DOI: 10.61782/fa.2023.0422

forum acusticum 2023

ANALYSIS AND ACOUSTIC EVENT CLASSIFICATION OF ENVIRONMENTAL DATA COLLECTED IN *SONS AL BALCÓ* PROJECT

Daniel Bonet-Solà[1] **Ester Vidaña-Vila**[1]* **Rosa Ma Alsina-Pagès**[1]*

[1] Human-Environment Research (HER) — La Salle, Universitat Ramon Llull
Sant Joan de la Salle, 42, 08022 Barcelona — Spain

ABSTRACT

A difficulty encountered in citizen science projects is the processing and analysis of data collected by participants in order to draw conclusions. The project *Sons al Balcó* started with the aim of studying the effect of lockdown due to the COVID-19 pandemic on the perception of noise in Catalonia, asking the citizens to evaluate the soundscape from their homes. In one of the activities of the project, citizens collaborated by sending short videos recorded with a mobile phone, together with a subjective questionnaire about the recorded soundscape on their home balcony or window. Following this purpose, the samples coming from citizens should be automatically analyzed in terms of acoustic event detection, in order to compare the objective data in the videos with the subjective impressions collected in the questionnaires. As a first step towards automatic acoustic event classification, this paper details and compares the acoustic samples of the two collecting campaigns of the project. While the 2020 campaign obtained 365 videos, the 2021 campaign obtained 237. Later, a convolutional neural network has been trained to automatically detect and classify acoustic events even if they occur simultaneously. The findings indicate that the detection rates of different categories are not uniform, with the prevalence percentage of an event in the dataset and its foreground-to-background ratio being important determining factors.

Keywords: *noise annoyance, acoustic event detection, citizen science, convolutional neural networks*

**Corresponding author*: *ester.vidana@salle.url.edu, rosamaria.alsina@salle.url.edu.*

1. INTRODUCTION

Noise pollution has a negative impact on the health and quality of life of millions of people worldwide, particularly in densely populated urban areas. Some of the reported detrimental effects include hypertension, heart diseases [1], diastolic blood pressure [2], sleep disorders [3], psychological stress [4], decreased work performance [5], learning impairment [6], and general annoyance [7]. As a first step to tackle the growing concern for the welfare of the population affected by environmental noise in their daily lives, it is paramount to assess the quality of urban soundscapes. Many municipal administrations map their streets and areas based solely on the LA_{eq} measured. However, not all sources of noise are equally annoying. Therefore, a tool that automatically detects different sound events at specific locations can be useful in determining the level of annoyance at those spots.

During the spring of 2020, the soundscape of many cities around the world changed dramatically due to the COVID-19 pandemic lockdowns, and Catalonia was no exception. TThere was a significant reduction in the urban noise level that was objectively measured by the sensors networks deployed in some of their most prominent cities, such as Barcelona [8] and Girona [9]. In this context, the *Sons al Balcó* project [10] launched its first campaign during the lockdown. Participants were asked to send short 30-second videos recorded with their smartphones or tablets and to answer a questionnaire about

their perception of the soundscape from their balconies. A year later, another campaign was conducted [11] in a post-lockdown, normalized context, enabling the *Sons al Balcó* team to compare both scenarios.

The datasets obtained through both collecting campaigns have been used to test an automatic sound event detection algorithm based on a Convolutional Neural Network (CNN). The main goal is to automatically detect the different types of sound present in each urban location, especially those appearing more frequently as they would probably have a greater impact on the subjective perception of the quality of the soundscape. Other goals include comparing the performance across those two widely different scenarios and searching for possible correlations between the prevalence of the sounds and their foreground or background placement and the detection performance of the system.

2. *SONS AL BALCÓ* CAMPAIGNS

A total of 365 videos were obtained during the 2020 campaign and another 237 during the subsequent 2021 one. The sounds in these videos were manually annotated using a hierarchical taxonomy [12]. The resulting compilations are complex polyphonic datasets with several sounds overlapping.

Only sound categories that appeared in four or more locations in any given campaign were considered for the detection algorithm, and they are listed in Table 1. The table shows the total aggregated duration in seconds (*Dur.(s)*) for each sound category in the corresponding campaign. The total annotated time was 8,302.85 seconds for 2020 and 6,951.57 seconds for 2021. As seen in Table 1, the datasets are highly imbalanced, with some categories having a prevalence of over 50%, while others are almost anecdotal. The percentage of time when each class is foreground placed (*%Fg.*) is presented in the columns. Once again, the foreground-to-background ratio significantly differs across classes. The 2021 dataset is more complex, with more categories detected and a higher polyphonic level (overlapping of events). This is to be expected due to the standing mobility and activity restrictions during the 2020 campaign.

The mean duration of the videos collected was 33.62 seconds for 2020 and 32.44 seconds for 2021. However, there are several outliers, both shorter and longer, as seen in Figure 1.

Table 1. Datasets composition

Campaign	2020		2021	
Class	Dur.(s)	%Fg.	Dur.(s)	%Fg.
Cough			4.2	100%
Steps	18.26	51.6%	107.48	55.4%
Music	111.3	9.2%	264.44	94.6%
Voice	1488.9	42%	1223.6	42%
Construct.	429.3	56%	190.85	95.3%
Industry			150.79	66.7%
Ventilation	350.24	28.9%	405.25	15.6%
Bird	4425.8	50.4%	3241	48.8%
Dog	181.59	68.6%	85.36	80.8%
Water	121.97	21.2%	545.79	78.9%
Wind	653.01	65.3%	559.96	78.6%
Bells	100.69	58.6%	117.83	67.9%
CarHorn			20.4	84.3%
Door	14.89	100%	4.31	100%
ThingsMv.	99.81	62.1%	319.47	93.9%
Rail			94.7	100%
Road	2586.8	56.5%	2713.2	54.9%
Non-motor			13.49	68.4%

3. METHODS AND SETTING

The annotated video files for both campaigns were split into four to implement a 4-fold cross-validation scheme. Each fold followed a 65% train, 10% validation and 25% test distribution. As seen in Table 2, a 30ms Hamming Window with a 50% overlap was used. The features extracted from the clips were GammaTone Cepstral Coefficients (GTCC) which offered optimal performance in a previous work by the authors [13]. Then, the 100 coefficients extracted for each frame were re-formatted into a 10x10 matrix to feed the machine learning (ML) algorithm.

A CNN was implemented using two convolutional layers with a 2x2 kernel followed by two pooling layers. The number of neurons used for the convolutional layers depended on the number of categories of the dataset. When the full dataset was used, 64 and 32 neurons were set. On the contrary, when only the most prevalent sound events were detected, the number of neurons was reduced to 16 and 8 respectively. The last two layers were a flattener layer and dense layer (DL). As stated in Table 2,

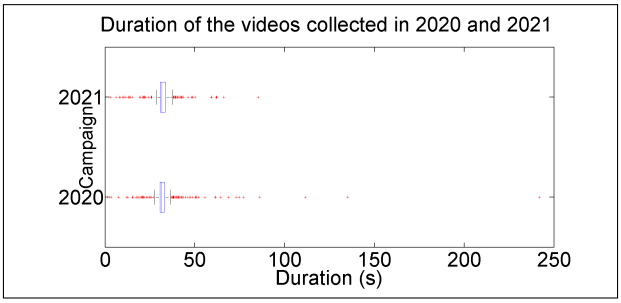

Figure 1. Duration of the video segments in seconds for both campaigns

the activation function used was Rectifier Linear Unit (ReLU), the loss function was binary_crossentropy and the optimizer chosen was Adam with a 0.001 learning rate.

Table 2. Setting for the Features Extraction and Machine Learning algorithms

Framing		Machine Learning	
Samp.Rt.	48 kHz	Agorithm	CNN
Framing	30 ms	Conv. layers	2
Window	Hamming	Kernel size	2x2
Overlap	15 ms	Pool. layers	MaxPool
Features Extraction		Neurons L1	64 or 16
Features	GTCC	Neurons L2	32 or 8
EarQ	9.26	Activ. Funct.	ReLU
BW	24.7	Neurons DL	34
LowFreq.	20 Hz	Optimizer	Adam
HighFrq.	24 kHz	Learn. Rate	0.001
Filters	48	Loss Funct.	binary
Order	4th		crossent.
Features	100	Threshold	0.1-0.5

The metrics used to evaluate the algorithm's performance were Accuracy and F1-Score, which are widely used in the literature. These metrics were adapted to a multi-label scenario [14]. However, due to the highly imbalanced datasets, F1-Score was preferred. It is recommended to report both instance-averaged F1-Score (micro F1-Score) and class-averaged F1-Score (macro F1-Score). The micro F1-Score takes into account the disparate amount of events for each class, meaning that classes with a higher number of samples have a greater impact on the performance. On the other hand, the macro F1-Score does not consider the amount of events for each class, meaning that all classes have the same impact on the performance.

It is also interesting to not only provide event-based metrics of the performance. For the final goal of using the detected sounds for the assessment of the quality of a given soundscape, it can be enough to detect the sounds appearing in a wider segment, regardless of the exact time frames where they occur. In fact, segment-based metrics are more robust to label subjectivity and can be preferable on complex polyphonic contexts. In this present work, the segments considered were the audio files provided by the contributors to the project. They have a duration of approximately 30 seconds with some variability already stated in Figure 1.

As a multilabel classifier was needed, a vector with 34 boolean values has been generated for each frame or segment (depending on the metrics used), one for each possible category in the taxonomy. A value of 1 was assigned to the detected categories in a given frame or segment and a 0 is assigned otherwise. For event-based metrics, the threshold chosen in the computed probabilities to consider that a sound category exists in the selected frame was 0.5. On the contrary, for the segment-based metrics, the mean value of the individual probabilities for all the frames in the segment was computed. In this case, a lower threshold of 0.1 achieved better F1-Scores.

4. RESULTS

4.1 General Performance

Focusing on the event-based metrics, a similar accuracy of over 90% is achieved for both campaigns, as seen in Figure 2. Regarding F1-Score, this proposal achieves instance averaged F1-Scores of 54.75% and 51.37% when all the classes in Table 1 are considered. The class averaged F1-Scores are significantly lower: 16.3% and 16.26%. These results are consistent with other recent works in the literature that deal with similar complex polyphonic datasets. For comparison, a study conducted in 2021 [15] achieved an instance averaged F1-Score of 46% and a class averaged F1-Score of 12% before data augmentation using a similar dataset of 21 outdoor urban sounds classes.

F1-Scores are improved when opting for a segment-

based approach. The micro F1-Scores rise to 57.82% and 53.38% for both campaigns and macro F1-Scores also increase to 23.5% and 21.40% respectively. On the contrary, accuracy experiments a modest decline to approximately 85%.

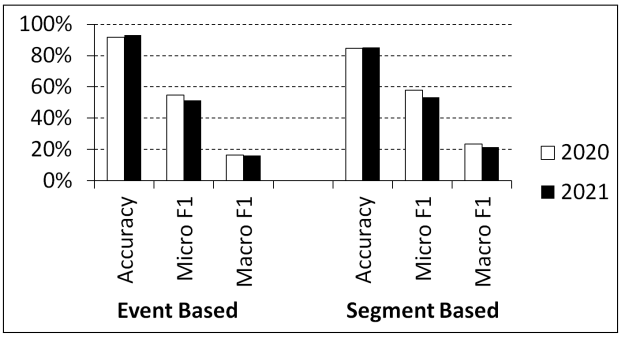

Figure 2. Classification Performance for 2020 and 2021 campaigns

4.2 Performance for the most prevalent sounds

It is interesting to study the performance with a reduced dataset of the most prevalent sounds. In order to assess the quality of a soundscape, the sound classes that appear more frequently are bound to have a greater impact. In Figure 3 only the four most prevalent sounds in both campaigns are considered, i.e., birds, road traffic noise, voice and wind.

F1-Score improves significantly in this subset of categories. Event-based micro F1-Score surpasses 65% for both campaigns. Moreover, macro F1-Score escalates from barely 16% to more than 50.23% in 2020 and 54.09% in 2021. These values are state-of-the-art and consistent with recent publications. The top-ranked work in the DCASE2020 Challenge Task 4 [16], the last year that used event-based metrics for evaluation, achieved 41.7% (prior to data augmentation) in the event-based F1-Score tested on a dataset of 10 categories of indoor sounds.

Segment-based F1-Scores, are even higher, scoring values from 68.47% to 70.45% for the micro F1-Score and from 63.31% to 66.55% for the macro F1-Score. In segment-based evaluation, the performance among different classes is more balanced. Thus, accuracy, micro F1-Score and macro F1-Score score similar values, close to 70% in most cases.

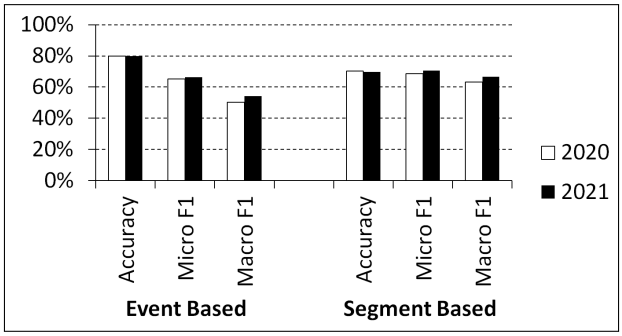

Figure 3. Classification Performance for the four most prevalent sounds during the 2020 and 2021 campaigns

4.3 Effects of the prevalence on the classification performance

There is a correlation between the prevalence of a sound in the dataset and its detection performance (F1-Score) as seen in Figure 4 and Figure 5. All the sound classes with a higher than 20% prevalence achieve event-based F1-Scores over 60% and segment-based F1-Scores over 72%. These percentages are even higher when focusing only on the 2021 campaign. The sound classes with a prevalence between 5% and 20% score lower values with event-based F1-Scores ranging from 10% to 47% and segment-based F1-Scores ranging from 21% to 60% with the sole exception of *ventilation* in the 2021 campaign which presents a very weak performance. Finally, the sound classes with less than a 5% prevalence have generally very poor performances with the notable exception of *rail* that, even though it has a low prevalence of barely 1.36%, it achieves F1-Scores over 70%.

Almost all the categories with more than 5% of prevalence follow a decreasing trend in the performance perfectly correlated with the decline in prevalence. However, *voice* registers a steeper dip than would be expected, probably due to a higher background placement compared to *wind*.

The majority of categories tend to have higher F1-Scores at the segment-based level compared to the event-based level, except for the categories of *construction* and *ventilation* where the differences are negligible. Of particular note is the significant improvement observed in the *voice* category. While it had a modest F1-Score of only just over 10% at the event level, the F1-Score rose to almost 50% at the segment level.

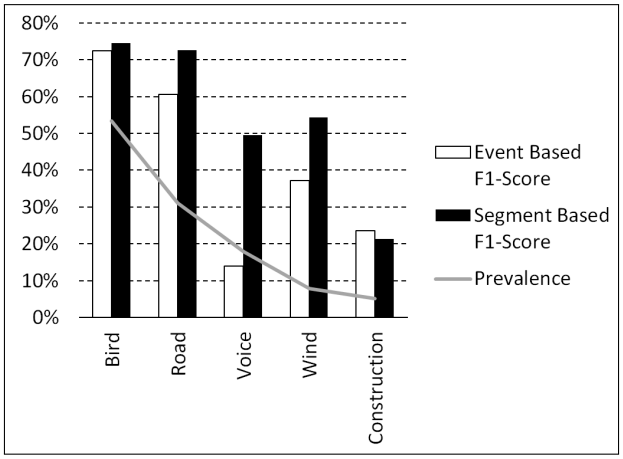

Figure 4. F1-Score for the categories that have a presence higher than 5% of the total time in the 2020 dataset

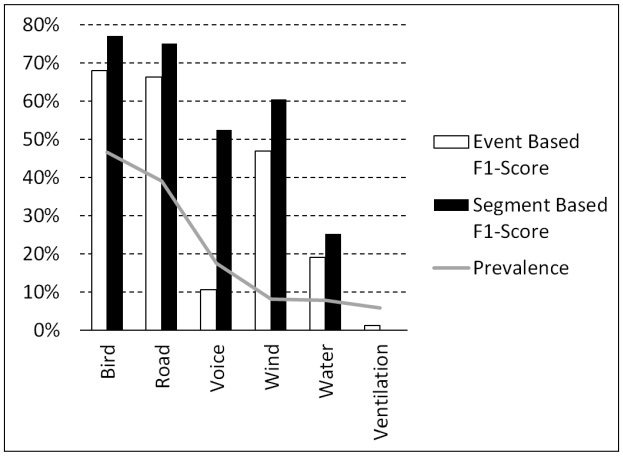

Figure 5. F1-Score for the categories that have a presence higher than 5% of the total time in the 2021 dataset

4.4 Effects of the foreground or background placement on the detection performance

Apart from the prevalence, another feature of the composition of the datasets that affects the performance is the foreground or background placement of the existing sounds. In most cases, a higher foreground-to-background ratio implies a higher F1-score. As shown in Figure 4 and Figure 5, *wind* performs better than *voice* in spite of having lower prevalence in the dataset. Moreover, *road*'s and *bird* score similar performances because the higher prevalence of *bird* related to *road* is partly counteracted by its lower foreground-to-background ratio.

To better state the effects of the sound placement on the performance, a discrete computing of the micro F1-Score has been executed to compare the detecting results for foreground-placed sounds against background-placed sounds in 2020 and 2021. A significant difference can be observed in Figure 6. On the one hand, F1-Scores for background placed sounds are scarcely over 35% for both campaigns. However, they jump to 58% in 2020 and 51% in 2021 for the foreground-placed ones.

5. CONCLUSIONS

A system to automatically detect sounds in urban soundscapes based on a CNN has been successfully implemented. It has been tested with data obtained from a collaborative citizen science project during two campaigns taking place in overly distinct contexts. The first campaign was conducted during the lockdown caused by the COVID-19 pandemic, full of mobility and activity restrictions that affected substantially the noise sources of the urban soundscapes. On the contrary, the second campaign was conducted during a back to the normal situation, without any of the restrictions imposed in 2020. In spite of these two vastly different contexts, the performance of the algorithm is similar for both years, indicating the robustness of the proposed implementation.

A segment-based approach achieves better F1-Scores than an event-based one. However, accuracy is slightly decreased. If the number of different classes in the dataset is high, the class-averaged F1-Score is significantly lower than the instance averaged F1-Score. In datasets with fewer categories and, especially with a segment-based evaluation, differences between both micro and macro F1-Scores are minimized.

The detection performance relies on several features of the dataset. There is a marked correlation between the prevalence and performance of the algorithm. The system improves exceptionally when it is focused on detecting the most prevalent sounds in the soundscapes, offering state-of-the-art scores for both campaigns.

This exceptional performance when detecting prevalent sounds such as *birds* and *road* is encouraging, as these classes are among the most relevant when assessing the quality of an urban soundscape.

Finally, the foreground-to-background ratio of the sound classes also affects the performance, being

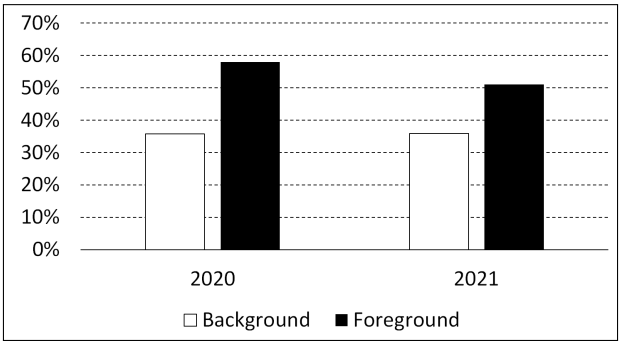

Figure 6. Micro F1-Score improvement for the foreground placed events compared to the background placed ones

the foreground-placed sounds better detected than the background-placed ones.

6. ACKNOWLEDGMENTS

The authors would like to thank all the contributors of both 2020 and 2021 collecting campaigns. Authors would also like to thank Universitat Ramon Llull, under the grants 2020-URL-Proj-054 and 2021-URL-Proj-053 (Rosa Ma Alsina-Pagès) and the Departament de Recerca i Universitats (Generalitat de Catalunya) under Grant Ref. 2021 SGR 01396.

7. REFERENCES

[1] J. Dratva, H. C. Phuleria, M. Foraster, J.-M. Gaspoz, D. Keidel, N. Künzli, L.-J. S. Liu, M. Pons, E. Zemp, M. W. Gerbase, and C. Schindler, "Transportation noise and blood pressure in a population-based sample of adults," *Environmental Health Perspectives*, vol. 120, no. 1, pp. 50–55, 2012.

[2] D. Petri, G. Licitra, M. A. Vigotti, and L. Fredianelli, "Effects of exposure to road, railway, airport and recreational noise on blood pressure and hypertension," *Int. J. Environ. Res. Public Health*, vol. 18, no. 17, 2021.

[3] M. Kohlhuber and G. Bolte, "Einfluss von umweltlärm auf schlafqualität und schlafstörungen und auswirkungen auf die gesundheit," *Somnologie*, vol. 16, pp. 10–16, 2012.

[4] I. van Kamp and H. Davies, "Environmental noise and mental health: Five year review and future directions," in *9th International Congress on Noise as Public Health Problem (ICBEN) - Foxwoods, CT*, 2008.

[5] L. Vukić, V. Mihanović, L. Fredianelli, and V. Plazibat, "Seafarers' perception and attitudes towards noise emission on board ships," *Int. J. Environ. Res. Public Health*, vol. 18, no. 12, p. 6671, 2021.

[6] F. Minichilli, F. Gorini, E. Ascari, F. Bianchi, A. Coi, L. Fredianelli, G. Licitra, F. Manzoli, L. Mezzasalma, and L. Cori, "Annoyance judgment and measurements of environmental noise: A focus on italian secondary schools," *Int. J. Environ. Res. Public Health*, vol. 15, no. 2, 2018.

[7] H. M. Miedema and C. Oudshoom, "Annoyance from transportation noise: relationships with exposure metrics dnl and denl and their confidence intervals," *Environmental Health Perspectives*, vol. 109, no. 4, pp. 409–416, 2001.

[8] D. Bonet-Solà, C. Martínez-Suquía, R. M. Alsina-Pagès, and P. Bergadà, "The soundscape of the covid-19 lockdown: Barcelona noise monitoring network case study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, 2021.

[9] R. M. Alsina-Pagès, P. Bergadà, and C. Martínez-Suquía, "Changes in the soundscape of girona during the covid lockdown," *Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3416–3423, 2021.

[10] R. M. Alsina-Pagès, F. Orga, R. Mallol, M. Freixes, X. Baño, and M. Foraster, "Sons al balcó: Soundscape Map of the Confinement in Catalonia," in *Engineering Proc.*, vol. 2, p. 77, 2020.

[11] X. Baño, P. Bergadà, D. Bonet-Solà, A. Egea, M. Foraster, M. Freixes, G. J. Ginovart-Panisello, R. Mallol, X. Martín, A. Martínez, *et al.*, "Sons al balcó, a citizen science approach to map the soundscape of catalonia," *Engineering Proc.*, vol. 10, no. 1, p. 54, 2021.

[12] D. Bonet-Solà, E. Vidaña-Vila, and R. M. Alsina-Pagès, "Analysis and acoustic event classification of environmental data collected in a citizen science project," *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, 2023.

[13] D. Bonet-Solà and R. M. Alsina-Pagès, "A comparative survey of feature extraction and machine learning

methods in diverse acoustic environments," *Sensors*, vol. 21, no. 4, 2021.

[14] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.

[15] E. Vidaña-Vila, J. Navarro, D. Stowell, and R. M. Alsina-Pagès, "Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors," *Sensors*, vol. 21, no. 22, 2021.

[16] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution-augmented transformer for semi-supervised sound event detection," *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2020.