [26]. The aforementioned methods differ from each other in terms of algorithmic complexity, and their suitability to various arrays and sound scenarios. MUSIC specifically is very generic with regards to array geometry, directional properties and can handle multiple simultaneously active narrow band sources. On the other hand, MUSIC and subspace methods in general, require a good estimate of the number of active sources, which are often unavailable or difficult to obtain. Furthermore, MUSIC is computationally intensive and can suffer at low signal to noise ratio (SNR) and in reverberant scenarios [27].

ICA based methods have been proposed to identify the delay between microphones corresponding to a source direction. Another popular algorithm is Steered Response Power with Phase transform (SRP-PHAT) [27], which is the generalization of Generalized Cross Correlation with Phase transform (GCC-PHAT) for more than two microphones. GCC-PHAT allows for a direct computation of the DOA using the time delay belonging to the maximum of the cross correlation function of the signals. SRP-PHAT evaluates the cross correlation functions of all pairs of microphones for each candidate direction, adds up the GCC-PHAT scores and finally performs a maximum search to obtain an estimate of the DOA. Another group of algorithms devises a distance measure between an observation and all possible candidate directions. It is employed a cosine distance between the observed phase and the candidate models, i.e., the expected observations in an anechoic sound field originating from a given direction.

For the purpose of determining the DOA of a sound in a scenario of multiple simultaneous sound sources, several approaches were considered balancing accuracy and computational load, keeping in mind the utilization of IoT devices. Among of the most common methods of tracking of sounds, we considered in this paper, the approaches of GCC-PHAT and the ICA technique. Although GCC-PHAT algorithm presents good results when used to estimate the direction of arrival based on that delay, it is not useful when used in an environment with multiple source signals. For this reason, the ICA method was considered, as it allows for the separation of multiple statistically independent signals, as long as the number of sources is not superior to the number of sensors.

Despite this limitation, the GCC-PHAT method, allows for a much faster estimation of the direction of arrival, given its lower complexity when compared to ICA. As such it would still be a viable option when considering only one sound source. Figure 2 shows the topology of microphone array setup, the Mictetrapus, designed and built in collaboration with FI-Sonic company. This device consists of four articulated arms, ending on omni-directional microphone capsules. Therefore, it has the ability of changing the distances and topology of the array easily. Although the distance between microphones can be arbitrary selected, for the purposes of this study, the distance of 2.5 cm and 5 cm were used. In the setup depicted in Figure 2, the microphones are 5 cm apart; tests with 2.5 cm distances are reported in Section III-A.

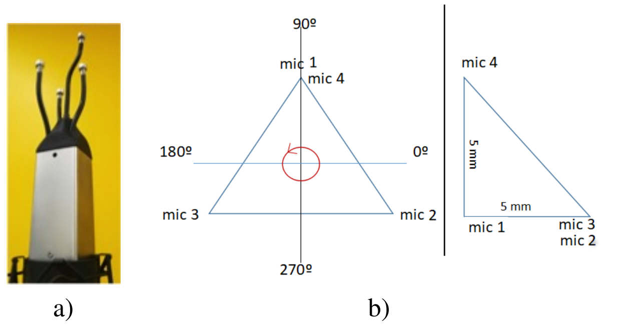


Figure 2. Mictetrapus: microphone array developed using a triangular topology. Mics 1, 2 and 3 are in horizontal plane, along a circumference, with an angle of 120 degrees between them, and mic4 is on the vertical of mic1. a) photo of the apparatus developed and b) top and lateral views of the microphone array.

Figure 3 depicts the procedures used to estimate the direction of arrival of an acoustic event and source separation.

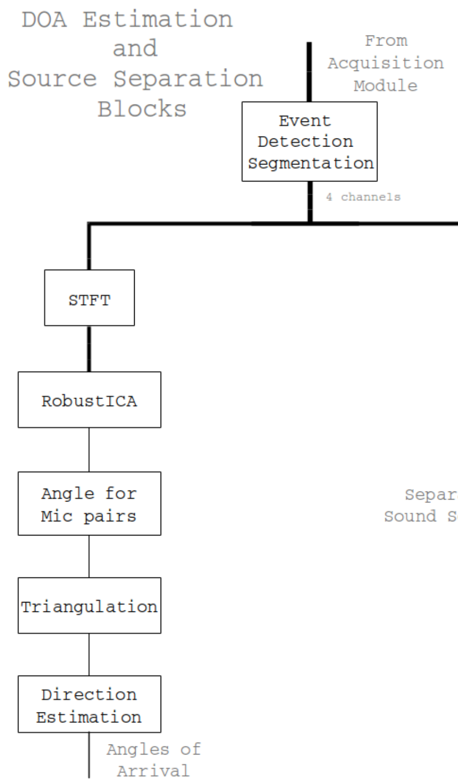


Figure 3. Simplified flowchart for the DOA estimation and source separation procedures.

Considering an acoustic urban environment, the four signals acquired by the Mictetrapus form a mixture of the different sources. Each of these signals will have delays for each source, corresponding to the microphones spacial position. Therefore, the signals are later separated by the ICA procedure. A conversion to frequency domain is applied before ICA, because of limitations of the ICA [28] technique when dealing with non-linear or non-instantaneous mixtures, as is the case, given the signals show delays in time domain. Since the convolution operation in the time domain turns into a multiplication in frequency domain, the signals are viewed as instantaneous mixtures.

To perform this transition into the frequency domain, the Short-Time Fourier Transform (STFT) is used on each mixture, resulting a matrix of frequencies indexes, bins, for different time frames. In other words, the spectrogram of the mixture is obtained, which represents the variation of each frequency index across a given time frame. This results, however, in mixtures which are composed of complex values, meaning that the ICA procedure, for our purpose, should be able to deal with these type of values. For this reason, the toolbox RobustICA [29] is considered in this work.

In fact, ICA is essentially a statistical technique often used for the separation of any signal that composes a given mixture that is often incomprehensible, with the purpose of obtaining a clean signal without the interference of the remaining signals present in the mixture. This is a Blind Source Separation problem, BSS. ICA method can make that separation as long as certain conditions are met [28], by computing a separation matrix that allows the retrieval of each individual source. Also, these separation matrices allow the current method can extract the direction of arrival of the source for each pair of microphones.

Therefore, after applying the STFT, the four mixtures are analyzed by ICA, one frequency at a time, and for each frequency index, a separation matrix is computed via ICA. The idea is that each separation matrix has information on the direction of arrival of the signal, so by partially isolating a signal through the use of frequency indexes and computing their corresponding matrix, it is possible to obtain the direction of arrival for each individual signal.

The direction of sound source is estimated according to [30], where the angle of arrival, θ_p, for each pair of microphones is obtained:

$$\theta_p = \cos^{-1}\left(\frac{\angle(H_{qp}/H_{q'p})}{2\pi f c^{-1}(d_q - d_{q'})}\right) \tag{1}$$

where $\angle$ is the angle operator, f and c are the frequency and the speed of sound, respectively. This angle is calculated based on the distance $d_q - d_{q'}$ between microphones q and q', and the frequency responses, H_{qp} and $H_{q'p}$, of the mixing system, $H(f)$, obtained via ICA.

This approach has, however, its limitations, namely the ambiguities of scaling and permutation of ICA which affect the accuracy of the reconstructed signals. Another limitation is the poor results observed when trying to reconstruct the sources after going through all these steps, where it was not possible to determine which of the sources correspond in each of the obtained directions.

After obtaining the estimated angles for each pair of microphones, it is necessary to calculate the final angle considering the given referential of the microphone array for the considered topology. Therefore, it is necessary to determine whether the angles obtained from Equation 1, are facing ’forward’ (0 to 180 degrees) or ’backward’ (180 to 360 degrees), since the signal from a source originated from a k degree angle is virtually the same as a signal originated from a (360-k) degree angle. This means that from Equation 1 the estimated angle will always be between 0 and 180 degrees.

To overcome this redundancy, the setup of the microphone array is used in a process of elimination to find out which of the angles are facing 'forward' and which are actually facing 'backwards' and need to be corrected with that information.

C. Source Separation

In the present study, sound (event) source separation refers to the task of estimating the signals produced by individual sound sources from a mixture of these signals. In our setup, the microphone array captures different delayed mixtures when there are co-occurring sound events. Our goal is to extract the clean source signals based solely on the captured mixtures. This is a hard problem, since by nature, the mixing is a convolutive process and the captured signals are linearly filtered versions of the sources. The task is to determine the inverse filters: this is a blind source separation (BSS) problem that can be solved via independent component analysis methods [31]. One of the approaches to BSS is to transform the signals to the frequency domain. Due to the properties of the Fourier transform, convolutive mixtures in the time domain can be obtained by multplying the signals spectra, which makes the BSS easier since now we deal wiht multiplicative instantaneous mixtures. However, the permutation ambiguity of ICA solutions becomes a problem. There are methods that deal with this permutation problem [30]. Another way to tackle the BSS problem is to directly estimate the de-mixing filters in the time domain [32].

D. Classification

The task of automatically classifying dynamic, complex urban sounds is an important aspect of many emerging applications in the context of IoT and Smart Cities, such as noise monitoring, surveillance, and soundscape characterization, and therefore has recently gained a lot of attention from the academic community [10], [33], [34], [35], [36]. In order to classify sound events, it is necessary to extract from the audio signal a set of representative acoustic features. These features are generally derived from audio signal processing techniques inspired from the research fields of automatic speech recognition and music information retrieval [37], [38]. Time-frequency features are obtained by dividing the time signal into overlapping, short-length segments, in which spectral descriptors are calculated, like the all-purpose Mel-frequency cepstral coefficients (MFCCs) [37]. In this work, the signal was divided into 43 milliseconds segments (2048 samples at a sampling frequency $F_s = 48\text{kHz}$), with 50% overlap, and we used 50 Mel bands to extract 20 MFCCs, plus the spectral roll-off, spectral centroid and the zero crossing rate, features also commonly used in content-based music information applications [39]. This process converts the audio signal into a sequence of 23-dimensional vectors. Standard machine learning classification algorithms only deal with vectors not vector sequences, and therefore the feature sequence has to be converted into a single vector. Our approach was to summarize the sequence of features using the median and the standard deviation. The median was used instead of the mean since this statistic is more robust to outliers. This way each audio excerpt is represented by a 46-dimensional vector. The feature vectors were pre-processed via principal component analysis, PCA, followed by variance normalization (whitening) [40]. In the context of urban sounds, the tests we conducted with normalized features showed better results versus non-normalized ones, which was also corroborated by other researchers [36], [41]. For classification we used a standard k-nearest neighbors (k-NN) algorithm [40], and a support vector machines (SVM) [42]. The k-NN was used in order to ascertain a baseline performance on urban sound event classification. The k-NN is an instance-based learning, where class membership is assigned based on a majority vote of its neighbors, and is possibly one of the simplest classification methods, and therefore it is well suited for a baseline system. The SVM were chosen for their performance and the generalization capabilities, particularly in high-dimensional spaces. Furthermore, SVM are binary classifiers and ideal for automatic tagging of sound events, which can be considered a binary classification problem (see Section III-B). A representation of the whole process leading to the classification of a sound excerpt is shown in Figure 4. The implementation was done in the Python programing language, using *librosa* package [43] for audio feature extraction, and *scikit-learn* package [44] for data preprocessing and classification.

III. Experimental results

A. DOA Estimation

The source of sound material used in this study was the *UrbanSound8K* dataset[1][16]. For the estimation of DOA, tests were performed in anechoic chamber facility as a compromise between ideal conditions and outdoors scenario in absence of any obstacles of significant dimensions capable to produce sound reflections and diffraction. Therefore, results are not affected by issues like the reverberation of the room, when using the proposed setup of the array of microphones for the tracking of the direction of arrival of the different sounds in a three-dimensional space. Moreover, the effects of diffraction of sound waves on the surface of the capturing sound device, the Mictetrapus, are therefore considered. Figure 5 shows the setup installed in the anechoic chamber for the experiments.

The results are obtained when testing the complete system with two sources separated by 45 degrees on the horizontal plane, for different types of signals. Source 1 can be lifted to modify the elevation angle.

Tests used a sampling frequency $F_s = 48000$ Hz, and a NFFT (number of points used for calculating the Fast Fourier Transform - FFT) of 2048. Tests were made for distance between microphones of 5 cm and 2.5 cm. The audio material used in the tests consists of a number of different type of daily sounds encountered in urban environment to simulate the city life, such as, Firetruck sirens, Ambulance sirens, General sirens and horns, Gunfire, Airplane, Dog barking, Truck engine, Music, Explosion, Shot bursts, Traffic, etc.

[1] publicly available at: https://urbansounddataset.weebly.com/urbansound8k.html

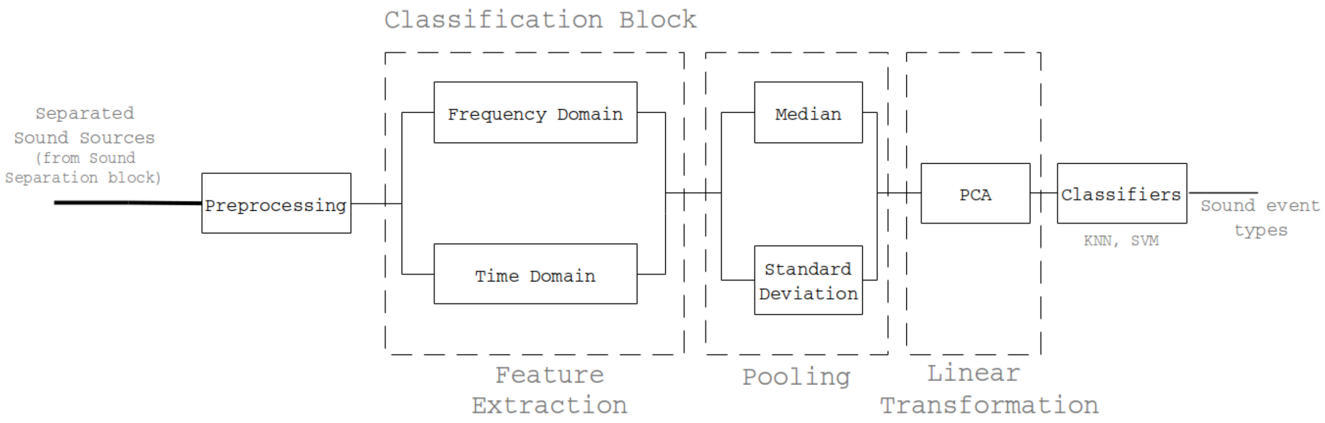


Figure 4. Block diagram for the sound event classification process.

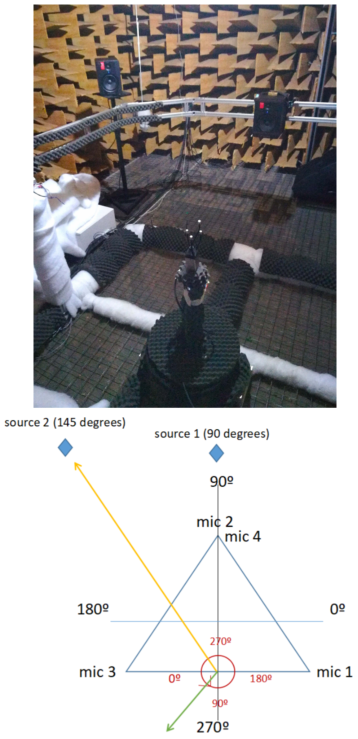


Figure 5. Photograph of the experimental setup in the anechoic chamber facility and configuration of Mictetrapus device with the locations of two sound sources (source 1 and 2) used in the experiments. Also, it is shown the ambiguity caused by this geometry of microphone array, 'mirror effect', when sound waves come from different locations (yellow and green directions).

The graphics shown in Figure 7 represent the results obtained when testing the system for different angles and using two different sources in every test, to confirm its performance when faced with diverse types of signals. Each point in the x-axis of Figure 7 represents a different combination of simultaneous signals that were used in the test. These combinations are as follows in Table I.

Considering the computation time of the direction estimation for non-simultaneous sound sources, the GCC-PHAT achieved 0.08 seconds against 2.2 seconds using ICA. Therefore, for scenarios of sound tracking of just one source GCC-PHAT should be used, leaving more computational resources of CPU for other tasks.

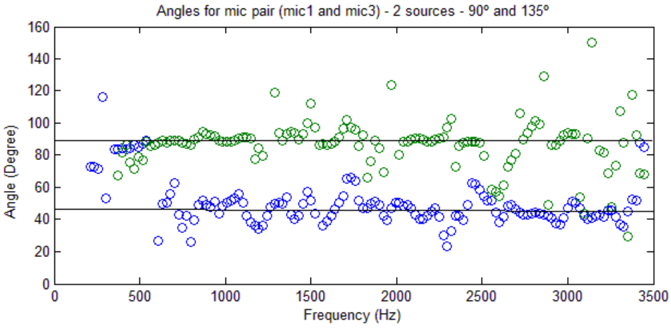
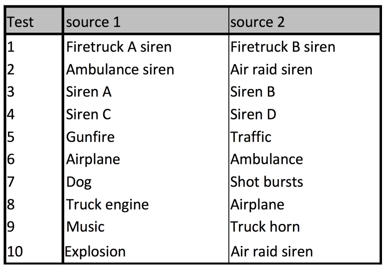


Figure 6. Angle estimation for one pair of microphones for each frequency index. The upper bound frequency is 3.5 kHz to avoid spatial alising.

Table I
COMBINATION OF SOUND TYPES USED IN THE TESTS.



Test	source 1	source 2
1	Firetruck A siren	Firetruck B siren
2	Ambulance siren	Air raid siren
3	Siren A	Siren B
4	Siren C	Siren D
5	Gunfire	Traffic
6	Airplane	Ambulance
7	Dog	Shot bursts
8	Truck engine	Airplane
9	Music	Truck horn
10	Explosion	Air raid siren

In practice, this sample rate is often too low to achieve a good angle resolution. Thus, interpolation approaches are required to obtain more precise angle estimates. In fact, a maximum delay of 3 and 7 samples is expected for a distance between capsules of 2.5 and 5 cm, respectively, which is usually very low. In this work, the re-sampling technique was used with a ratio of 4, which is a reasonable balance between angle resolution improvement and load computation.

Figure 6 represents the DOA estimation for each frequency of two sources using Equation 1, for the microphone pair composed by mic1 and mic3. In this example, for source 2, where the green arrow corresponds to 45 degrees (obtained), from the pair mic3 and mic1 (referential in red color), the yellow arrow (expected) is calculated as (360 - 45) = 315. Moreover, after adjusting the red referential of the microphone pair, to the blue referential of the microphone array setup in Figure 2, the final angle will be (315-180) = 135.

The final angle is calculated based on all the DOA angles previously obtained for a given source, using trigonometry relations and triangulation methods.

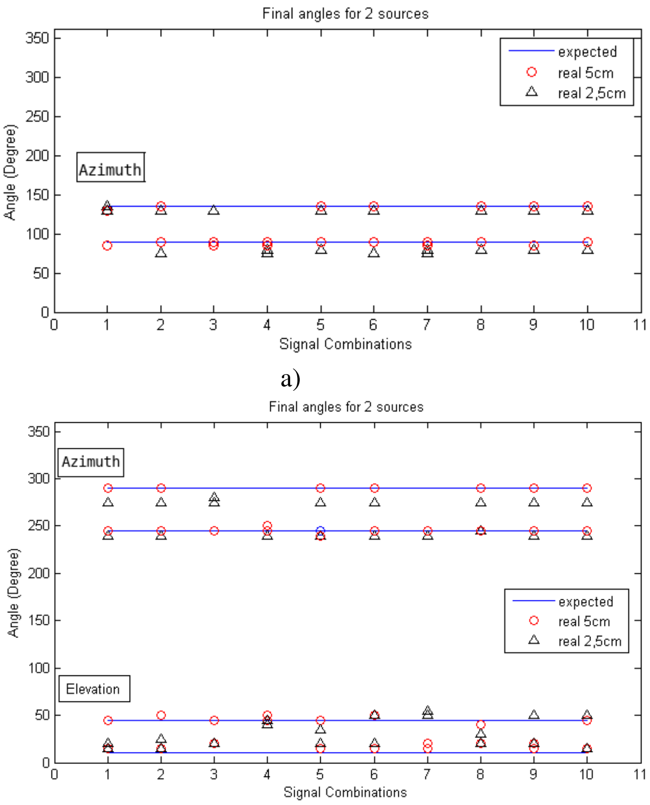
As shown in Figure 6, the angle of one source (blue) sometimes overlaps with the angle of the other source (green), which is a consequence of ICA permutation ambiguity. Another key factor, is that despite the second source being at 135 degrees, the result points to 45 degrees. This is because, as previously stated, the angles returned by Equation 1 only vary between 0 and 180 degrees, and that when the angle is above 180, there is a 'mirror effect' as shown in Figure 5 (green arrow):

In Figure 7.b), elevation angles are introduced in one of the sources, to test the use of the 4th microphone which will return the angle for each source, from a vertical point-of-view, thus turning a direction in a two-dimensional space, into an estimation in a three-dimensional space.

From the results obtained in Figure 7, it is possible to conclude that the distance between microphones affects the outcome of the system, and causes a greater margin of error for shorter distances. This stems from the fact that this distance will influence the resolution of the final angle, that is correlated with the maximum possible number of samples between each microphone and the speed of propagation of the signal (speed of sound), meaning that the resolution becomes smaller for shorter distances. However, the increase of the distance of the microphones results in a reduction of the effective bandwidth of the audio contents due to the effect of spacial aliasing.

Additionally, it can be concluded that the worst results were observed when using two sources with very similar spectra (refer to Table I combinations: 3 and 4) and for sources that show very impulsive profiles such as, the barking of a dog or several periodic bursts of shots (refer to Table I combination 7), since these types of signals tend to have several pauses between bursts.

The system was also tested using real sound signals with simulated delays based on the location of a virtual source. These tests, consisting of changing synthetically the time delay from each virtual source to the microphone array, are very close to the results in which a real recording was used, thus, are not shown in the paper.



a)

b)

Figure 7. Estimated directions of sound sources for a combination of two simultaneous acoustic signals. The blue lines correspond to the true directions of the two sources (in degrees), and the circle and triangular markers correspond to the estimates obtained with 5 cm and 2.5 cm distances between microphones, respectively. a) Azimuth angle obtained using three microphones for two simultaneous occurring sources at different positions. b) Azimuth and elevation angles obtained using four microphones for two simultaneous occurring sources at different positions.

B. Classification

We tested the effectiveness of the classification system on the *UrbanSound8K* dataset through several systematic tests. The dataset is designed for a multi-class problem, where each example belongs to one of ten predefined, mutually exclusive classes. A brief description of the dataset is given in the next paragraph and the results obtained with multi-class problem afterwards. Urban sounds are dynamic, chaotic and can be composed of several co-occurring events, and the multi-class scenario is a bit too restrictive for our goals. Ideally we would like our system to be able to classify simultaneously occurring audio events. Hence a more realistic setting is the multi-label setting where sound clips can be assigned multiple labels, or tags. This problem is known as auto-tagging which refers to the task of assigning each audio excerpt a set of high level concepts (the tags). This problem is usually divided into sets of binary classification problems, one for each tag. In the experiments we conducted, the tags were the class labels and we present the results at the end of this section.

1) Data and Testing Methodologies: In recent years, several new datasets for environmental sound classification tasks have been released (*e.g.* [16], [45], [46]). In order to evaluate the proposed approach, we chose the *UrbanSound8K* dataset which includes ten classes of urban sounds with 8732 real-world sound clips. A brief taxonomy of the classes, along with the number of audio segments per class is shown in Figure 8. The clips span ten environmental sound classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. The reason for choosing this dataset is because it encompasses urban noises and emergency sounds which are highly related to city life and thus suitable for testing urban sound classification algorithms, and also because we can compare our results with previously published approaches evaluated on the same dataset [36], [47], [48]. We used a ten-fold cross-validation in all our test with the folds provided by the *UrbanSound8K* dataset. All the performance values are obtained by averaging the results on the test folds.

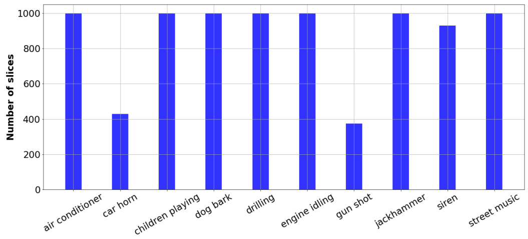


Figure 8. *UrbanSound8K* dataset - number of samples per class.

2) Multi-Class Classification: The confusion matrices pertaining to the classification results of k-NN and SVM are shown in Figures 9 and 10. These results were obtained after the parameters of the classifiers were adjusted to yield the best performances. We tested the k-NN with a different number of neighbors and chose $k=9$ empirically. The overall accuracy for this classifier was 54%. For the SVM, we tested several kernels and opted for a radial basis function one, and through experimentation we also calibrated the regularization parameter. The overall accuracy for this classifier was 65%. Analyzing the figures, we can see that there is some confusion between classes with mechanical, repetitive sounds, namely air conditioner, drilling, engine idling and the jackhammer. There is also some error between children playing and street music, since they also have similar acoustic patterns. The dataset also provides a salience level for the audio clips indicating whether the event was perceived to be in the foreground or the background of the recording. Foreground sounds are salient and normally undistorted by other sounds, while background sounds are often mixed with environmental or other types of noise. For the SVM model, the classification was done for the whole dataset and also for foreground and background sounds separately. The accuracies per class are shown in Figure 11.

The performances of our model are still below some of the results found in the literature. In [36], an mixture of expert models combined with local and global features achieves an accuracy of 77%, while [47] and [48] use deep convolutional neural networks (CNN) and obtain an accuracy of 73% and 79% respectively. Our sub-optimal performances are mainly due to two factors. The first has to do with the features used in our models. The descriptors are based on short-time spectral characteristics of the acoustic signals and do not take in account medium or long term temporal relationships that differentiate urban sounds. This is due to the pooling process that converts the feature sequence into a single vector, discarding temporal dependencies. This process is commonly known as the bag of frames approach, and methods that use this formulation have limited capabilities [49]. The second reason has to due with the type of classification models used. As the referenced works show, better performances can be obtained with CNN. Furthermore, single multi-class models may not be the best strategy to tackle the task of urban sound classification. Specific urban sounds can be grouped in broad categories such as mechanical or motorized sounds, pitch sounds such as sirens or car horns, impulsive sounds such as bursts and shots, nature sounds, human sounds, and many others. These categories all have distinctive acoustic signatures and therefore it may be beneficial to train models separately for each type of broad category rather than building a model for all sound classes.

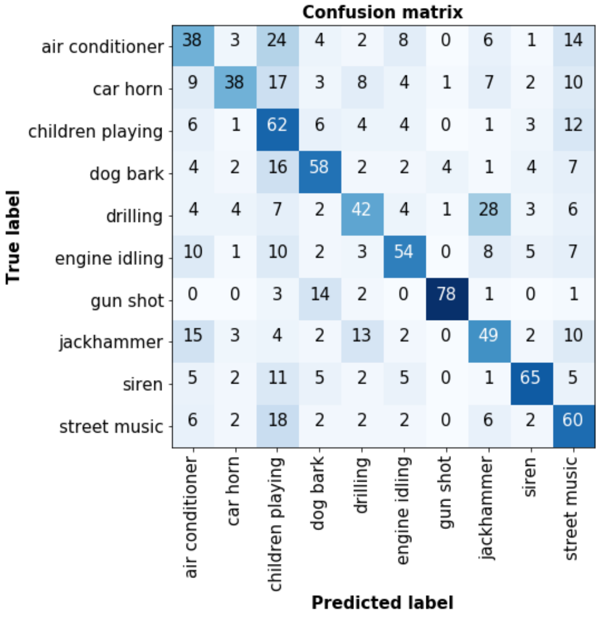
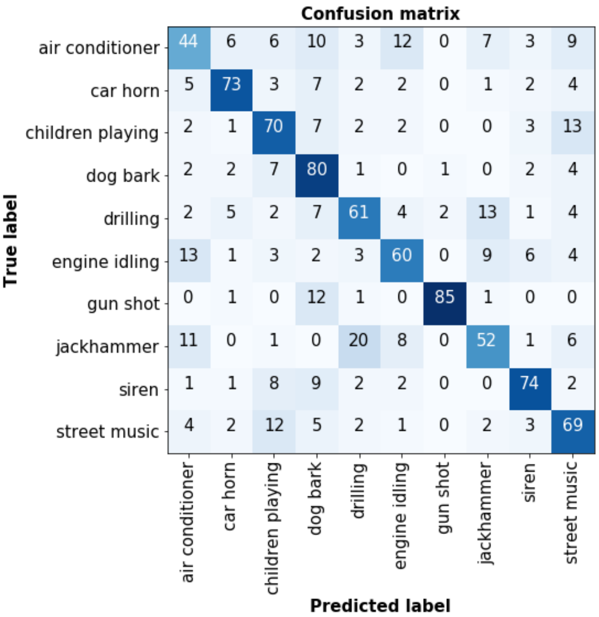


Figure 9. *UrbanSound8k* - Confusion Matrices for the k-NN classifier (values in percentages). k-NN overall accuracy: 54%.

3) Multi-Label Classification: In this section, we present the results for the multi-label classification. For this task, ten SVM binary classifiers were trained, one for each class, using as the positive examples the class members, and as negative examples the sound clips of all other classes. Compared to the previous experiments, this is a more realistic scenario since a single sound clip can be annotated with multiple classes, a situation which is applicable to several audio excerpts present in the *UrbanSound8K* dataset. Note that in this binary classification problem we are dealing with highly imbalanced sets, and therefore, the accuracy is a misleading metric since assigning a negative label to all test examples results in accuracies around 90%. More reliable performance measures are obtained by analyzing receiver operating characteristics (ROC) curves [50]. ROC curves are two-dimensional graphs that depict the trade-offs between benefits (true positives) and costs (false positives). The results for our models are shown in Figure 12. SVMs, like many other classification models, can output a score reflecting the degree of certainty in the decision along with the predicted class. Using different threshold with the classification scores can produce more "conservative" classifiers that make positive classifications only with strong



	air conditioner	car horn	children playing	dog bark	drilling	engine idling	gun shot	jackhammer	siren	street music
air conditioner	44	6	6	10	3	12	0	7	3	9
car horn	5	73	3	7	2	2	0	1	2	4
children playing	2	1	70	7	2	2	0	0	3	13
dog bark	2	2	7	80	1	0	1	0	2	4
drilling	2	5	2	7	61	4	2	13	1	4
engine idling	13	1	3	2	3	60	0	9	6	4
gun shot	0	1	0	12	1	0	85	1	0	0
jackhammer	11	0	1	0	20	8	0	52	1	6
siren	1	1	8	9	2	2	0	0	74	2
street music	4	2	12	5	2	1	0	2	3	69

Figure 10. *UrbanSound8k* - Confusion Matrices for the SVM classifier (values in percentages). SVM overall accuracy: 65%.

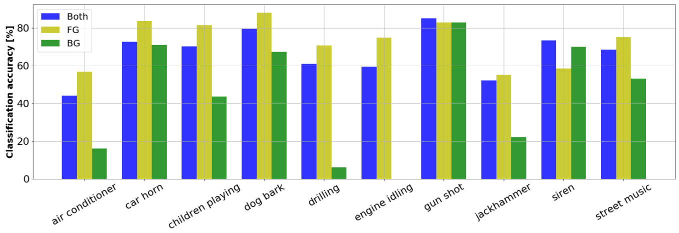


Figure 11. *UrbanSound8k* - Class accuracies for the SVM classifier. FG - class accuracies on foreground sounds. BG - class accuracies on background sounds.

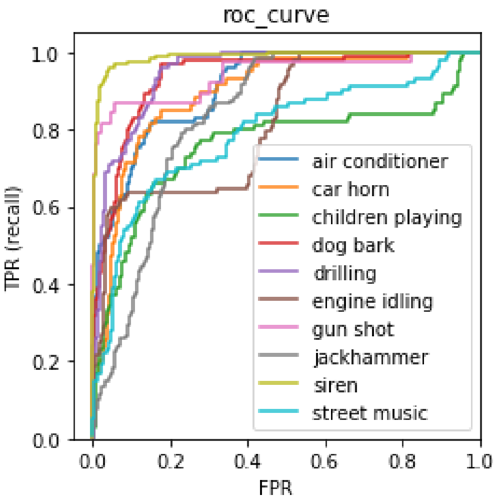


Figure 12. *UrbanSound8k* - ROC curves for the binary classifiers.

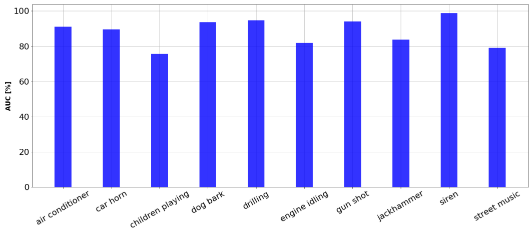


Figure 13. *UrbanSound8k* - Area under the ROC curve for the binary classifiers.

evidence or more “liberal” ones making positive classifications with weak evidence. The threshold can be adjusted to reach the optimal operational point, which is located on the top left-hand corner of the ROC curve (coordinate (0,1)). The information in the ROC curve can be summarized by calculating the area under the ROC curve (AUC) [50], and the results for our models are shown in Figure 13. Analyzing the results, siren sounds are the ones with better predictions but other classes like dog bark and gun shots also show good results. On the other hand, classes like children playing, street music or engine idling, have worse performances.

IV. Conclusions

This paper focuses on urban sound content retrieval and sound event location which is regarded as an essential building block of smart cities. The proposed system assesses the sound field and extracts the relevant information about the sound events occurring in the city such as excessive sound levels, type of sound sources and location of the sound events. In fact, the urban sound content retrieval and sound event location is regarded as an essential building block of Smart Cities, or the new term, Happy Cities, in the sense to provide comfort and security to the citizens. Therefore, a proof of concept of a low-cost, intelligent sound station for automatic urban sound classification and tagging and event location was studied and developed. The design took in consideration the final intent to operate in a IoT concept, where several of these devices could be deployed around a city to collect urban acoustic information.

The multi-label classification was developed to able to handle more than one sound event simultaneously, which is a more realistic scenario in conformity urban environments.

The results obtained for the accuracy for the identification of sound types, considering all elements of the database (both), is relatively low, 65%. However, if we consider a situation where we guarantee better segmentation and separation of sound events and a better SNR(foreground), the results reach 73%. In a real-time operating situation and using the event localization and source separation module, the classification results should improve since we are dealing with separated sources instead of a mixture of sounds.

The method used to the event location (DOA) achieved good results in estimating the direction of arrival up to two simultaneous sound sources, with an error of less than 3 degrees, for anechoic environment conditions.

Although these results are promising, much remains to be done and improved.

V. Future Work

So far we have implemented the detection, direction of arrival, and classification blocks of the system. The detection and segmentation need to be improved and the separation block of the system is still in a preliminary phase of development and further tests need to be done to determine the method that better suits our purposes.

There are also other considerations that must be addressed in order to have a fully functioning smart audio sensor. The first has to do with the processing capabilities of the computing platform. We need to guarantee that algorithms can operate in real time without consuming the computing resources of the system. We undertook some tests that show that the classification and tagging algorithms run in real-time on the Raspberry Pi. The process included the whole classification chain beginning with the raw audio, extracting features, representing the audio segment with a feature vector, and finally predicting the tag or class label.

In fact, after the execution of the classification routine, the Raspberry Pi took less than 7 secs to perform the classification against 2 secs when running on an Intel (R) Core (TM) i7-4710HQ CPU, for the k-NN and SVM classifiers. Therefore, improvements in algorithms need to be done.

We still have to test other parts of the system such as the DOA estimation and the source separation blocks, and it is foreseeable that we will need to make some concessions in terms of the complexity of the methods we choose.

Another aspect we need to address is how to manage and transmit the information produced by the sensor when it is continuously capturing audio. We have to determine how much data the system is capable of transmitting to a centralized computer, and still be able to fulfill its other processing requirements. Finally, we need to tackle practical issues such as devices housing, long-term exterior exposure, and power requirements.

Acknowledgements

This project, research and development, has been partially supported by the Audio and Acoustics Laboratory of ISEL for the development of the Mictetrapus multichannel sound capturing device and CAPS/IST for the availability of using the anechoic chamber to perform the acoustic tests.

References

[1] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami. An information framework for creating a smart city through internet of things. *IEEE Internet of Things Journal*, 1(2):112–121, 2014.

[2] B. Lau, N. Wijerathne, B. Ng, and C. Yuen. Sensor fusion for public space utilization monitoring in a smart city. *IEEE Internet of Things Journal*, 5(2):473–481, 2018.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino. Audio surveillance: A systematic review. *ACM Comput. Surv.*, 48(4):52:1–52:46, 2016.

[4] R. Radhakrishnan, A. Divakaran, and A. Smaragdis. Audio analysis for surveillance applications. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 158–161, 2005.

[5] G. Nagy, R. Rodigast, and D. Hollosi. Energy based traffic density estimation using embedded audio processing unit. In *Audio Engineering Society Convention 136*, 2014.

[6] A. Torija, D. Ruiz, and Á. Ramos-Ridao. A tool for urban soundscape evaluation applying support vector machines for developing a soundscape classification model. *Science of The Total Environment*, 482-483:440 – 451, 2014.

[7] M. Rychtáriková and G. Vermeir. Soundscape categorization on the basis of objective acoustical parameters. *Applied Acoustics*, 74(2):240 – 247, 2013.

[8] A. Agha, R. Ranjan, and W.-S. Gan. Noisy vehicle surveillance camera: A system to deter noisy vehicle in smart city. *Applied Acoustics*, 117:236 – 245, 2017.

[9] I. Kivelä, C. Gao, J. Luomala, and J. Ihalainen. Design of networked low-cost wireless noise measurment sensors. *Sensors and Transducers*, 10:171–190, 2011.

[10] C. Mydlarz, J. Salamon, and J. Bello. The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics*, 117:207 – 218, 2017.

[11] H. Ising and B. Kruppa. Health effects caused by noise : Evidence in the literature from the past 25 years. *Noise and Health*, 6(22):5–13, 2004.

[12] Arduino. https://www.arduino.cc/, 2018.

[13] BeagleBone. https://beagleboard.org/, 2018.

[14] Raspberry Pi. https://www.raspberrypi.org/, 2018.

[15] M. Bell and F. Galatioto. Novel wireless pervasive sensor network to improve the understanding of noise in street canyons. *Applied Acoustics*, 74(1):169 – 180, 2013.

[16] J. Salamon, C. Jacoby, and J. Bello. A dataset and taxonomy for urban sound research. In *ACM Multimedia*, 2014.

[17] P.-J. Chung, M. Viberg, and J. Yu. Chapter 14 - DOA estimation methods and algorithms. In *Academic Press Library in Signal Processing*, volume 3, pages 599 – 650. Elsevier, 2014.

[18] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[19] L. Lu, R. Cai, and A. Hanjalic. Audio elements based auditory scene segmentation. In *International Conference on Acoustics, Speech and Signal Processing*, volume 5. IEEE, 2006.

[20] M. Wölfel and J. McDonough. *Distant speech recognition*. John Wiley & Sons, 2009.

[21] T. Sainath, R. Weiss, K. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):965–979, 2017.

[22] A. Politis, J. Vilkamo, and V. Pulkki. Sector-based parametric sound field reproduction in the spherical harmonic domain. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):852–866, 2015.

[23] J. Nikunen and T. Virtanen. Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):727–739, 2014.

[24] Y. Huang, J. Benesty, G. Elko, and R. Mersereati. Real-time passive source localization: a practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, 9(8):943–956, 2001.

[25] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

[26] R. Roy and T. Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on acoustics, speech, and signal processing*, 37(7):984–995, 1989.

[27] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In *Microphone Arrays*, pages 157–180. Springer, 2001.

[28] G. Naik and D. Kumar. An overview of independent component analysis and its applications. *Informatica*, 35(1), 2011.

[29] V. Zarzoso and P. Comon. Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size. *IEEE Transactions on Neural Networks*, 21(2):248–261, 2010.

[30] H. Sawada, R. Mukai, and S. Makino. Direction of arrival estimation for multiple source signals using independent component analysis. In *Seventh International Symposium on Signal Processing and Its Applications*, volume 2, pages 411–414, 2003.

[31] J. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[32] H. Buchner, R. Aichner, and W. Kellermann. Trinicon-based blind system identification with application to multiple-source localization and separation. In *Blind speech separation*, pages 101–147. Springer, 2007.

[33] G. Hancke, B. Silva, and G. Hancke Jr. The role of advanced sensing in smart cities. *Sensors*, 13(1):393–425, 2012.

[34] H. Phan, M. Maaß, R. Mazur, and A. Mertins. Random regression forests for acoustic event detection and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):20–31, 2015.

[35] A. Rakotomamonjy and G. Gasso. Histogram of gradients of time-frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(1):142–153, 2015.

[36] J. Ye, T. Kobayashi, and M. Murakawa. Urban sound event classification based on local and global features aggregation. *Applied Acoustics*, 117:246 – 256, 2017.

[37] B.-H. Juang and L. Rabiner. Automatic speech recognition–a brief history of the technology development, 2005.

[38] M. Schedl, E. Gómez, and J. Urbano. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, 2014.

[39] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer-Verlag, 2006.

[40] R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[41] A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th international conference on machine learning*, pages 921–928, 2011.

[42] C. Cortes and V. Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[43] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[45] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, pages 85–92, 2017.

[46] K. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 1015–1018, New York, NY, USA, 2015. ACM.

[47] K. Piczak. Environmental sound classification with convolutional neural networks. In *IEEE 25th International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2015.

[48] J. Salamon and J. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24:279–283, 2017.

[49] M. Lagrange, G. Lafay, B. Défréville, and J.-J. Aucouturier. The bag-of-frames approach: A not so sufficient model for urban soundscapes. *The Journal of the Acoustical Society of America*, 138(5):EL487–492, 2015.

[50] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.