euronoise 2021
25th to 27th of October
Madeira, Portugal - online

Multilabel acoustic event classification for urban sound monitoring at a traffic intersection

Ester Vidaña-Vila[1], Dan Stowell[2], Joan Navarro[3], Rosa Ma Alsina-Pagès[1]

[1]GTM- Grup de Recerca en Tecnologies Mèdia, La Salle – Universitat Ramon Llull, Barcelona, Spain
{ester.vidana, rosamaria.alsina}@salle.url.edu
[2]Department of Cognitive Sciences & Artificial Intelligence, Tilburg University, Tilburg, Netherlands
[3] GRITS- Grup de Recerca en Internet Technologies and Storage, La Salle – Universitat Ramon Llull, Barcelona, Spain
jnavarro@salle.url.edu

Abstract
Persistent exposure to city noise has a great impact on the population's well-being. Due to their intrinsic characteristics, different noise sources have different effects on citizens' health. Automatically detecting and classifying acoustic events in urban environments would allow public administrations to monitor the city soundscape and, thus, to identify harmful noise sources and quantify their impact on people. One of the main challenges when classifying acoustic data in real-operation environments, such as urban scenarios, is the presence of simultaneous noise sources. The purpose of this paper is to propose a system able to detect and classify, in real-time, a predefined set of urban acoustic events that may occur simultaneously. More specifically, the proposed approach features a multi-label deep-learning-based algorithm that runs over a low-cost wireless acoustic sensing node. The system has been tested using real-world recorded data to evaluate its feasibility and accuracy.

Keywords: multilabel classification, real-world data, acoustic event detection.

1 Introduction

It is estimated than 20% of European Union (EU) population might be exposed to levels of noise pollution that are above the limits of current regulations. Indeed, citizen concerns regarding environmental health and noise pollution have been consistently rising in the recent years. Acoustic noise (or pollution) can be defined as any sound that is loud or unpleasant enough that causes some kind of disturbance [1]. Such disturbance may range from difficulties in understanding a voice message to some serious adverse health effects such as heart diseases or psychological disorders derived from lack of rest or sleep [2]. Nonetheless, it is well-known that not all sound sources have the same impact on human disturbance as the sound level is not the only parameter that indicates the extent and intensity of noise pollution [3]. Therefore, identifying the sources of those potentially harmful sounds has emerged as a hot research topic nowadays.

So far, several efforts have been made by public and private entities on identifying acoustically polluted environments in urban areas [4]. Typically, this is done by either analysing the distribution of noise-related complains in a certain area, or by deploying a wireless acoustic sensor network (WASN) to automatically monitor the environment [4]. Both approaches entail the same underlying challenge: identifying the acoustic sources—considering that several events coming from different sources may occur concurrently—that populate a given soundscape. In this regard, this paper

proposes an automatic classification system based on a deep neural network that is targeted to analyse acoustic frames in real-time and distinguish the events that appear in them—not only on the foreground soundscape but also on the background. Hence, the proposed system has been trained to identify different events that may occur concurrently(referred to as a "multi-label" classifier system). The deep network architecture has been selected so it can meet the computing constraints typically found in the potential application domain of this system (i.e., low-cost WASN [5]). In order to assess the classification performance, real-world data has been collected and annotated.

The remainder of this paper is organized as follows. Section 2 reviews the related work on identification of acoustic events in urban environments. Section 3 describes the real-world data collection and labelling processes that have led to the training and test sets used to assess the classification performance. Section 4 details the proposed multi-label classifier system and its evaluation. Finally, Section 5 concludes the paper.

2 Related work

There is an increasing demand of an automatic monitoring of noise levels in urban areas, especially if this monitoring can give information about the noise source of the measured levels. In this sense, several WASN-based projects are being developed in several parts of the world, mainly adapted to their requirements. There are some projects that do not only concentrate in noise monitoring, but also in air pollution. The IDEA project (Intelligent Distributed Environmental Assessment) [6] analyses and describes both pollutants in several urban areas of Belgium. It integrates a sensor network based on a cloud platform, and it measures noise and air quality [7]. The MESSAGE project, which stands for Mobile Environmental Sensing System Across Grid Environments, [8] not only monitors noise, carbon monoxide, nitrogen dioxide, temperature, and they go further for gathering also real-time humidity and traffic occupancy in the United Kingdom. Also, the MONZA project [9] follow both the idea of monitoring urban noise real-time together with other air pollutants in the Italian city of Monza.

One of the projects that face our challenge in a closer way is Sounds of New York City Project (SONYC), which monitors the city using a low-cost static acoustic sensor network [10]. The goal of this project is to describe the acoustic environment, identifying noise sources, while monitoring noise pollution real-time in a more standard method. It collects longitudinal urban acoustic data, in order to process the audios and have generous sampling to work with acoustic event detection [4].

Another project with a similar conceptual principle is the DYNAMAP project [11], with two pilot WASN deployed in Rome and Milan, in suburban and urban areas respectively. The noise sensors aim to remove any specific audio event but road traffic noise, by means of the Anomalous Noise Event Detector (ANED) [12] to compute an only-road traffic noise map.

Deep learning has been applied to urban audio datasets, with encouraging performance [5, 13]. However, many research studies are limited to datasets which are unrealistic because they are curated from audio libraries rather than urban monitoring, or are single-label annotated, neglecting the simultaneous occurrence of sounds [14]. Recent work suggests that multi-label data can improve performance [15].

3 Real-world dataset

3.1 Recording campaign

To assess the feasibility of a multilabel classifier, the first step is to gather multilabel data. For this purpose, two recording campaigns took place in a metropolitan area (centre of Barcelona, Spain). To have a wider variety of data, each recording campaign took place on a different season of the year. Whereas the first recording campaign was conducted during Autumn 2020 (17 November 2020), the second one took place during Spring 2021 (31 May 2021). It must be considered that during the first recording campaign there were mobility restrictions due to COVID-19 pandemic (mobility was allowed only inside municipalities and only essential workers were allowed to work in-situ), whereas during the second campaign the restrictions were softened (no mobility restrictions at all and some people working properly at their workplace).

To have even more diversity in data, the hours in which the recording campaigns took place were different: whereas the Autumn campaign was recorded from 12:00 to 14:30, the Spring campaign was recorded from 15:30 to 18:00.

The scenario in which we decided to record the acoustic samples was a specific crossroad of the Barcelona city centre: the crossroad between Villarroel Street and Diputació Street (plus code 95M5+H9). This crossroad is located on the *Eixample* area of Barcelona, which is the expansion district of the city. The location was chosen to be able to validate the architecture proposed in [5] in a future work. From now on, these recordings will be referred to as Eixample Dataset.

Figure 1. Recording campaign and Zoom recorder.

Each recording campaign resulted in about 2 hours and 30 minutes of acoustic data. However, due to problems with the batteries of the recorders, the recordings taken on the Spring campaign were fragmented into two audio files (one lasting about 1 hour and the other one lasting about 1 hour and a half).

Four Zoom H5 recorders (see Figure 1) were used to record data: one placed on the middle of each corner of the street intersection. Again, the reason behind this decision is to have simultaneous audio recordings to validate in future work if physical redundancy helps increasing the classification results of the end-to-end system proposed in [5]. Nevertheless, in the machine learning study presented in this work, used data comes from the recording of only one sensor, which aims to give a clear evaluation of performance at the single-device level.

3.2 Data labelling

After the recording campaigns, we manually labelled the data corresponding to one specific corner, in order to maintain the location and recording conditions. This way, data from about 5 hours of recordings (2 hours and a half of each recording campaign) was used for the experimental evaluation. As the idea was to use a classification algorithm like the deep neural network proposed in [5], we decided to directly label the audio files in blocks of 4-seconds as this is the window size selected in [5] as well. Hence, the audio files were split in fragments of that length. As a result, the labels file for the dataset contained the starting and ending time of the 4-seconds window and the multilabels assigned to that fragment.

The manual labelling task led the team to this taxonomy, with the following number of classes:

Table 1 – Number of events labelled on the dataset.

Label	Description	Number of occurrences		
		1st Campaign	2nd Campaign	Total
rtn	Background traffic noise	2177	2118	4295
peop	Noise produced by people	300	612	912
brak	Car brakes	489	424	913
bird	Bird vocalizations	357	960	1317
motorc	Motorcycles	769	565	1334
eng	Engine idling	203	913	1116
cdoor	Car door	133	161	294
impls	Undefined impulsional noises	445	170	615
cmplx	Complex noises that the labeller could not identify	85	73	158
troll	Trolley	162	152	314
wind	Wind	8	23	31
horn	Car or motorbike horn	43	33	76
sire	Sirens from ambulances, the police, etc.	18	57	75
musi	Music	8	30	38
bike	Non-motorized bikes	51	24	75
hdoor	House door	25	60	85
bell	Bells from a church	24	27	51
glass	People throwing glass on the recycling bin	17	32	49
beep	Beeps from trucks during reversing	31	0	31
dog	Dogs barking	3	25	28
drill	Drilling	0	14	14

4 Multilabel classification

4.1 Feature extraction

As features, and to maintain compatibility with [5], a spectrogram was obtained from each 4-second window of the dataset. Audio files were originally recorded at a sampling rate of 44,100 Hz. First, we considered down-sampling the audio files to 22,050 Hz, but after analyzing the labelled events we realized that the *brak* label had all its frequential information at the band of ~17,000 Hz. Considering the Nyquist theorem, if the *brak* event is aimed to be detected, a sampling rate of 22,050 Hz is not high enough. Hence, we finally decided to keep the original 44,100 Hz frequency even if it required more computational resources.

Each spectrogram was calculated generated with an FFT (Fast Fourier Transform) window of 1,024 points and using the *librosa* python library [16]. Next, each spectrogram was individually normalized to have a minimum value of 0 and a maximum value of 1, for compatibility with the input format of the neural network.

4.2 Train/Validation/Test split

The audio files obtained on the recording campaign had to be divided into Train/Validation/Test subsets. As soundscapes have temporal continuity, and so to evaluate the machine learning algorithm correctly, it is important to make sure these three data subsets are taken from different times of day. Therefore, we tried to avoid or mitigate the fact that different audio samples with similar background noise were placed, for example, on both the Training and Testing set.

Concretely, the division was done as shown in Figure 2: with divisions into contiguous regions of 5—71 minutes length.

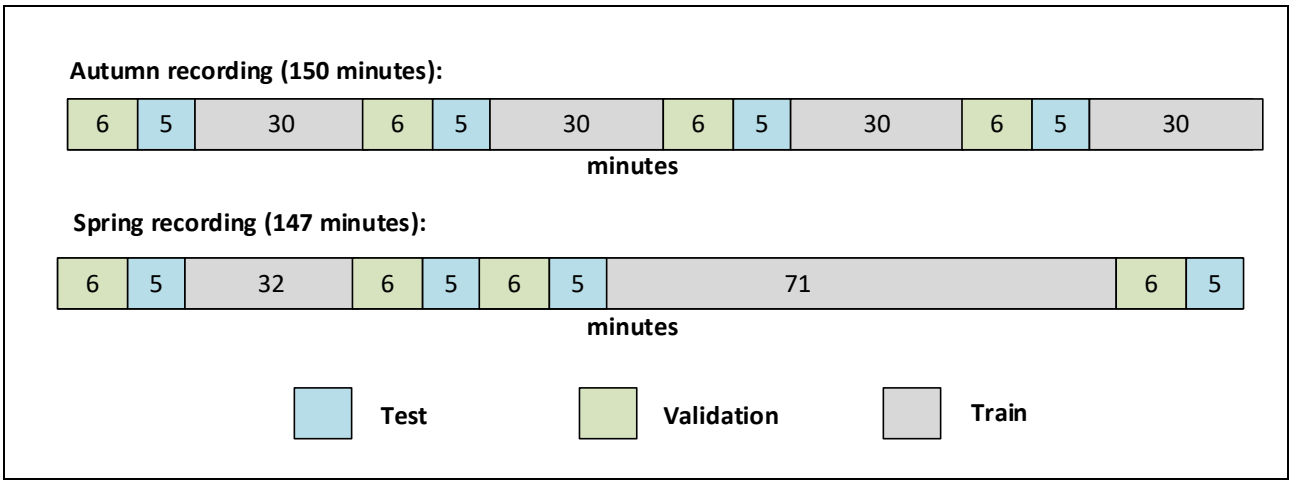

Figure 2 – Train/Validation/Test split of the dataset.

Table 2 – Number of events on the Training/Validation/Testing set.

Label	Dataset		
	Train	Validation	Test
rtn	3029	583	683
peop	954	100	181
brak	627	137	149
bird	913	196	208
motorc	954	183	197
eng	864	73	179
cdoor	190	51	53
impls	457	67	91
cmplx	128	16	14
troll	229	53	32
wind	19	4	8
horn	49	17	10
sire	69	0	6
musi	34	0	4
bike	55	8	12
hdoor	65	12	8
bell	34	4	13
glass	40	6	3
beep	9	13	9
dog	23	4	1
drill	14	0	0

This division left the dataset with 209 minutes for Training, 40 minutes for Validating and 48 minutes for Testing. Note that the division of the two datasets was not exactly even due to the distribution of the events. We tried to maximize the variety of the events on each of the datasets while keeping their temporal evolution.

As it can be appreciated in Table 2, the three datasets are highly unbalanced. Note that due to the lack of *drilling* events during the recording campaigns (only 14 consecutive events) we were unable to test that category. We discarded the option of splitting the 14 events in the Train and Test sets as they belonged to the same drilling machine used in the same location. Also, we decided to remove the *cmplx* sounds from the dataset. As when labelling those sounds, we could not identify their specific source, so the conclusion is that they may confuse the system.

4.3 Data augmentation

To mitigate the potential effects of class imbalance while training, we decided to add more training data and to apply data augmentation techniques to obtain more samples on the poorer classes. Additional data was obtained from the BCNDataset [17], that is a dataset containing real-word urban and leisure events recorded at night in Barcelona. As BCNDataset was labelled differently than the Eixample Dataset, their labels were standardized.

Table 3 – Macro and micro average F-1 scores for the experimental evaluation.

Dataset used	F1- Macro average	F1- Micro average
Experiment 0	12%	46%
Experiment 1	**39%**	**70%**
Experiment 2	36%	75%
Experiment 3	33%	67%

More concretely, on the BCNDataset the labels are provided as [`start_second end_second label`] per each of the acoustic events. To make it compatible with the work presented in this paper, the labels were fragmented and grouped in windows of 4-seconds. This way, we were able to obtain one-hot encoded multilabel labels.

The concrete data augmentation technique used in this work consisted on audio mixing, sometimes known as mixup [18]. That is, two spectrograms (one belonging to the Eixample Dataset and the other belonging to BCNDataset) were added and then divided by two to maintain 0-to-1 normalization values. As the newly generated sample would contain information of all the events tagged in both spectrograms, the labels file was generated by aggregating the one-hot-encoding values as well. This process was carried out using pseudo-random spectrogram selection until all the classes had about 500 samples on the Training set.

4.4 Multilabel classification

The classification of the events was carried out using a deep neural network with a MobileNet v2 architecture [19] with a size of 8.8MB—which should fit on a low-cost computing node for a WASN. The last layer of the classifier was replaced by a fully connected layer with one neuron per class and a Sigmoid activation function on each of them. As a result, for each input data, the output neurons showed the probability of that class being present on the input spectrogram. Once the probabilities were obtained, custom thresholds for each class were applied to determine if the event was actually present on the 4-seconds fragment. The thresholds were obtained by maximizing the F1-measure of each class on the validation set. As hyperparameters, an ADAM optimizer was used with a learning rate of 1e-4 and a weight decay regularization of 1e-5.

We evaluated the effect of training data on performance. Concretely, four experiments were conducted, differing only in the training datasets used:

- **Experiment 0**: We used the Training set of the Eixample Dataset and the entire BCNDataset, without using data augmentation techniques.
- **Experiment 1**: We used the Training set of the Eixample Dataset and the entire BCNDataset using data augmentation techniques to have around 500 samples for each class.
- **Experiment 2**: We used the same data as in Experiment 1 and we added also data from the UrbanSound 8K dataset [14]. The sampling frequency of most of the audio files of the UrbanSound dataset is lower than the one used on the recording campaign (i.e., 44100 Hz). In order to avoid having half of the spectrogram empty for the UrbanSound samples, each audio file was combined with an audio file from Experiment 1 using mix-up aggregation (that is, two spectrograms are aggregated, each of them having a different weight on the

final image). Concretely, the audio files from the UrbanSound 8K dataset have only between a random 10% to 30% on the final weight of the spectrogram.

- **Experiment 3**: We used the same data as in Experiment 2, but we combined the audio files from the UrbanSound 8K dataset 10 times to increase the size of the Training data.

The metrics that we used to compare the results are the macro and micro average F1-scores [20]. Whereas the first metric gives an overall classification result without taking into account the number of samples of each class (i.e., all the classes have the same importance), the second one considers the number of samples of each class of the dataset (i.e., those classes that have a greater number of samples on the Test set have more importance). We present both results as, on the one hand, the macro average could be biased because of the limitations of the Test set in some classes (e.g., there is only one *dog* event, which means that the F1-measure for that class will be binary); and, on the other hand, the micro average could be biased as well as the *rtn* class is present in almost all the audio samples. Hence, whereas the first metric is mostly affected by the performance of the smaller classes of the dataset, the second one is mostly affected by the performance of the larger classes of the dataset. Table 3 shows the classification results for each of the experiments.

Table 4 – Evaluation metrics of the system.

Label	True Negative	False Positive	False Negative	True Positive	F1-score
rtn	0	37	10	673	**0,97**
peop	445	94	65	116	**0,59**
brak	485	86	78	71	**0,46**
bird	473	39	47	161	**0,79**
motorc	397	126	67	130	**0,57**
eng	492	49	44	135	**0,74**
cdoor	655	12	40	13	0,33
impls	527	102	35	56	0,45
troll	651	37	17	15	0,36
wind	693	19	0	8	0,46
horn	703	7	5	5	0,45
sire	697	17	5	1	0,08
musi	698	18	4	0	0
bike	665	43	11	1	0,04
hdoor	667	45	4	4	0,14
bell	707	0	3	10	0,87
glass	696	21	1	2	0,15
beep	708	3	9	0	0
dog	718	1	1	0	0

We think that the data used on Experiment 1 offers the fairest trade-off between the performance of the system on large and small classes. Table 4 shows the individual classification metrics per each class of the dataset based on the results obtained in Experiment 1. As it can be seen, the system has a good performance when classifying events with more than 100 instances on the Validation and Test set (values highlighted in Table 4). However, it behaves poorly when classifying those classes with few instances except for the *bell* event. This may be due to the fact that in the recording

location, the saliency of the recorded bells was higher than the background noise, so all the recorded bells are foreground events. On the contrary, events such as sirens or music were occasionally mixed with background noise depending on the distance between the noise source, the sensor and the simultaneous acoustic events happening at the same time.

5 Conclusion

In this work, progress has been made in the training, testing and validation of deep neural networks algorithms with a very relevant focus on the use of real-world data. The data gathering process has been detailed and the strategies to enrich these data (i.e., data augmentation) to balance the corpus and, thus, improve the performance of the classifier have been shown. Upon the conducted experiments, we foresee that adding a memory layer to the system may increase the classifier performance. That is, we believe that knowing the probability of certain events in certain cases may help. This hypothesis will be further evaluated in future works.

Acknowledgements

We would like to thank Gerard Ginovart for his valuable assistance on the recording campaign in both seasons. Also, they would like to thank Secretaria d'Universitats i Recerca of the Department d'Empresa i Coneixement of the Generalitat de Catalunya for partially funding this work under grants 2017-SGR-966 and 2017-SGR-977.

References

[1] Moudon, A.V. Real noise from the urban environment: How ambient community noise affects health and what can be done about it. *American journal of preventive medicine.* Vol 37 (2), 2009, pp 167–171.

[2] Tsafnat, T., et al. The influence of hearing impairment on sleep quality among workers exposed to harmful noise. *Sleep.* Vol 34 (1), 2011, pp 25-30.

[3] Abbaspour, M., et al. Hierarchal assessment of noise pollution in urban areas–A case study. *Transportation Research Part D: Transport and Environment.* Vol 34, 2015, pp 95-103.

[4] Bello, J.P., Silva, C., Nov, O., et al. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, Vol 62, 2019, pp 68–77.

[5] Vidaña-Vila, E., Navarro, J., Borda-Fortuny, C., Stowell, D., & Alsina-Pagès, R. M. Low-cost distributed acoustic sensor network for real-time urban sound monitoring. *Electronics*, *9* (12), 2020.

[6] Botteldooren, D., De Coensel, B., Oldoni, D., et al. Sound monitoring networks new style. *Acoustics 2011: Breaking New Ground: Proceedings of the Annual Conference of the Australian Acoustical Society*; Queensland, Australia, 2011; pp 93:1–93:5.

[7] Domínguez, F., Dauwe, S., Cuong, N.T., et al. Towards an environmental measurement cloud: Delivering pollution awareness to the public. *International Journal of Distributed Sensor Networks.* Vol 10 (3), 2014, pp 541360.

[8] Bell, M.C., Galatioto, F. Novel wireless pervasive sensor network to improve the understanding of noise in street canyons. *Applied Acoustics*. 2013, pp 169–180.

[9] Bartalucci, C., Borchi, F., Carfagni, M., et al. The smart noise monitoring system implemented in the frame of the Life MONZA project. *Proceedings of EuroNoise 2018*; EAA – HELINA: Heraklion, Crete – Greece. 2018, pp 783–788.

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

[11] Sevillano, X., Socoró, J.C., Alías, F., Bellucciet al. DYNAMAP–Development of low cost sensors networks for real time noise mapping. *Noise mapping*, Vol 1, 2016.

[12] Socoró, J.C., Alías, F., Alsina-Pagès, R.M. An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments. *Sensors,* Vol 17 (10), 2017, pp 2323

[13] Mesaros, A., Heittola, T., Virtanen, T. TUT database for acoustic scene classification and sound event detection. *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*. Budapest, Hungary, 2016, pp. 1128-1132.

[14] Salamon, J., Jacoby, C., Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. *22nd ACM International Conference on Multimedia*, Orlando USA, 2014, pp. 1041-1044.

[15] Gontier, F., Lostanlen, V., Lagrange, M., et al. Polyphonic training set synthesis improves self-supervised urban sound classification. *Journal of the Acoustical Society of America.* Vol 149 (6), 2021, pp. 4309-4326.

[16] McFee, B., Metsai, A., McVicar, M., et al. Audio and music signal analysis in python. *Proceedings of the 14th python in science conference*, 2015.

[17] Vidaña-Vila, E., Duboc, L., Alsina-Pagès, R. M., Polls, F., & Vargas, H. BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset. *Sustainability*, Vol 12 (19), 2020.

[18] Stowell, D., Petrusková, T., Šálek, M., Linhart, P. Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *Journal of the Royal Society Interface*. Vol 16 (153), 2019.

[19] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[20] Mesaros, A., Heittola, T., Virtanen T. Metrics for Polyphonic Sound Event Detection. *Applied Sciences*. Vol 6 (6), 2016, pp 162.