Real-time acoustic event classification in urban environments using low-cost devices

Ester Vidaña Vila

UNIVERSITAT RAMON LLULL

DOCTORAL THESIS

Title	Real-time acoustic event classification in urban environments using low-cost devices
Presented by	Ester Vidaña Vila
Centre	La Salle International School of Commerce and Digital Economy
Department	Engineering
Directed by	Dr. Rosa Ma. Alsina-Pagès

It is a mistake to think that the practice of my art has become easy to me. I assure you, dear friend, no one has given so much care to the study of composition as I. There is scarcely a famous master in music whose works I have not frequently and diligently studied.

— Wolfgang Amadeus Mozart

Dedicada als meus pares Ana i Andreu.

Abstract

In the modern and ever-evolving society, the presence of noise has become a daily threat to a worrying amount of the population. Being overexposed to high levels of noise may interfere with day-to-day activities and, thus, could potentially bring severe side-effects in terms of health such as annoyance, cognitive impairment in children or cardiovascular diseases. Some studies point out that it is not only the level of noise that matters but also the type of sound that the citizens are exposed to. That is, not all the acoustic events have the same impact on the population.

With current technologies used to track noise levels, for both private and public administrations, it is hard to automatically identify which sounds are more present in most polluted areas. Actually, to assess citizen complaints, technicians are typically sent to the area to be surveyed to evaluate if the complaint is relevant. Due to the high number of complaints that are generated every day (specially in highly populated areas), the development of Wireless Acoustic Sensor Networks (WASN) that would automatically monitor the noise pollution of a certain area have become a research trend. Currently, most of the networks that are deployed in cities measure only the equivalent noise level by means of expensive but highly accurate hardware but cannot identify the noise sources that are present in each spot. Given the elevated price of these sensors, nodes are typically placed in specific locations, but do not monitor wide areas.

The purpose of this thesis is to address an important challenge still latent in this field: to acoustically monitor large-scale areas in real-time and in a scalable and cost efficient way. In this regard, the city centre of Barcelona has been selected as a reference use-case scenario to conduct this research. First, this dissertation starts with an accurate analysis of an annotated dataset of 6 hours corresponding to the soundscape of a specific area of the city (*l'Eixample*). Next, a scalable distributed architecture using low-cost computing devices to recognize acoustic events is presented. To validate the feasibility of this approach, a deep learning algorithm running on top of this architecture has been implemented to classify 10 different acoustic categories. As the sensing nodes of the proposed system are arranged in such a way that it is possible to take advantage of physical redundancy (that is, more than one node may hear the same acoustic event), data has been gathered in four spots of the city centre of Barcelona respecting the sensors topology. Finally, as real-world events tend to occur simultaneously, the deep learning algorithm has been enhanced to support multilabel (i.e., polyphonic) classification. Results show that, with the proposed system architecture, it is possible to classify acoustic events in real-time. Overall, the contributions of this research are the following: (1) the design of a low-cost, scalable WASN able to monitor large-scale areas and (2) the development of a real-time classification algorithm able to run over the designed sensing nodes.

Keywords: Acoustic Event Detection, Urban Noise, Wireless Acoustic Sensor Network, Real-time Classification, Polyphonic Event Classification.

Ester Vidaña Vila

Barcelona, February 2022

Resumen

En la sociedad moderna y en constante evolución, la presencia de ruido se ha convertido en una amenaza diaria para una cantidad preocupante de la población. Estar sobreexpuesto a altos niveles de ruido puede interferir en las actividades cotidianas y, por tanto, podría acarrear graves efectos secundarios en términos de salud como mal humor, deterioro cognitivo en niños o enfermedades cardiovasculares. Hay estudios que señalan que no solo afecta el nivel de ruido al que están expuestos los ciudadanos, sino que también es importante el tipo de sonido. Es decir, no todos los eventos acústicos tienen el mismo impacto en la población.

Con las tecnologías que se utilizan actualmente para la monitorización de la contaminación acústica, es difícil identificar automáticamente qué sonidos están más presentes en las zonas más contaminadas. De hecho, para evaluar las quejas de los ciudadanos, normalmente se envían técnicos a la zona donde se ha realizado la queja para evaluar si ésta es relevante. Debido al elevado número de quejas que se generan a diario (especialmente en zonas muy pobladas), el desarrollo de Redes de Sensores Acústicos Inalámbricos (WASN) que monitoricen automáticamente la contaminación acústica de una determinada zona se ha convertido en una tendencia de investigación. En la actualidad, la mayoría de las redes desplegadas en entornos urbanos solo miden el nivel de ruido equivalente mediante un equipos caros pero muy precisos, pero no son capaces de identificar las fuentes de ruido presentes en cada lugar. Dado el elevado precio de estos sensores, los nodos suelen colocarse en lugares estratégicos, pero no monitorizan zonas amplias.

El objetivo de esta tesis es abordar un importante reto aún latente en este campo: monitorizar acústicamente zonas de gran tamaño en tiempo real y de forma escalable y económica. En este sentido, se ha seleccionado el centro de la ciudad de Barcelona como caso de uso de referencia para llevar a cabo esta investigación. En primer lugar, esta tesis parte de un análisis preciso de un conjunto de 6 horas de datos anotados correspondientes al paisaje sonoro de una zona concreta de la ciudad (*l'Eixample*). A continuación, se presenta una arquitectura distribuida escalable que utiliza dispositivos de bajo coste para reconocer eventos acústicos. Para validar la viabilidad de este enfoque, se ha implementado un algoritmo de aprendizaje profundo que se ejecuta sobre esta arquitectura para clasificar 10 categorías acústicas diferentes. Como los nodos del sistema propuesto están dispuestos en una topología con redundancia física (es decir, más de un nodo puede escuchar el mismo evento acústico a la vez), se han recogido datos en cuatro puntos del centro de la ciudad de Barcelona respetando la arquitectura de los sensores. Por último, dado que los eventos del mundo real tienden a producirse de forma simultánea, se ha mejorado el algoritmo de aprendizaje profundo para que soporte la clasificación multietiqueta (es decir, polifónica). Los resultados muestran que, con la arquitectura del sistema propuesto, es posible clasificar eventos acústicos en tiempo real. En general, las contribuciones de esta investigación son las siguientes (1) el diseño de una

WASN de bajo coste y escalable, capaz de monitorizar áreas a gran escala y (2) el desarrollo de un algoritmo de clasificación en tiempo real ejecutado sobre los nodos de detección diseñados.

Palabras clave: Detección de Eventos Acústicos, Ruido Urbano, Red de Sensores Acústicos Inalámbricos, Clasificación en Tiempo Real, Clasificación de Eventos Polifónicos.

Ester Vidaña Vila
Barcelona, February 2022

Resum

En la societat moderna i en constant evolució, la presència de soroll s'ha convertit en un perill diari per a una quantitat preocupant de la població. Estar sobreexposats a alts nivells de soroll pot interferir en activitats quotidianes i, per tant, podria causar greus efectes secundaris en termes de salut com mal humor, deteriorament cognitiu en nens o malalties cardiovasculars. Hi ha estudis que assenyalen que no només afecta el nivell de soroll al qual estan exposats els ciutadans, sinó que també és important el tipus de so. És a dir, no tots els esdeveniments acústics tenen el mateix impacte en la població.

Amb les tecnologies que es fan servir actualment per a monitorar la contaminació acústica, és difícil identificar automàticament quins sorolls estan més presents en les zones més contaminades. De fet, per avaluar les queixes dels ciutadans, normalment s'envien tècnics a la zona on s'hi ha produït la queixa per avaluar si aquesta és rellevant. A causa de l'elevat nombre de queixes que es generen diàriament (especialment en zones molt poblades), el desenvolupament de Xarxes de Sensors Acústics Sense Fils (WASN) que monitorin automàticament la contaminació acústica d'una determinada zona s'ha convertit en una tendència d'investigació. En l'actualitat, la majoria de les xarxes desplegades en entorns urbans només mesuren el nivell de soroll equivalent fent servir equipaments cars, però molt precisos, però no permeten d'identificar les fonts de soroll presents a cada lloc. Donat l'elevat cost d'aquests sensors, els nodes solen col · locar-se en llocs estratègics, però no monitoren zones àmplies.

L'objectiu d'aquesta tesi és abordar un important repte que encara està latent en aquest camp: monitorar acústicament zones de gran envergadura en temps real i de forma escalable i econòmica. En aquest sentit, s'ha seleccionat el centre de la ciutat de Barcelona com a cas d'ús de referència per a dur a terme aquesta investigació. En primer lloc, aquesta tesi parteix d'una anàlisi precís d'un conjunt de 6 hores de dades anotades corresponents al paisatge sonor d'una zona concreta de la ciutat (*l'Eixample*). A continuació, es presenta una arquitectura distribuïda escalable que fa servir dispositius de baix cost per a reconèixer esdeveniments acústics. Per a validar la viabilitat d'aquest enfocament, s'ha implementat un algorisme d'aprenentatge profund que s'executa sobre aquesta arquitectura per a classificar 10 categories acústiques diferents. Com que els nodes del sistema proposats estan disposats en una topologia amb redundància física (és a dir, que més d'un node pot escoltar el mateix esdeveniment acústic simultàniament), s'han recollit dades en quatre punts del centre de la ciutat de Barcelona respectant l'arquitectura dels sensors. Per últim, donat que els esdeveniments del món real tendeixen a produir-se de forma simultània, s'ha millorat l'algorisme d'aprenentatge profund perquè suporti la classificació multietiqueta (és a dir, polifònica). Els resultats mostren que, amb l'arquitectura del sistema proposat, és possible classificat esdeveniments acústic en temps real. En general, les contribucions d'aquesta investigació són les següents: (1) el

disseny d'una WASN de baix cost i escalable, que pugui monitorar àrees a gran escala i (2) el desenvolupament d'un algorisme de classificació en temps real executat sobre els nodes de detecció dissenyats.

Paraules clau: Detecció d'Esdeveniments Acústics, Xarxa de Sensors Acústics Sense Fils, Classificació en Temps Real, Classificació d'Esdeveniments Polifònics.

Ester Vidaña Vila

Barcelona, February 2022

Acnowledgements

M'agradaria agrair profundament a totes les persones brillants que m'han ajudat i guiat durant aquest viatge. En primer lloc, mai tindré prou paraules de gratitud cap als meus pares, que sempre m'ho han donat tot i han sigut el millor model a seguir que podria haver tingut. Evidentment, també mereixen tota la meva gratitud la resta de la meva família, el meu germà Àlex, els meus avis (en especial els que ja no hi són, però, d'alguna manera, encara estan amb nosaltres), i tots els meus tiets i cosins.

M'agradaria continuar agraint al Joan tot el suport, l'ajuda i la paciència que ha tingut amb mi durant tots aquests anys, tan personal com professionalment. Sense tu, res d'això hauria estat possible. Literalment. Recordo com si fos ahir mateix quan, sis anys enrere, ens vas convocar a la Rosa i a mi a una saleta de reunions i em vaig iniciar en el món de la recerca.

Rosa, no has sigut només la meva directora, has sigut molt més que això. Moltes gràcies per totes les oportunitats que m'has donat, els teus consells i la teva ajuda 365/24/7. Ha sigut un honor poder treballar sota el teu lideratge.

A continuació, m'agradaria agrair a tots els companys de departament, que fan que l'ambient de treball sigui increïble. Gràcies a tots els membres del GTM, en especial al Xuti per acollir-me al grup des del primer moment. I gràcies als membres del meu (quasi) segon grup: el GRITS. Gracias Agustín por estar siempre abierto a compartir vuestra infraestrutura cuando la he necesitado. I als companys que m'heu donat consells de valor incalculable per poder acabar la tesi: Selene, Gerard, Marc, Oriol, Xavi Sevillano, Roger, Letícia, Cris... I molts més!

També, gràcies als companys de docència, que heu fet que donar classe sigui infinitament més divertit. Gràcies Carme, Nacho, Alejandro, Gonçal, Xavi Solé, i tots els equips de monitors de pràctiques que hem tingut aquests últims anys. Gracias Lisa por todas tus correcciones, por tu paciencia y por no tener nunca un *no* por respuesta. I agraïr també al grup d'aventures de divendres nit que van fer que el confinament fos (una miqueta) més lleu: Edus, Pol, Víctor, Adrià, Alan i Marta.

Finally, big thanks to Dan Stowell and his mini-group from QMUL. Thank you for letting me join the group and for all your help. It has been a pleasure to work under your superivision.

Curriculum Vitae

Education

- **PhD in Communication and Information Technologies and their Application to Management, Architecture and Geophysics.** 2018–present. La Salle, Universitat Ramon Llull. Research visits:
  - Queen Mary University, Centre for Digital Music (C4DM) research group, under the supervision of Dr. Dan Stowell (June - July 2020).
  - Tilburg University, Department of Cognitive Science and Artificial Intelligence, under the supervision of Dr. Dan Stowell (June - July 2021).

  Dissertation title: *"Acoustic event detection and classification in urban environments using low-cost devices"*. Supervised by Dr. Rosa Ma Alsina-Pagès.
- **Masters Degree in Telecommunications.** 2017–2018. La Salle, Universitat Ramon Llull. Thesis title: *"Scalable system for real-time acoustic monitoring of partially dependent people in distributed environments"*. (A with Honors). Supervised by Dr. Rosa Ma Alsina-Pagès and Dr. Joan Navarro.
- **Telecommunications Engineering Degree, majoring in Audiovisual Systems.** 2012–2017. La Salle, Universitat Ramon Llull. Thesis title: *"An authomatic audio classifier for Picidae bird species"*. (A with Honors). Supervised by Dr. Rosa Ma Alsina-Pagès and Dr. Joan Navarro.

Participation in research projects

- Researcher in the **Cow-talk Pro** project. Public reference number: SNEO-20211301. Internal reference number: D-PROJ-52710. (2021–present).
- Researcher in the **Andorra Eco Urban Lab** project. Internal reference number: D-PROJ-52709. (2021–present).
- Researcher in the **Sons de Sabadell** project. Internal reference number: D-PROJ-52707. (2021–present).
- Researcher in the **EMMA** (Environmental Monitoring and Measurement Application) project. Public reference number: ACE014/20/000044. Internal reference number: D-PROJ-52704. (2021–present).

- Researcher in the **Sons a l'aeroport** project, spin-off of the Sons al balcó project. Internal reference number: D-PROJ-52103. (2021–2021).
- Researcher in the **SUARAMAP** project. Public reference number: IDI-20200768. Internal reference number: D-PROJ-52703. (2020–present).
- Researcher in the **DLANED** project. Internal reference number: 2020-URL-Proj-053. (2020–2021).
- Intern in the **DYNAMAP** (Dynamic Acoustic Mapping). Public reference number: LIFE13 ENV/IT/001254. (2017–2018).
- Researcher in the **HomeSound** Project. (2017–2018).
- Intern in the **LSMaker** Project (2016–2018).

Work experience

- **Associate lecturer** at La Salle, Universitat Ramon Llull (2017 – present).
- **Member** of the Grup de Recerca in Media Technologies (GTM) at La Salle, Universitat Ramon Llull (2018 – present).

Publications

Journals

- Vidaña-Vila, E., Navarro, J., Stowell,D., Alsina-Pagès, R.M. Multilabel Acoustic Event Classification Using Real-World Urban Data and Physical Redundancy of Sensors. Sensors 2021, (Q1) (JCR IF = 3.576 2020). Citations in SCOPUS (December 2021): 0 Citations in JCR (December 2021): 0.
- Vidaña-Vila, E., Navarro, J., Borda-Fortuny, et al. Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring. Electronics 2020, (Q3) (JCR IF = 2.397 2020). Citations in SCOPUS (December 2021): 5 Citations in JCR (December 2021): 3.
- Vidaña-Vila, E., Navarro, J., Alsina-Pagès, R.M, Ramírez, A. A two-stage approach to automatically detect and classify woodpecker (Fam. Picidae) sounds, Applied Acoustics 2020, (Q2) (JCR IF = 2,639 2020) Citations in SCOPUS (December 2021): 5 Citations in JCR (December 2021): 4.
- Vidaña-Vila, E., Luboc, L., Alsina-Pagès, R.M, et al. BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset, Sustainability 2020, (Q2) (JCR IF = 3.251) Citations in SCOPUS (December 2021): 4 Citations in JCR (December 2021): 3.

- Navarro, J., Vidaña-Vila E., Alsina-Pagès R.M., Hervás M. Real-Time Distributed Architecture for Remote Acoustic Elderly Monitoring in Residential-Scale Ambient Assisted Living Scenarios. Sensors 2018, 18, 2492.(Q1) (JCR IF = 3.031). Citations in SCOPUS (December 2021): 16 Citations in JCR (December 2021): 14.
- Vidaña-Vila, E.; Navarro, J.; Alsina-Pagès, R.M. Towards Automatic Bird Detection: An Annotated and Segmented Acoustic Dataset of Seven Picidae Species. Data 2017, 2, 18. Citations in SCOPUS (December 2021): 5 Citations in JCR (December 2021): 5.

Conferences

- Vidaña-Vila, E., Navarro, J., Stowell,D., Alsina-Pagès, R.M. Multilabel acoustic event classification for urban sound monitoring at a traffic intersection, Poster presented in: Deep Learning Barcelona Symposium 2021.
- Blanch, J., Vidaña-Vila, E., Alsina-Pagès, R.M. Analysis of the Noise Impact of the Airport of Barcelona to the Llobregat Delta Natural Environment during the 2021 Lockdown period, 8th Electronic Conference on Sensors and Applications.
- Vidaña-Vila, E., Alsina-Pagès, R.M, Navarro, J. Improving classification accuracy of acoustic real-world urban data using sensors physical redundancy, IEEE International Workshop on Distributed and Intelligent Systems DistInSys 2021.
- Vidaña-Vila, E., Alsina-Pagès, R.M, Navarro, J. Prototyping a low-cost wireless acoustic sensor network with physical redundancy to automatically classify acoustic events in urban environments, UrbanSound Symposium 2021.
- Ginovart, G., Vidaña-Vila, E., Caro S., et al. Low-Cost WASN for Real-Time Soundmap Generation. Proceedings of the 8th International Symposium on Sensor Science 2021.
- Vidaña-Vila, E., Navarro, J., La evaluación como refuerzo positivo en una *flipped classroom*. IV Simposi sobre innovació docent i noves tecnologies 2019. ISBN: 978-84-946960-2-3.
- Navarro J., Amo D., Canaleta X., Vidaña-Vila E., Martínez C. Utilizando Analítica del Aprendizaje en una Clase Invertida: Experiencia de Uso en la Asignatura de Sistemas Digitales y Microprocesadores. III Simposi sobre innovació pedagògica i noves tecnologies 2018, ISBN: 978-84-946969-3-0
- Hervás M., Alsina-Pagès R.M., Vidaña-Vila E., et al. LSMaker 2.0: An improved Educational Robot based on previous academic experiences. Jornada d'intercanvi d'experiències didàctiques 2017, ISBN: 978-84-697-4182-5.
- Navarro, J., Amo, D., Canaleta, X., Vidaña-Vila, E., Martínez, C. (2018). Utilizando analítica del aprendizaje en una clase invertida: Experiencia de uso en la asignatura de Sistemas Digitales y Microprocesadores. Actas De Las Jornadas Sobre Enseñanza Universitaria De La Informática, 3.

Índex / Contents

Llistat de Figures / List of Figures

Llistat de Taules / List of Tables

Llistat d'acrònims / List of Abbreviations

ADC Analogue-to-Digital Converter o Convertidor Analògic-Digital. 171, 181

ANN Artificial Neural Network o Xarxa Neuronal Artificial. 43, 44, 62, 63

ASC Acoustic Scene Classification o Classificació d'Escena Acústica. 44, 62, 63

CNN Convolutional Neural Network o Xarxa Neuronal Convolucional. 41, 44–46, 59, 63, 64, 170, 180, 181

CPU Central Processing Unit o Unitat de Processament Central. 46, 64

DALYs Disability-Adjusted Life Years o Anys de Vida Ajustats per Discapacitats. 1, 2, 19, 20

dBA Decibels with A-weighting o Decibels amb Ponderació A. 1, 19

DCASE Detection and Classification of Acoustic Scenes and Events o Detecció i Classificació d'Escenes i Esdeveniments Acústics. 49, 67

DL Deep Learning o Aprenentatge Profund. 7, 11, 14, 17, 25, 29, 33, 35, 40, 41, 44, 45, 48, 58, 59, 62–64, 67, 169, 170, 173, 174, 179, 180, 183, 184

DNN Deep Neural Network o Xarxa Neuronal Profunda. 14, 16, 32, 35, 44, 46, 63, 64, 171, 173, 181, 184

DT Decision Tree o Arbre de Decisió. 43, 62

EBD Estimated Burden of Disease o Càrrega Estimada de les Malalties. 1, 2, 19, 20

CE1 Criteris d'Exclusió 1. 38, 56

CE2 Criteris d'Exclusió 2. 38, 56

CE3 Criteris d'Exclusió 3. 38, 56

CE4 Criteris d'Exclusió 4. 38, 56

FBC Frequency Bank Coefficients o Coeficients de Bancs de Freqüències. 42, 60

GAN Generative Adversarial Network o Xarxa Generativa Antagònica. 175, 186

GMM Gaussian Model Mixture. 40, 43, 58, 61

GPU Graphics Processing Unit o Unitat de Processament de Gràfics. 11, 30, 41, 59

GTCC GammaTone Cepstral Coefficients o Coeficients Cepstrals de Tons Gamma. 43, 61

HMM Hidden Markov Models o Models Ocults de Markov. 40, 42, 58, 61

IC1 Inclusion Criteria 1 o Criteris d'Inclusió 1. 38, 56

IC2 Inclusion Criteria 2 o Criteris d'Inclusió 2. 38, 56

IC3 Inclusion Criteria 3 o Criteris d'Inclusió 3. 38, 56

IC4 Inclusion Criteria 4 o Criteris d'Inclusió 4. 38, 56

IoT Internet of Things o Internet de les Coses. 45, 63

IR Intermittency Ratio o Ratio d'Intermitència. 15, 34

KNN K-Nearest Neighbors o K-Veïns Propers. 42, 60

LPCC Linear Prediction Cepstrum Coefficients o Coeficients Cepstrals de Predicció Lineal. 42, 60

MEMS Micro Electro Mechanical System o Micro Electret Sistema Mecànic. 47, 48, 65, 66, 171, 176, 181, 186

MFCC Mel Frequency Cepstral Coefficients o Coeficients Cepstrals de Freqüència Mel. 42–46, 59–63, 65

ML Machine Learning o Aprenentatge Automàtic. 7, 14, 25, 33, 40–44, 48, 58, 59, 61, 62, 67, 169, 171, 179, 181

MSC Mel-Spectral Coefficients o Coefficients Espectrals Mel. 42, 60

NB Naive Bayes. 43, 62

OS Operating System o Sistema Operatiu. 171, 181

PCA Principal Component Analysis o Anàlisi de Components Principals. 44, 63

PCB Printed Circuit Board o Placa de Circuit Imprès. 12, 30

PCEN Per-Channel Energy Normalization o Normalització d'Energia Per Canal. 45, 64, 170, 180

PICOC Population, Intervention, Comparison, Outcome, Context o Població, Intervenció, Comparació, Resultats, Context. 37, 55

RF Random Forest o Bosc Aleatòri. 43, 62

RQ1 Research Question 1 o Pregunta de Recerca 1. 7, 8, 25, 26, 37, 55, 173, 183

RQ2 Research Question 2 o Pregunta de Recerca 2. 7, 25, 37, 55, 173, 184

RQ3 Research Question 3 o Pregunta de Recerca 3. 7, 8, 25, 26, 37, 55, 174, 184

SMO Sequential Minimal Optimization o Optimització Mínima Seqüencial. 43, 61

SVM Support Vector Machine o Màquina de Vectors de Suport. 40, 42, 43, 58, 60, 61

TO1 Thesis Objective 1 o Objectiu de Tesi 1. 7, 9, 11, 25, 27, 29, 170, 180

TO2 Thesis Objective 2 o Objectiu de Tesi 2. 7, 14, 25, 33, 171, 181

TO3 Thesis Objective 3 o Objectiu de Tesi 3. 8, 9, 14, 26, 27, 33, 171, 172, 181, 182

TO4 Thesis Objective 4 o Objectiu de Tesi 4. 8, 11, 14, 26, 29, 33, 172–174, 182–184

TRL Technology Readiness Level o Nivell de Maduresa Tecnològica. 8, 26

VAE Variational AutoEncoder o AutoCodificador Variacional. 175, 186

WASN Wireless Acoustic Sensor Network o Xarxa de Sensors Acústics sense Fils. 2, 7, 12, 20, 24, 30, 40, 43, 46, 47, 58, 61, 64–66, 169, 171, 174, 177, 179, 181, 184, 187, 189, 191

WHO World Health Organization o Organització Mundial de la Salut. 1, 19

WOS Web Of Science. xxv, 37, 38, 55, 56

ZCR Zero Crossing Rate o Taxa de Creuament per Zero. 42, 60

Capítol 1

Introducció

1.1 Context i Motivació

La primera entrada de soroll al diccionari, el defineix com "un so, especialment un so alt i desagradable" (University 2021). A la societat moderna i en constant evolució en la que vivim, la presència de soroll s'ha convertit en una amenaça diària per a una quantitat preocupant de la població (WHO 2011). No obstant això, no són només els humans els que es veuen afectats pel soroll: alguns estudis han demostrat que estar exposats al soroll causat pels humans té un impacte negatiu en la vida animal, causant migracions no naturals, problemes reproductius i, fins i tot, amenaces de supervivència de l'espècie a llarg termini (Radle 2007).

En el cas concret de la població humana, un conjunt d'estudis (WHO 2011) duts a terme per la World Health Organization o Organització Mundial de la Salut (WHO) confirmen que estar exposat a nivells excessius de soroll interfereix negativament amb les activitats del dia a dia, com ara treballar, assistir a l'escola o descansar durant el temps lliure. A més, els mateixos estudis calculen el Estimated Burden of Disease o Càrrega Estimada de les Malalties (EBD) de la població causada pel soroll ambiental segmentat en el soroll de trànsit, el soroll dels avions i el soroll ferroviari. Aquesta estimació té com a objectiu quantificar els anys potencials de vida que una persona pot perdre per una mort prematura, a més dels anys de vida saludable perduts per males condicions sanitàries o discapacitats. Aquesta estimació es mesura en unitats de Disability-Adjusted Life Years o Anys de Vida Ajustats per Discapacitats (DALYs). Els principals efectes secundaris de la sobreexposició al soroll són:

- **Risc cardiovascular**: El risc de patir una afecció cardíaca (fins i tot una cardiopatia isquèmica) s'incrementa al estar sobreexposat tant al soroll de trànsit com al soroll dels avions. A més, aquests dos tipus de sorolls estan correlacionats amb un increment del risc de patir una alta pressió arterial anòmala. Concretament, s'estima que l'EBD per a malalties cardiovasculars en països europeus d'alta renda és de 61 000 anys.

- **Deficiència cognitiva en joves**: S'ha analitzat mitjanant estudis experimentals i epidemiològics. Aquest deteriorament es produeix mentre els nens estan exposats al soroll i persisteix durant algun temps després que el soroll acabi. Concretament, per dur a terme els estudis, es va observar que, mentre que tots els nens que estaven exposats a un nivell de 95 Decibels with A-weighting o Decibels amb Ponderació A (dBA) es van veure afectats cognitivament, cap nen es va veure afectat a un nivell de 59 dBA. La EBD per als països europeus és de 45 000 anys per als joves d'entre 7 i 19 anys.

- **Pertorbacions del son**: Per calcular el EBD causat per la pertorbació del son en

la població, es van tenir en compte dos tipus d'estudis. En primer lloc, mesures electrofisiològiques. En segon lloc, autoinformes realitzats mitjançant enquestes en diferents estudis. La EBD per als ciutadans europeus que viuen en ciutats amb una població superior a 50 000 habitants és de 903 000 DALYs perduts a causa de la pertorbació del son causada pel soroll.

- **Tinnitus**: Tinnitus es pot definir com la sensació d'escoltar un so concret quan aquest so no està passant realment (per exemple, un clic o brunzit). Aquest efecte pot derivar en altres patologies com la pertorbació del son, la frustració o l'ansietat. L' EBD per a ciutadans europeus adults és de 22 000 anys a causa de Tinnitus causat per estar sobreexposat al soroll.

- ***Annoyance* o Molèsties**: Per mesurar la molèstia causada per entorns sorollosos, s'utilitzen qüestionaris personals. L' EBD per als ciutadans que viuen en ciutats amb una població superior a 50 000 habitants és de 587 000 DALYs perduts a causa de molèsties causades pel soroll.

A més, a part de tots aquests efectes secundaris, s'ha estudiat que el que importa no és només el nivell de soroll, sinó també el tipus de so al qual estan exposats els ciutadans. És a dir, no tots els esdeveniments acústics tenen el mateix impacte en la població (Abbaspour et al. 2015). No obstant això, quan les administracions públiques o privades intenten identificar quines són les àrees més contaminades de les ciutats, el paràmetre que poden quantificar i tenir en compte és normalment el nivell de so conjuntament amb un seguit d'indicadors acústics o psico-acústics (com el the Traffic Noise Index, el Noise Pollution Level o Intermittency Ratio), però no la font específica que està generant el so. En realitat, el procediment principal per saber quins són els entorns més contaminats són (Bello et al. 2019):

1. Analitzar les queixes dels ciutadans relacionades amb el soroll. Això requereix recursos humans experts que han de traslladar-se a la zona concreta per a estudiar l'entorn i prendre mesures acústiques utilitzant equips específics. No obstant això, a causa de l'elevat nombre de reclamacions que pot rebre una administració pública, no és possible que els experts atenguin aquestes reclamacions en temps real anat a fer les mesures in situ. A més, a causa de les característiques intrínseques volàtils del so (només succeeix quan la font de so és present), en algunes ocasions, quan els experts mesuren l'entorn, la font ja no hi és o ha parat de fer soroll.

2. Estudiar una àrea concreta de la ciutat utilitzant una Wireless Acoustic Sensor Network o Xarxa de Sensors Acústics sense Fils (WASN). En aquest cas, els sensors acústics es despleguen a la ciutat i normalment mesuren el nivell de soroll equivalent.

3. Mapes de soroll fets a partir de mesures en llocs específics de la ciutat, d'acord amb la normativa 2002/49/EC (Parliament 2002). Actualment, aquests mapes usualment consideren només les categories de soroll de trànsit, soroll de tren, soroll d'avions i soroll industrial, però no mostren totes les fonts acústiques que estan sonant a temps real.

A causa de les respostes amb retard obtingudes pel primer mètode i la novetat del segon, l'obtenció d'un sistema (1) de baix cost en termes de maquinari, (2) fiable en termes de precisió i (3) amb temps de resposta baix ha sorgit com un repte de recerca modern.

Per aquesta raó, aquesta dissertació té com a objectiu fer un pas més en la investigació d'aquest camp cap a una implementació real d'un sistema classificador que compleixi les tres característiques abans esmentades. A causa de la complexitat del problema, l'abast d'aquesta tesi es limita a la creació d'un prototipus *hardware* capaç de classificar els esdeveniments acústics (fins i tot si es produeixen simultàniament) i mostrar el resultat de la classificació en temps real en un escenari concret.

1.2 *Baix-cost* i *Temps-real* en Xarxes de Sensors Acústics sense Fils.

El títol d'aquesta dissertació engloba dos termes que poden ser ambigus per a diferents lectors. Aquesta subsecció té com a objectiu debatre i definir-les per a aquesta dissertació.

1.2.1 Baix-cost

Hi ha dos enfocaments arquitectònics principals en el disseny d'una xarxa de sensors distribuïda. El primer, anomenat jeràrquic, té com a objectiu utilitzar un dispositiu de gamma alta (car) que es comporta com un líder (també conegut com a mestre) de tots els altres dispositius de gamma baixa anomenats seguidors. El segon enfocament, anomenat homogeni, té com a objectiu alleujar les tasques de computació del dispositiu mestre mitjançant l'eliminació d'aquest paper del sistema i permetre que tots els dispositius interactuin a voluntat. Per descomptat, en termes de cost, tots dos enfocaments pateixen la mateixa qüestió: el cost de la xarxa de sensors creix linealment amb el nombre de dispositius. Aquest creixement és a un ritme més baix en els sistemes jeràrquics (és a dir, els dispositius de seguiment són més barats) que en els sistemes homogenis. No obstant això, per als sistemes jeràrquics hi ha un punt en el qual el maquinari del dispositiu líder ja no es pot actualitzar i no pot coordinar amb èxit tots els seguidors, resultant en una degradació del rendiment a causa de l'efecte de coll d'ampolla. A més, els sistemes jeràrquics porten altres reptes (per exemple, un sol punt de fracàs o tolerància a les particions de xarxa) que poden fer-los inadequats per a gran escala. Per tant, en aquesta recerca s'ha seleccionat l'enfocament homogeni. Val la pena assenyalar que, per a arquitectures distribuïdes homogènies com la que es proposa en aquesta tesi, minimitzar el cost global de la proposta és equivalent a minimitzar el cost d'un sol dispositiu. Per tant, considerem que el cost d'un sol dispositiu de detecció ha de ser un ordre de magnitud inferior a qualsevol mesurador de nivell de so de classe 1.

1.2.2 Temps-real

Temps-real (o *real-time* en anglès) és un terme estretament connectat al domini d'aplicació en el qual s'està utilitzant. Fins i tot en el mateix domini, diferents concepcions de temps real

poden coexistir per a diferents serveis. Típicament, quan es processen els fluxos de dades, les dades del flux són serialitzats i emmarcats en una finestra (també coneguda com a *chunk*) d'una mida predefinida per tal de ser processades. Quan el processament de dades triga sistemàticament més que omplir una finestra, es requereix una cua infinita per emmagatzemar totes les finestres que cal processar. En aquest context, s'assumeix que el comportament en temps real es produeix quan les dades es processen més ràpid que el temps que triga a omplir una finestra. És a dir, no hi ha finestres a la cua. Per aquesta dissertació, ja que la mida de la finestra seleccionada és de 4 segons, l'abast del temps real es limita a proporcionar una sortida del sistema en menys d'aquesta quantitat de temps.

1.3 Escenari

El centre de la ciutat de Barcelona ha estat seleccionat com a escenari de treball atès que és una de les ciutats més sorolloses d'Europa. De fet, Barcelona ha estat classificada com la setena ciutat més sorollosa del món (*Worldwide Hearing Index* 2017), sent una de les dues ciutats europees classificades en el top-10 de ciutats més sorolloses. El rànquing està encapçalat per Guangzhou (Xina), seguit per Nova Delhi (Índia), El Caire (Egipte), Bombai (Índia), Istanbul (Turquia), Pequín (Xina), Barcelona (Espanya), Ciutat de Mèxic (Mèxic), París (França) i Buenos Aires (Argentina).

Donada l'extensió de la ciutat, s'ha seleccionat una àrea en concret: L'Eixample de la ciutat, que és el districte d'expansió que ocupa una àmplia àrea de la ciutat. Concretament, segons l'ajuntament (*El districte i els seus barris* 2021), aquest districte ocupa 747.60 ha i alberga 266 754 habitants. Per tant, la densitat de població és de 356 inhab./ha.

1.3.1 Pla Cerdà

Aquest districte va ser dissenyat per l'enginyer Ildefons Cerdà el 1860 (Permanyer 2008). Concretament, el projecte d'expansió de la ciutat va ser anomenat *Pla Cerdà*, i l'enginyer va crear i va seguir l'eslògan *Urbanizar el campo y ruralizar la ciudad* (Urbanitzar el camp i *ruralitzar* la ciutat en català). Per sobre de tot, Cerdà tenia com a objectiu construir una ciutat pensant en el futur mitjançant l'anàlisi de les necessitats socials i polítiques de la població.

En el seu pla original, Cerdà va proposar construir una cruïlla cada 113.3 metres, amb l'objectiu de construir un districte en el qual tots els carrers serien camins ràpids. D'aquesta manera, ja que tots els carrers tindrien longituds i amplades similars, el trànsit estaria equilibrat i, per tant, el soroll estaria més o menys distribuït per igual.

L'alç màxima de cada bloc seria de 16 metres, i la superfície total del districte podria albergar 800 000 habitants. L'expansió màxima del districte seria de 7 500 metres, i hi hauria una carretera per creuar tota la ciutat anomenada *Diagonal*. Tots els carrers serien perpendiculars entre ells, i tindrien una amplada d'almenys 20 metres: 10 metres per als vianants (5 metres a cada costat) i 10 per als vehicles. D'aquesta manera, des d'un punt de vista aeri, la ciutat semblaria una xarxa gairebé perfecta de blocs d'edificis i carrers. D'altra

banda, els blocs d'edificis tindrien les cantonades retallades en forma de xamfrà per millorar la visibilitat dels vehicles en les interseccions de trànsit, la qual cosa faria més segurs els encreuaments dels vianants.

Considerant l'amplada de 20 metres dels carrers i l'alçada de 16 metres dels blocs d'edificis, tots els veïns tindrien llum solar directa en algun moment del dia i els blocs veïns no projectarien ombres entre ells.

La Figura 1.1 mostra un mapa topogràfic elaborat per Ildefons Cerdà el 1855, abans de l'expansió de la ciutat. En aquella època, el centre de la ciutat estava envoltat de muralles, que van ser demolides abans de l'expansió de la ciutat.

Per altra banda, la Figura 1.2 mostra el pla original després del seu disseny. Aquest segon mapa il · lustra la idea original de l'Eixample de Barcelona, i il · lustra perfectament la idea que Cerdà tenia per construir la ciutat amb blocs de dimensions idèntiques.

1.3.2 Eixample de Barcelona a l'actualitat

Més de 160 anys després del disseny original del pla, el centre de la ciutat segueix els patrons originals. No obstant això, el districte no és tan gran com Cerdà planejava originalment. A més, l'alçada dels blocs s'ha incrementat de 16 a 20 metres. Actualment, aquesta zona està dividida en els 7 barris que es poden veure a la Figura 1.3:

- *El Fort Pienc*

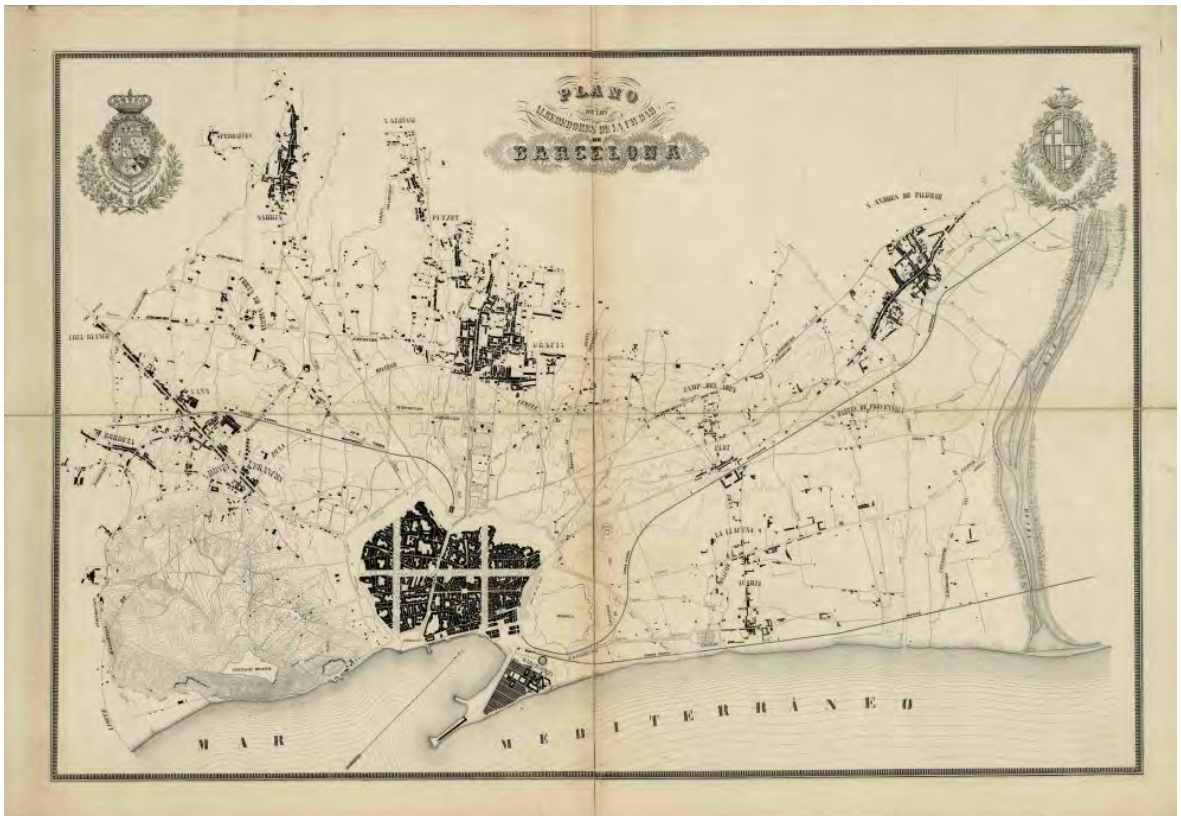


Figura 1.1: Mapa topogràfic el · laborat per Ildefons Cerdà el 1855.

- *La Sagrada Família*
- *La Dreta de l'Eixample*
- *L'Antiga Esquerra de l'Eixample*
- *La Nova Esquerra de l'Eixample*
- *Sant Antoni*

Com es pot observar, la simetria dels blocs facilita la tasca de dissenyar una WASN altament escalable i, per tant, és ideal per al propòsit d'aquesta tesi.

Després d'analitzar els diferents barris de la ciutat i juntament amb el Departament de Qualitat Ambiental de l'Ajuntament, *L' Antiga Esquerra de l'Eixample* s'ha definit com l'àrea d'interès d'aquesta tesi, ja que és la zona que concentra més queixes relacionades amb el soroll. La raó principal d'aquestes queixes són els locals d'oci que ocupen la zona. És a dir, restaurants, bars o llocs de música (la majoria d'ells amb terrasses).

1.4 Preguntes de recerca i objectius de la Tesi

Aquesta subsecció exposa les preguntes de recerca i descriu l'abast dels objectius de la tesi d'aquesta dissertació tenint en compte el context i la motivació exposats a la secció anterior.

Figura 1.2: Pla original de la ciutat dissenyat per Idelfons Cerdà el 1859.

- Research Question 1 o Pregunta de Recerca 1 (RQ1): Podem detectar i identificar esdeveniments acústics en un univers predefinit usant informació espectral i temporal encara que els esdeveniments es produeixin simultàniament?
- Research Question 2 o Pregunta de Recerca 2 (RQ2): És possible encabir un algorisme classificador d'àudio en un dispositiu de baix cost per tal que la classificació doni resultats en temps real?
- Research Question 3 o Pregunta de Recerca 3 (RQ3): Fins a quin punt la redundància física dels sensors pot ajudar a millorar un algorisme classificador d'esdeveniments acústics?

Aquestes qüestions de recerca deriven en els següents objectius de tesi:

- Thesis Objective 1 o Objectiu de Tesi 1 (TO1): **Desenvolupar un sistema classificador automàtic capaç de detectar esdeveniments acústics en ambients urbans utilitzant informació espectral i temporal.**

  Aquest primer objectiu pretén donar una resposta a la pregunta RQ1. El propòsit és concebre un algorisme *software* capaç de classificar aquells esdeveniments que poden ocórrer en un entorn urbà. L'abast d'aquest objectiu es limita a classificar 10 categories diferents de sons. La idea és utilitzar algorismes de Machine Learning o Aprenentatge Automàtic (ML) o Deep Learning o Aprenentatge Profund (DL) per realitzar la classificació.

- Thesis Objective 2 o Objectiu de Tesi 2 (TO2): **Dissenyar una plataforma de maquinari de baix cost capaç de classificar esdeveniments acústics en temps real.**

  Aquest segon objectiu pretén donar una resposta preliminar a la pregunta RQ2. La idea és dissenyar i prototipar una arquitectura de maquinari capaç d'acollir el sistema classificador desenvolupat a l'objectiu TO1. A més, la plataforma hauria de poder

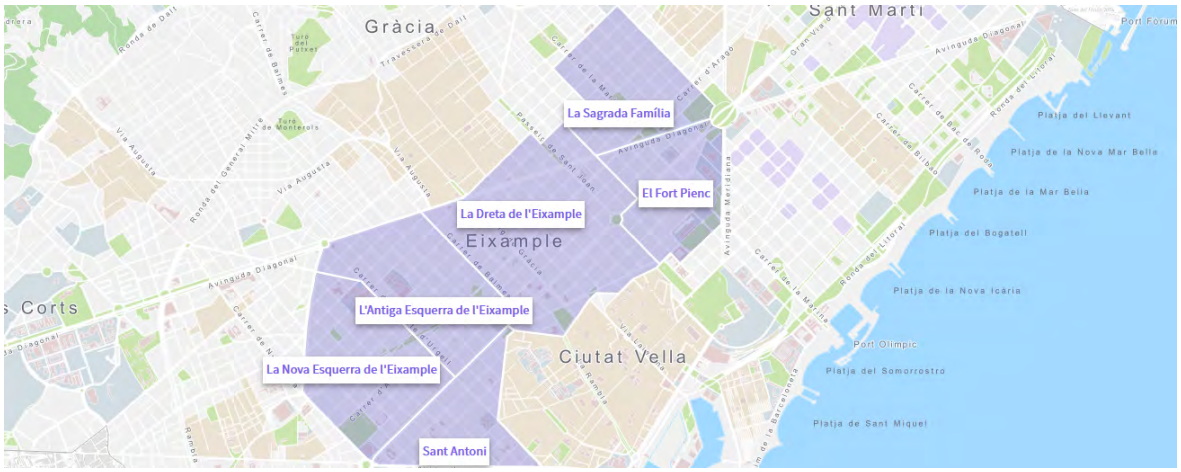


Figura 1.3: Barris de l'Eixample de Barcelona (*El districte i els seus barris* 2021).

generar un resultat de classificació en temps real. Com que el temps real pot tenir diferents interpretacions en funció del lector, per aquesta dissertació es considerarà que un sistema genera un resultat en temps real si és capaç de proporcionar els resultats en un període de temps més curt que una finestra de classificació predefinida. Concretament, en aquesta tesi, el temps de finestra predefinit serà de 4 segons. Més detalls sobre la selecció de finestres s'explicaran més endavant a l'Article II.

- Thesis Objective 3 o Objectiu de Tesi 3 (TO3): **Utilitzar dades del món real per entrenar i avaluar la plataforma de classificació (programari i maquinari) per estudiar la viabilitat d'un desplegament en el món real.**

  Aquest tercer objectiu té com a objectiu avaluar el rendiment del sistema utilitzant dades del món real recopilades en un entorn urbà de cas d'ús. Per tant, com que en ambients urbans els sons solen ocórrer simultàniament, aquest objectiu donarà una resposta final a RQ1.

  Abans del desenvolupament de l'algorisme de classificació, serà necessari estudiar la ciutat seleccionada per a realitzar una anàlisi exhaustiva dels sons que es produeixen al centre de la ciutat. Per limitar l'abast d'aquest objectiu, i a causa de les restriccions de recursos humans per recopilar i anotar dades del món real, només s'utilitzaran un nombre limitat d'hores d'enregistraments.

- Thesis Objective 4 o Objectiu de Tesi 4 (TO4): **Quantificar fins a quin punt la redundància física dels sensors millora la precisió del classificador.**

  En general, i a causa de l'elevat preu dels sensors acústics d'alta qualitat, per estudiar una àmplia àrea urbana només es poden prendre mesures en alguns punts concrets de la ciutat. L'objectiu final és comprovar si una (sobre)població de sensors acústics de baix cost pot millorar els resultats de la classificació per respondre RQ3. En el context d'aquesta tesi, la redundància física dels sensors fa referència a utilitzar els sensors en una certa topologia, de manera que el mateix esdeveniment acústic pugui ser escoltat per múltiples nodes simultàniament. Concretament, l'abast d'aquest objectiu es limita a comprovar una única i concreta topologia de sensors.

D'aquests objectius, s'espera obtenir un prototip com a prova de concepte a Technology Readiness Level o Nivell de Maduresa Tecnològica (TRL) 6. Això vol dir que s'espera que "El model representatiu o sistema prototipus, situat molt més enllà de TRL 5, és testejat en un entorn rellevant. Representa un avenç important en la demostració de la maduresa tecnològica. Exemples inclouen testar un prototip en un entorn de laboratori d'alta fidelitat o en un entorn operacional simulat."(DoD 2011). En aquest sentit, les dades captades al món real es faran servir com a un entorn d'alta fidelitat. Tanmateix, el prototip encara no estarà llest per ser desplegat a qualsevol part o ciutat del món, per a fer-ho caldria arribar a un nivell de TRL més elevat fent més proves i ajustos amb més dades.

1.5 Contribucions de la Tesi

Aquesta subsecció té com a objectiu oferir una visió general de les contribucions presentades en aquesta dissertació juntament amb una explicació del flux de treball utilitzat per a la investigació per aconseguir els objectius de la tesi i respondre a les preguntes de la investigació.

D'acord amb l'actual Reglament del Doctoral de la Universitat Ramon Llull, aquesta tesi es presenta en forma de compendi de publicacions. S'utilitzen tres articles per al compendi: Article I, Article II i Article III.

No obstant això, com a material de suport o complementari, al final d'aquesta dissertació s'adjunten quatre articles i dos pòsters (Article IV, Article VI, Article VII, Article VIII i Article IX), ja que han estat passos complementaris per obtenir les tres contribucions principals i poden ajudar al lector a entendre el flux de treball de la tesi. En concret, mentre que els dos primers articles complementaris estan directament relacionats amb el *firmware* i el *software* desenvolupat per aconseguir els objectius de la tesi i els dos articles intermedis són pòsters presentats en simposis per a propòsits de difusió, els dos últims articles exploren la detecció acústica d'esdeveniments i la classificació en un entorn natural, utilitzant una taxonomia de sons radicalment diferent de les que es poden trobar en àrees urbanes.

A continuació s'explica el flux de treball de la dissertació i la relació de cada article amb les qüestions de recerca i objectius de la tesi.

1.5.1 Anàlisi i definició de taxonomia del paisatge sonor

Com que el tema principal d'aquesta tesi és la detecció d'esdeveniments acústics i la classificació en entorns urbans centrats en l'escenari de cas d'ús del centre de la ciutat de Barcelona, la primera investigació realitzada durant aquesta tesi va ser un estudi i anàlisi del paisatge sonor de la ciutat i el seu impacte en la població. Un dels problemes que es van trobar en aquesta fase va ser la falta de disponibilitat de dades acústiques d'alta qualitat, d'accés obert i amb una etiquetació exacta i precisa captada en l'àrea d'estudi (centre de la ciutat de Barcelona). Per a aquest propòsit, es va anotar i analitzar manualment un conjunt de dades de 6 hores de duració gravades en un balcó d'un carrer situat a *l' Esquerra de l'Eixample*. Les campanyes d'enregistrament van ser dutes a terme per altres investigadors de la universitat en el marc d'un Treball de Final de Màster, però l'autora de la dissertació va dur a terme les etapes d'anotació i anàlisi. Una de les dificultats trobades quan es realitzava la part d'etiquetatge era que, en algunes ocasions, es podien escoltar múltiples esdeveniments simultàniament —incloent tant sorolls de trànsit com de lleure. En aquesta primera etapa, aquests esdeveniments es van considerar esdeveniments rars o estranys i no es van fer distincions entre ells.

La ubicació exacta per a la recopilació de les dades va ser escollida en un carrer envoltat tant per sons de trànsit com d'oci, i els enregistraments es van dur a terme a la nit (entre les 22:00 i les 03:00) i en dissabte per maximitzar la presència de persones en els establiments d'oci. D'aquesta manera, l'anàlisi comprèn els dos tipus d'activitats (tràfic i oci) i permet quantificar quines activitats poden impactar més a la població veïna (considerant la intensitat dels sons i també la seva durada) i la distribució de temps dels esdeveniments (és a dir, en el

moment en què va ocórrer cada esdeveniment).

Aquesta anàlisi va permetre tenir una visió general de la taxonomia de l'àrea d'estudi (el centre de la ciutat de Barcelona), que va constituir el primer pas per aconseguir els objectius de la tesi TO1 i TO3.

Cal destacar que aquesta primera anàlisi del paisatge sonor només inclou paràmetres objectius: és a dir, no es van dur a terme proves perceptives o subjectives als veïns de la ciutat, excepte a la propietària del balcó en el qual va tenir lloc la campanya de gravació, que es va queixar de que la zona era massa sorollosa per dormir a la nit. Una anàlisi perceptiva exhaustiva de la percepció dels veïns que viuen a l'Eixample està fora de l'abast d'aquesta dissertació. També s'ha de destacar que aquest primer treball avalua el paisatge sonor de l'escenari d'ús de la tesi amb l'objectiu de definir una taxonomia que ens permeti començar a definir l'algorisme de classificació de la tesi. Un estudi complet i exhaustiu del paisatge sonor de la ciutat sencera està fora de l'abast d'aquesta tesi, ja que fer-lo comportaria fer més campanyes de gravació distribuïdes per tota la ciutat i en diferents horaris.

Com a contribució a la comunitat científica, els resultats d'aquesta anàlisi, juntament amb el conjunt de dades anotades (que va rebre el nom BCNDataset), es van publicar a l'Article I en la modalitat d'accés obert.

Per tant, l'Article I constitueix el primer article del compendi d'aquesta dissertació. La contribució de l'autora de la tesi en aquest primer treball de compendi va consistir en l'anotació i l'anàlisi (càlculs de mètriques i definició de taxonomia) del conjunt de dades i l'escriptura de l'article. Els altres autors de l'article van dur a terme les campanyes d'enregistrament.

1.5.2 Disseny de la WASN i desenvolupament d'un algorisme de classificació *single-class*

Un cop es va acabar d'analitzar l'àrea geogràfica a estudiar i es va definir la primera taxonomia, el següent pas va consistir en dissenyar una plataforma distribuïda de baix cost capaç de reconèixer esdeveniments acústics. Els detalls concrets sobre els nodes de detecció seleccionats s'expliquen a Article IV. Els principals criteris de selecció per als nodes de detecció seguien les següents premisses:

- Les unitats de computació haurien de poder realitzar la classificació en temps real.
- El cost total dels sensors hauria de ser inferior a 100€ cadascun per ser classificat com a sensors de baix cost.
- La resposta en freqüència dels micròfons ha de ser el més plana possible, mantenint la premissa de baix cost.
- La premissa de baix cost és més important que la precisió en les mesures (és a dir, els nodes no necessiten ser classificats com a sensors de classe A).
- Fer servir una unitat de computació genèrica de baix cost que suporti les llibreries de *software* utilitzades per programar un sistema classificador facilitarà la tasca de desplegar

els algoritmes. A més, si la unitat de computació genèrica té una gran comunitat de suport en línia, serà més fàcil resoldre problemes potencials.

Després d'una anàlisi exhaustiva i diferents proves, la Raspberry Pi va ser seleccionada com a unitat de computació per a cadascun dels sensors. A més, en quant a microfonia, es va escollir fer servir un micròfon USB plug-and-play.

Una vegada seleccionades les unitats de computació, la segona contribució principal de la tesi, que utilitza els mateixos nodes de detecció de Article IV, s'explica a l'Article II. Més concretament, en aquesta segona contribució, es va proposar una arquitectura de computació distribuïda i es va provar utilitzant un conjunt de dades descarregades d'un repositori en línia. En l'arquitectura de computació distribuïda, els nodes de detecció es van organitzar en una topologia que permet aprofitar la redundància física. Aprofitar la redundància física significa que un esdeveniment acústic pot ser escoltat per diferents sensors. Aquest treball té com a objectiu avaluar si la redundància física és útil per al classificador, i, per tant, és el primer pas cap a la consecució del TO4. Un problema que es va trobar a l'intentar avaluar aquesta idea va ser que, normalment, en conjunts de dades en línia, les dades s'han recopilat en un sol lloc. De fet, l'autora de la tesi creu que no existeix cap conjunt de dades acústiques en línia recopilades en llocs simultanis propers a Barcelona que permetin avaluar la hipòtesi i validar la topologia proposada. Per aquesta raó, en aquest treball, una vegada que el sensor va ser dissenyat, es va entrenar un algorisme de DL per classificar 10 categories diferents de sons urbans. Més concretament, els 10 sons que es van avaluar van ser els presents en el conjunt de dades d'UrbanSound8K (Salamon et al. 2014): aire condicionat, clàxon de cotxe, nens jugant, gos bordant, so de taladre, so de motor de cotxe, dispar de pistola, martell pneumàtica, sirena i música de carrer. Aquest conjunt de dades en línia és considerat per la comunitat una base per a la classificació de sons urbans. A més, aquest conjunt de dades (UrbanSound) subministra els arxius d'àudio en finestres de fins a 4 segons. Això significa que, quan hom es descarrega el conjunt de dades, hi ha uns 8 000 fitxers .wav, i la durada de cada fitxer és igual (o inferior) a 4 segons.

En el context d'aquesta tesi, després d'algunes proves de concepte en les quals la mida de la finestra va ser variada duent a terme una recerca sistemàtica (que va des d'uns 100 mil · lisegons a 4-segons), 4-segons és la mida de les finestres que va resultar en millors resultats de classificació tenint en compte el compromís entre la precisió dels resultats i la velocitat de classificació. Per tant, 4 segons és la mida de la finestra utilitzada en els experiments fets en aquesta tesi.

Aquest algorisme de classificació vol satisfer l'objectiu TO1. No obstant això, com que el conjunt de dades UrbanSound consisteix només en dades sintètiques d'un únic sensor i no conté el soroll de trànsit de fons típic que es pot escoltar a les ciutats, una vegada que el sistema classificador va ser avaluat en un únic sensor, es va dur a terme una emulació d'un entorn real mitjançant la mescla dels arxius d'àudios del conjunt de dades UrbanSound amb dades urbanes de trànsit del món real. Específicament, el soroll de trànsit seleccionat va ser el de la categoria *road traffic noise*, que principalment consisteix en soroll de cotxes en marxa. Per al procediment de mescla, i per poder avaluar si la redundància física de la

topologia proposada millora els resultats de la classificació, els àudios es van barrejar imitant la propagació del so d'un esdeveniment acústic a quatre sensors veïns propers d'acord amb l'arquitectura de sensors dissenyada en la dissertació. Per dur a terme el procés d'imitació, es va considerar únicament la distància entre els esdeveniments acústics i els punts de mesura. L'autora de la tesi és conscient que hi ha altres factors —a més a més de la distància— que poden afectar la propagació del so, tal com poden ser les reflexions en edificis o objectes dinàmics del paisatge com vehicles, arbres, o inclús vianants caminant. Tanmateix, a causa de les dificultats que es van trobar en plantejar fer una caracterització de la resposta impulsional del carrer (per exemple, soroll de fons constant, dificultats a l'hora d'estar a peu dret en el lloc on se suposa que ocorren els esdeveniments acústics a causa del pas de vehicles, l'esdeveniment transitori que s'hauria de generar hauria de tenir un nivell de so molt alt, etc.), aquestes variables es van ometre. Per a garantir un nivell alt de fidelitat en els sons imitats, però, el conjunt de dades modificat es va escoltar atentament de forma manual.

Per assegurar que l'algorisme seleccionat funcionaria en els nodes *hardware* seleccionats, tot i que l'entrenament dels sensors es va dur a terme en un ordinador amb una GPU potent, els experiments es van córrer en el sensor seguint la estructura que seguirien si s'haguessin de desplegar al carrer: adquisició de dades, processament de dades, transformació d'espectrogrames i classificació. A més, per a la comunicació entre sensors, es va proposar l'ús d'una antena ad hoc de tipus *bespoke* a cada node de computació.

La contribució de l'autora de la tesi en aquest segon treball del compendi va consistir principalment en el disseny dels nodes de detecció de baix cost fet a partir de materials comercials (és a dir, Raspberry Pi i micròfon USB) així com el desenvolupament i l'avaluació del *software* de classificació. La idea del protocol distribuït per enviar bytes entre sensors i el disseny de l'antena *bespoke* personalitzada presentada en el treball va ser duta a terme pel altres autors. De fet, l'antena *bespoke* no s'ha implementat a la vida real, només s'ha simulat. La implementació física de l'antena està fora de l'abast d'aquesta dissertació. La comunicació entre nodes d'aquesta tesi s'ha dut a terme a través d'Internet (Ethernet o Wi-Fi).

Per comprovar visualment si la classificació estava passant en temps real sobre els nodes de detecció, es va dissenyar una Printed Circuit Board o Placa de Circuit Imprès (PCB) que contenia 10 LEDs (un per categoria) més una capa serigràfica amb el nom de cada esdeveniment acústic i una petita pantalla LCD. Cada vegada que un sensor detecta un esdeveniment acústic, el LED amb l'etiqueta que coincideix amb la sortida de la classificació s'activa. A més, la pantalla LCD mostra la probabilitat que aquest esdeveniment sigui cert. Per tant, si a la pantalla es mostra un alt valor de probabilitat, el classificador està bastant segur sobre la seva predicció. Per contra, un valor de baixa probabilitat indica que l'algorisme no està segur sobre la seva decisió. Per provar el sistema general, vam reproduir diversos sons ambientals de categories específiques com ara sirenes, clàxons de cotxes o sorolls d'obres procedents de vídeos. La imatge de la Figura 1.4 mostra un sensor amb la PCB dissenyada. A https://youtu.be/NQiwXDrfyUc es pot veure un vídeo de demostració del funcionament de l'un sensor de la WASN classificant esdeveniments a temps real.

Figura 1.4: Sensor i PCB mostrant com un so de gos és classificat correctament.

1.5.3 Dades del món real i classificació polifònica

Després de la primera avaluació de l'arquitectura proposada utilitzant el conjunt de dades en línia de 10 categories i després de validar que la topologia proposada ens permet comprovar si la redundància física millora la precisió de la classificació, el següent pas de la dissertació va consistir a recollir enregistraments d'àudio del món real en l'escenari seleccionat. Per a aquest propòsit, es van dur a terme dues campanyes d'enregistrament en quatre localitats simultànies amb redundància física als carrers de *L' Antiga Esquerra de l'Eixample*. Les quatre ubicacions simultànies compleixen amb els requisits establerts en la topologia proposada: cada node és una cantonada d'un edifici en una intersecció de trànsit.

La raó per dur a terme dues campanyes d'enregistrament en lloc d'una va ser que, durant la primera campanya d'enregistrament, hi havia restriccions de mobilitat a causa de la pandèmia COVID-19 que podrien esbiaixar els esdeveniments acústics que ocorren als carrers (es permetia a la gent moure's per raons justificades com ara anar a treballar si el teletreball no era possible, però no hi havia persones al carrer amb finalitats d'oci). La segona campanya d'enregistrament va tenir lloc uns mesos després. Durant la segona campanya d'enregistrament, les restriccions es van suavitzar. L'hora del dia i la estació de l'any de les dues campanyes d'enregistrament també van ser diferents, perseguint una varietat més rica de paisatge sonor al carrer (l'estació tardorenca a la tarda i la temporada primaveral al matí). En total, es van obtenir unes 20 hores de dades acústiques (5 hores per sensor). Actualment, aquest conjunt de dades s'està

processant i analitzant amb l'objectiu de ser publicat i compartit amb la comunitat científica. No obstant això, com que aquest treball encara no ha acabat, l'anàlisi profund del conjunt de dades no forma part d'aquesta tesi.

Com a primera fase de prova, després de la primera campanya d'enregistrament, però abans de realitzar la segona, 1 hora d'una ubicació dels arxius d'àudio recollits es va avaluar utilitzant el sistema de classificació automàtic de l'Article II sense fer cap modificació. Això significa que l'algorisme divideix l'àudio en finestres de 4 segons de longitud i assigna a cada finestra una etiqueta provisional que consisteix en 1 de les 10 categories del conjunt de dades UrbanSound8K. Aquesta etiqueta va ser revisada manualment per l'autora de la tesi. Després, es van obtenir les mètriques de classificació de l'algorisme i es van presentar en l'article de conferència complementari Article V. El procés també es va il · lustrar en el pòster presentat al' Article VI. La raó de fer aquesta prova abans d'operar amb totes les dades disponibles era comprovar si el treball realitzat amb dades sintètiques podia ser extrapolat a dades del món real de l'escenari d'ús. No obstant això, els resultats d'aquest estudi van indicar que hi havia diversos esdeveniments acústics que no pertanyien a cap de les 10 categories predefinides, i també que la majoria dels fragments d'àudio contenen més d'un esdeveniment acústic. Això va fer que la recerca de l'autora conclogués que seria convenient utilitzar un classificador polifònic o multietiqueta en lloc del que s'utilitzava fins a aquest moment.

De nou, trobar dades en línia que coincideixin amb les característiques desitjades va ser un problema. L'autora va haver d'etiquetar manualment els arxius d'àudio de les dues campanyes d'enregistrament utilitzant un enfocament multietiqueta (és a dir, etiquetar tots els esdeveniments acústics que es poden escoltar en cadascun dels fragments, encara que es produeixin simultàniament). Com que l'etiquetatge manual d'arxius d'àudio és una tasca exhaustiva, que consumeix molt de temps i que sensible a errors, es va dissenyar i implementar un procés d'anotació mitjançant un script de Python. La idea era minimitzar la quantitat de temps dedicat a la tasca minimitzant les interaccions de l'usuari amb el ratolí i/o el teclat de l'ordinador. A més, es van utilitzar etiquetes febles en lloc d'etiquetes fortes amb l'objectiu de reduir el temps dedicat a la tasca. És a dir, en un fragment donat de 4 segons (que seria l'entrada de l'algorisme), totes les etiquetes van ser etiquetades, independentment de la segmentació exacta de l'esdeveniment acústic. Això significa que si un esdeveniment curt només durava uns pocs mil · lisegons, l'etiqueta s'assignaria al fragment complet de 4 segons de totes maneres. La lògica subjacent a aquesta idea és que, encara que es perd precisió en l'etiquetatge, el sistema generaria un resultat de classificació per cada fragment de 4 segons. Així doncs, sabent quins esdeveniments ocorren en una finestra de temps és suficient, no es necessita tenir la informació exacta de on comença i acaba l'esdeveniment. Cal tenir en compte que la mida de la finestra es va mantenir a partir de l'article anterior (és a dir, 4 segons).

Una vegada que el conjunt de dades va ser etiquetat, el classificador polifònic es va implementar utilitzant una DNN en cadascun dels sensors. Aquest classificador va ser dissenyat perquè pogués classificar tots els esdeveniments etiquetats del conjunt de dades (i no només les 10 categories del conjunt de dades UrbanSound). Un problema notable que es

va trobar quan es va entrenar el classificador basat en DL va ser que les dades del món real presenten un alt desequilibri de classe (és a dir, no tots els esdeveniments apareixen la mateixa quantitat de vegades en el conjunt de dades). Alguns esdeveniments es repeteixen gairebé constantment, mentre que altres apareixen només en unes poques ocasions. Això va fer que el classificador tingués problemes en avaluar la classificació de les dades mal representades en el conjunt de dades. Per mitigar aquest problema, es van aplicar tècniques d'augmentació de dades. La tècnica específica que es va seleccionar va ser *mix-up augmentation*, que consisteix en barrejar (per mitjà d'una suma ponderada) diferents fragments d'àudio i després combinar les seves etiquetes també. Aquest procés de mescla es va fer utilitzant el BCNDataset, (dataset analitzat i publicat a l'Article I), i el conjunt de dades UrbanSound8k. Els resultats obtinguts en aquest procés s'exposen al pòster presentat a l'Article VII.

Una vegada que el sistema polifònic estava corrent en cadascun dels sensors, per avaluar la redundància física dels sensors, es va afegir una capa intel · ligent basada en ML al sistema. Aquesta capa recull els resultats de la classificació de la xarxa neuronal que s'executa en cada node veí i dona una sortida de classificació final. Es van calcular diverses mètriques sobre els resultats de la classificació per poder discutir si s'aconsegueixen els objectius TO3 i TO4. A més, el sistema es va provar en tres unitats de computació diferents (models de Raspberry Pi diferents) per validar també si el sistema és capaç de proporcionar un resultat en temps real i, per tant, si també s'aconsegueix l'objectiu TO2. Tot aquest sistema polifònic, juntament amb el sistema intel · ligent basat en ML, es detalla en la tercera contribució del compendi: l'Article III.

La contribució de l'autora de la tesi en aquest tercer treball del compendi ha consistit principalment en el desenvolupament del *software* de classificació multietiqueta i les proves sobre els sensors físics. A més, l'autora ha organitzat les campanyes de recopilació de dades i ha dut a terme l'anotació de dades.

1.5.4 Treball complementari en bioacústica

Independentment dels tres articles principals que componen el compendi, durant la tesi s'han dut a terme dos articles més relacionats amb la bioacústica.

En primer lloc, Article VIII explora i proposa un sistema *software* que distingeix les vocalitzacions i sons dels ocells picots (*woodpeckers*) que habiten a la Península Ibèrica. La contribució de l'autora de la tesi per a aquest treball ha estat el disseny d'un sistema de classificador de dues capes, l'extracció de característiques i la programació del classificador.

En segon lloc, Article IX explora el paisatge sonor d'un entorn natural proper a l'Aeroport de Barcelona i té com a objectiu proposar un sistema de classificació dels sons detectats. En aquest treball, l'autora de la tesi ha contribuït en la campanya de recopilació de dades, el disseny d'avaluació experimental i l'escriptura de l'article. Un altre autor ha dut a terme la programació del *software*.

1.6 Organització de la memòria de Tesi

La dissertació s'organitza en els següents capítols:

Capítol 2 : Explora l'estat de l'art en el camp de la detecció acústica d'esdeveniments i la classificació en entorns urbans mitjançant una revisió sistemàtica de la literatura. Les publicacions més rellevants s'estudien per saber què han desenvolupat els investigadors de tot el món en els últims anys.

Capítol 3 : Inclou els tres articles principals del compendi de la tesi:

Article I : 'BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset', exposa el treball realitzat per concebre i analitzar un conjunt de dades de 6 hores d'enregistraments obtinguts en una àrea del centre de la ciutat de Barcelona. L'anàlisi inclou la durada dels esdeveniments detectats en el conjunt de dades, la relació senyal-soroll, el nombre d'ocurrències, l'impacte de cada ocurrència en el soroll de fons $L.Aeq$, i l' Intermittency Ratio o Ratio d'Intermitència (IR) de les dades, que són mètriques que poden estar correlacionades amb els efectes del soroll en la població.

Article II : 'Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring', presenta una arquitectura distribuïda de baix cost altament escalable que compta amb una xarxa de sensors acústics per controlar els sons urbans. Per validar analíticament la viabilitat de l'arquitectura *hardware* proposada, els experiments de classificació es duen a terme utilitzant un conjunt de dades que conté 10 categories d'àudio diferents. D'altra banda, per comprovar si la redundància física millora els resultats de la classificació, els arxius d'àudio estan adaptats sintèticament per imitar la propagació del so en un cert emplaçament al centre de la ciutat de Barcelona.

Article III : 'Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors'. Fent servir dades enregistrades i annotades recollides en quatre punts simultanis en una ubicació concreta del centre de la ciutat de Barcelona, aquest treball avalua l'arquitectura del sistema proposada a l'Article II utilitzant un classificador multietiqueta de dues etapes. Mostra com les tècniques d'augment de dades ajuden al sistema a obtenir una major mètrica de classificació i la quantitat de temps que triga al sistema a classificar una mostra quan s'utilitzen tres unitats de computació diferents. A més, aquest article explica una nova metodologia d'etiquetatge per a accelerar el procés d'anotació.

Capítol 4 : Finalitza la tesi *i)* resumint les principals contribucions d'aquesta recerca, *ii)* discutint els resultats obtinguts i *iii)* proposant possibles línies de futur. A més, relaciona els resultats obtinguts amb els objectius de la tesi i respon a les preguntes de recerca que s'han proposat a la Secció 1.4.

Capítol 5 : Inclou alguns articles complementaris al compendi de la tesi:

Article IV : 'Low-Cost WASN for Real-Time Soundmap Generation', presenta una arquitectura *hardware* de baix cost concebuda per recollir dades acústiques per construir un mapa de so en temps real 24/7. Cada node de la xarxa es compon d'un micròfon omnidireccional i una unitat de computació (Raspberry Pi), que processa informació acústica localment per obtenir dades no sensibles (és a dir, nivells de soroll equivalents o etiquetes d'esdeveniments acústics) que més tard s'envien a un servidor al núvol. L'objectiu final del sistema és permetre les següents funcions: *i)* mesurar l' L_{eq} o altres paràmetres similars en temps real en una finestra predefinida, *ii)* identificar patrons canviants en les mesures anteriors de manera que es puguin detectar situacions anòmales i *iii)* per prevenir i assistir a possibles situacions irregulars.

Article V : 'Improving classification accuracy of acoustic real-world urban data using sensors physical redundancy', avalua el rendiment de la xarxa de sensors acústics de baix cost que s'aprofita de la redundància física presentada a l'Article II. Per fer-ho, el treball avalua més d'1 hora de dades acústiques del món real recollides al centre de la ciutat de Barcelona. L'article vol avaluar si la redundància física ajuda a obtenir resultats de classificació més robustos. El sistema avaluat incorpora una xarxa neuronal profunda que funciona en cada node i un protocol de consens distribuït que implementa un conjunt d'heurístiques per beneficiar-se dels resultats de la classificació dels nodes veïns desplegats a la mateixa àrea (és a dir, la redundància física).

Article VI : 'Prototyping a low-cost Wireless Acoustic Sensor Network with physical redundancy to automatically classify acoustic events in urban environments', mostra un pòster presentat en un simposi internacional sobre sons urbans. Aquest treball va ser el pas intermedi a nivell de *software* entre els treballs Article II i Article III. Analitzant una hora d'enregistraments del món real, es va provar la DNN de l'Article II i es van analitzar les seves febleses.

Article VII : 'Multilabel acoustic event classification for urban sound monitoring at a traffic intersection', mostra un pòster presentat en un simposi local a Barcelona sobre el tema de DL. Aquest pòster resumeix els resultats de la primera capa de classificació obtinguda a l'Article III per a propòsits de difusió i promoció.

Article VIII : 'A Two-Stage Approach To Automatically Detect and Classify Woodpecker (Fam. *Picidae*) Sounds', proposa un sistema de classificador de dues capes per classificar els sons dels ocells picot que habiten a la península Ibèrica. Més específicament, l'arquitectura proposada compta amb un sistema de classificació d'aprenentatge de dues etapes que utilitza *i)* Coeficients de Cepstral de Freqüència Mel i Taxa de Creuament Zero per detectar sons d'ocells sobre soroll ambiental, i *ii)* Coeficients Predictius lineals Perceptuals i Coeficients Cepstrals de Freqüència Mel per identificar les espècies d'ocells i el tipus de so (és a dir, sons vocals o sons fets a les branques dels arbres amb el bec) associat.

Article IX : 'Analysis of the Noise Impact of the Airport of Barcelona to the Llobregat Delta Natural Environment during the 2021 Lockdown period', explora el paisatge sonor acústic d'un parc natural prop de l'aeroport de Barcelona i aplica tècniques d'aprenentatge automàtic per classificar els esdeveniments acústics produïts tant per l'activitat aeroportuària com per la vida salvatge. Per a l'anàlisi, s'utilitzen dades registrades en tres punts simultanis d'interès biològic (segons els comissaris del parc) prop de l'aeroport. Els enregistraments i l'anàlisi posterior es van fer el 5 de març de 2021, quan l'activitat aeroportuària encara es veia molt minvada per les restriccions de mobilitat.

Referències

Abbaspour, Majid, Karimi, Elham, Nassiri, Parvin, Monazzam, Mohammad Reza i Taghavi, Lobat (2015). 'Hierarchal assessment of noise pollution in urban areas–A case study'. A: *Transportation Research Part D: Transport and Environment* vol. 34, pàg. 95 - 103.

Bello, Juan P, Silva, Claudio, Nov, Oded, Dubois, R Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles i Doraiswamy, Harish (2019). 'Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution'. A: *Communications of the ACM* vol. 62, núm. 2, pàg. 68 - 77.

DoD, US (2011). 'Technology readiness assessment (TRA) guidance'. A: *Revision posted* vol. 13.

El districte i els seus barris (2021). URL: https://ajuntament.barcelona.cat/eixample/ca/el-districte-i-els-seus-barris/el-districte-i-els-seus-barris. (accessed: 18.11.2021).

Parliament, European (2002). *Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise*. European Parliament.

Permanyer, Lluís (2008). *L'Eixample : 150 anys d'història / Lluís Permanyer ; selecció fotogràfica de Daniel Venteo*. cat. Barcelona.

Radle, Autumn Lyn (2007). 'The effect of noise on wildlife: a literature review'. A: *World Forum for Acoustic Ecology Online Reader*, pàg. 1 - 16.

Salamon, J., Jacoby, C. i Bello, J. P. (nov. de 2014). 'A Dataset and Taxonomy for Urban Sound Research'. A: *22nd ACM International Conference on Multimedia (ACM-MM'14)*. Orlando, FL, USA, pàg. 1041 - 1044.

University, Cambridge (2021). 'Cambridge dictionary'. A.

WHO (2011). *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*. World Health Organization. Regional Office for Europe.

Worldwide Hearing Index (2017). URL: https://www.mimi.io/en/blog/2017/3/8/worldwide-hearing-index-2017. (accessed: 22.11.2021).

Chapter 1

Introduction

1.1 Context and Motivation

The first entry of noise in the dictionary defines it as "a sound, especially a loud, unpleasant sound" (University 2021). In the modern and ever-evolving society, the presence of noise has become a daily threat to a worrying amount of the population (WHO 2011). However, it is not only humans who are affected by noise: some studies have proved that being exposed to noise caused by humans has a negative impact on wildlife, causing non-natural migrations, reproductive problems and even long-term survival threats (Radle 2007).

In the concrete case of human population, a set of studies (WHO 2011) carried out by the World Health Organization o Organització Mundial de la Salut (WHO) confirm that being exposed to excessive noise levels interferes with day to day activities such as working, attending to school or resting during leisure time. Moreover, the same studies calculate the Estimated Burden of Disease o Càrrega Estimada de les Malalties (EBD) of the population caused by environmental noise segmented in road traffic noise, aircraft noise and railway noise. This estimation aims to quantify the potential years of life that a person may lose by a premature death plus the years of healthy life lost by poor health conditions or disabilities. This estimation is measured in units of Disability-Adjusted Life Years o Anys de Vida Ajustats per Discapacitats (DALYs). The main side effects of being over-exposed to noise are:

- **Cardiovascular diseases**: The risk of suffering from ischaemic heart disease (even myocardial infraction) is increased by being overexposed to both road traffic noise and aircraft noise. Also, these two types of noises are correlated with an increment of the risk of suffering from anomalous high blood pressure. Concretely, it is estimated that the EBD for cardiovascular diseases in high-income European countries is of 61 000 years.

- **Cognitive impairment in children**: It has been studied by means of experimental and epidemiological studies. This impairment occurs while children are exposed to noise and persists for some time after the noise finishes. Concretely, to carry out the studies, it was observed that while all the children that were exposed to a level of 95 Decibels with A-weighting o Decibels amb Ponderació A (dBA) were cognitively affected, no children were affected at a level of 59 dBA. The EBD for European countries is of 45 000 years from children ranging from 7 to 19 years old.

- **Sleep disturbance**: To calculate the EBD caused by sleep disturbance on the population, two types of studies were taken into account. First of all, electro-physiological

measurements. Secondly, by means of self-reports made with surveys in different studies. The EBD for European citizens living in towns with a population greater than 50 000 inhabitants is of 903 000 DALYs lost due to sleep disturbance caused by noise.

- **Tinnitus**: Tinnitus can be defined as the sensation of hearing a concrete sound when that sound is not actually happening (i.e., a ringing, clicking or buzzing sound). This affection may derive in other patologies such as sleep disturbance, annoyance, frustration or anxiety. The EBD for adult European citizens is of 22 000 years due to tinnitus caused by being overexposed to noise.

- **Annoyance**: To measure annoyance caused by noisy environments, personal questionnaires are used. The EBD for citizens living in towns with a population greater than 50 000 inhabitants is of 587 000 DALYs lost due to annoyance caused by noise.

Also, apart from all these side effects, it has been studied that it is not only the level of noise that matters but also the type of sound that the citizens are exposed to. That is, not all the acoustic events have the same impact on population (Abbaspour et al. 2015). However, when public or private administrations try to identify which are the most polluted areas of the cities, the parameter that they can quantify and take into consideration is usually the sound level of a certain area together with a set of acoustic indicators (such as the Traffic Noise Index, Noise Pollution Level, Intermittency Ratio), but not the specific source that is generating the sound. Actually, the main procedure to know which are the more polluted environments are (Bello et al. 2019):

1. Analysing the complaints from the citizens related to noise. This requires expert human resources that have to move to the concrete area to be studied and take acoustic measurements using specific equipment. However, due to the high amount of complaints that a public administration may receive, it is not possible for the experts to be present in real-time and make the measurements. Also, due to the intrinsic volatile characteristics of sound (it happens only when the sound source is present), in some of the occasions, when the experts measure the environment the noise source has already finished.

2. Surveying a concrete area of the city using a Wireless Acoustic Sensor Network o Xarxa de Sensors Acústics sense Fils (WASN). In this case, acoustic sensors are deployed on the city and they typically measure the equivalent noise level.

3. Noise mapping developed by means of measurements in specific spots of the city, according to the regulations of the Directive 2002/49/EC (Parliament 2002). Currently, these maps usually consider only the categories of road traffic, railway noise, aircraft noise and industrial noise, but they do not find all the acoustic sources occurring in real-time.

Due to the delayed responses obtained by the first method and the novelty of the second one, obtaining a (1) low-cost in terms of hardware, (2) reliable in terms of accuracy and (3) responsive system has emerged as a modern research challenge.

For this reason, this dissertation aims to give a step further in the research of this field towards a real-operation implementation of a classifier system that accomplishes the three aforementioned features. Due to the complexity of the problem, the scope of this thesis is limited to prototyping a hardware system capable of classifying acoustic events (even if they occur simultaneously) and output the classification results in real-time in a concrete use-case scenario.

1.2 *Low-cost* and *Real-time* in Wireless Acoustic Sensor Networks

The title of this dissertation englobes two terms that may be ambiguous to different reader. This subsection aims at discussing and defining them for this dissertation.

1.2.1 Low-cost

There are two main architectural approaches when designing a distributed sensor network. The first one, named hierarchical, aims to use a high-end device (i.e., expensive) which behaves as a leader (also referred to as master) of all the other low-end devices referred to as followers. The second approach, named homogeneous, aims to alleviate the computing tasks of the master device by erasing this role from the system and enable all the devices to interact at will. Certainly, in terms of cost, both approaches suffer from the same issue: the cost of the sensor network grows linearly with the number of devices. This growth is at a lower rate in hierarchical systems (i.e., follower devices are cheaper) than in homogeneous systems. However, for hierarchical systems there is a point in which the hardware of the leader device cannot be upgraded anymore and fails to successfully coordinate all the followers, resulting in a performance degradation due to the bottle-neck effect. Additionally, hierarchical systems bring other challenges (e.g., single point of failure or tolerance to network partitions) that may make them unsuitable for large-scale WASN. Therefore, in this research the homogeneous approach has been selected. It is worth noting that, for homogeneous distributed architectures like the one herein proposed, minimizing the overall cost of the proposed WASN is equivalent to minimize the cost of a single device. Therefore, we consider that low-cost is achieved when the cost of a single sensing device is an order of magnitude lower than any of-the-shelf Class 1 sound level meter.

1.2.2 Real-time

Real-time is a term tightly connected to the application domain in which it is being used. Even in the same domain, different conceptions of real-time may coexist for different services. Typically, when processing data streams, data from the stream are serialized and framed into a window (i.e., also referred to as chunk) of a predefined size in order to be processed. When the data processing systematically takes longer than filling a window, an infinite queue is required to store all the windows that need to be processed. In this context, real-time behavior

is assumed to happen when data are processed faster than the time that it takes to fill a window. That is, no windows are queued. For the sake of this dissertation, as the selected window size is 4 seconds, the scope of real-time is limited to providing a system output in less than this amount of time.

1.3 Use-case scenario

The city centre of Barcelona has been selected as a use-case scenario given that it is one of the noisiest cities Europe. Actually, Barcelona has been categorized as the seventh noisiest city in the world (*Worldwide Hearing Index* 2017), being one of the two European cities ranked in the top-10 noisiest cities ranking. The ranking is leaded by Guangzhou (China), followed by New Delhi (India), Cairo (Egypt), Bombay (India), Istanbul (Turkey), Beijing (China), Barcelona (Spain), Mexico City (Mexico), Paris (France) and Buenos Aires (Argentina).

Given the extension of the city, a concrete area has been selected: the "Eixample" of the city, which is the expansion district that occupies a wide area of the city. Concretely, according to the city hall (*El districte i els seus barris* 2021), this district occupies 747.60 ha and hosts 266 754 inhabitants. Hence, the population density is of 356 inhab./ha.

1.3.1 Pla Cerdà

This district was designed by the engineer Ildefons Cerdà in 1860 (Permanyer 2008). Concretely, the project of expanding the city was named *Pla Cerdà* (Cerdà plan in English), and the engineer created and followed the slogan *Urbanizar el campo y ruralizar la ciudad* (Urbanize the countryside and *ruralize* the city in English). Among all, Cerdà aimed at building a city thinking in the future by means of analysing the social and political needs of the population.

On his original plan, Cerdà proposed to build a crossroad each 113.3 meters, aiming to build a district in which all the streets were fast paths. This way, as all the streets would have similar lengths and widths, the traffic would be balanced and, hence, the noise would be more or less equally distributed.

The maximum height of each block would be of 16 meters, and the total area of the district would be able to host 800 000 inhabitants. The maximum expansion of the district would take 7 500 meters, and there would be a road to cross all the city named *Diagonal*. All the streets would be perpendicular between them, and had a width of, at least, 20 meters: 10 meters for pedestrians (5 meters on each side) and 10 for vehicles. This way, in an aerial view, the city would look like an almost perfect grid of building blocks and streets. Moreover, the building blocks would have chamfered corners to improve the visibility of vehicles at traffic intersections, which would make the crossing of pedestrians safer.

Considering the width of 20 meters of the streets and the height of 16 meters of the building blocks, all the neighbours would have direct sun light at some moment of the day and the neighbouring blocks would not project shadows between them.

Figure 1.1 shows a topographic map elaborated by the same Ildefons Cerdà in 1855 before the expansion of the city. By that time, the city centre was surrounded by walls, that were

demolished before the expansion of the city started.

On the contrary, Figure 1.2 shows the original map plan after its design. This second map illustrated the original idea of the Eixample of Barcelona, and perfectly illustrated the idea that Cerdà had for constructing the city with blocks of identical dimensions.

1.3.2 Eixample of Barcelona nowadays

More than 160 years after the original design of the plan, the center of the city still follows the original patterns. However, the district is not as big as Cerdà originally planned. Also, the height of the blocks has been increased from 16 meters to 20 meters. Nowadays, this zone is divided in the 7 neighbourhoods that can be seen in Figure 1.3:

- *El Fort Pienc*
- *La Sagrada Família*
- *La Dreta de l'Eixample*
- *L'Antiga Esquerra de l'Eixample*
- *La Nova Esquerra de l'Eixample*
- *Sant Antoni*

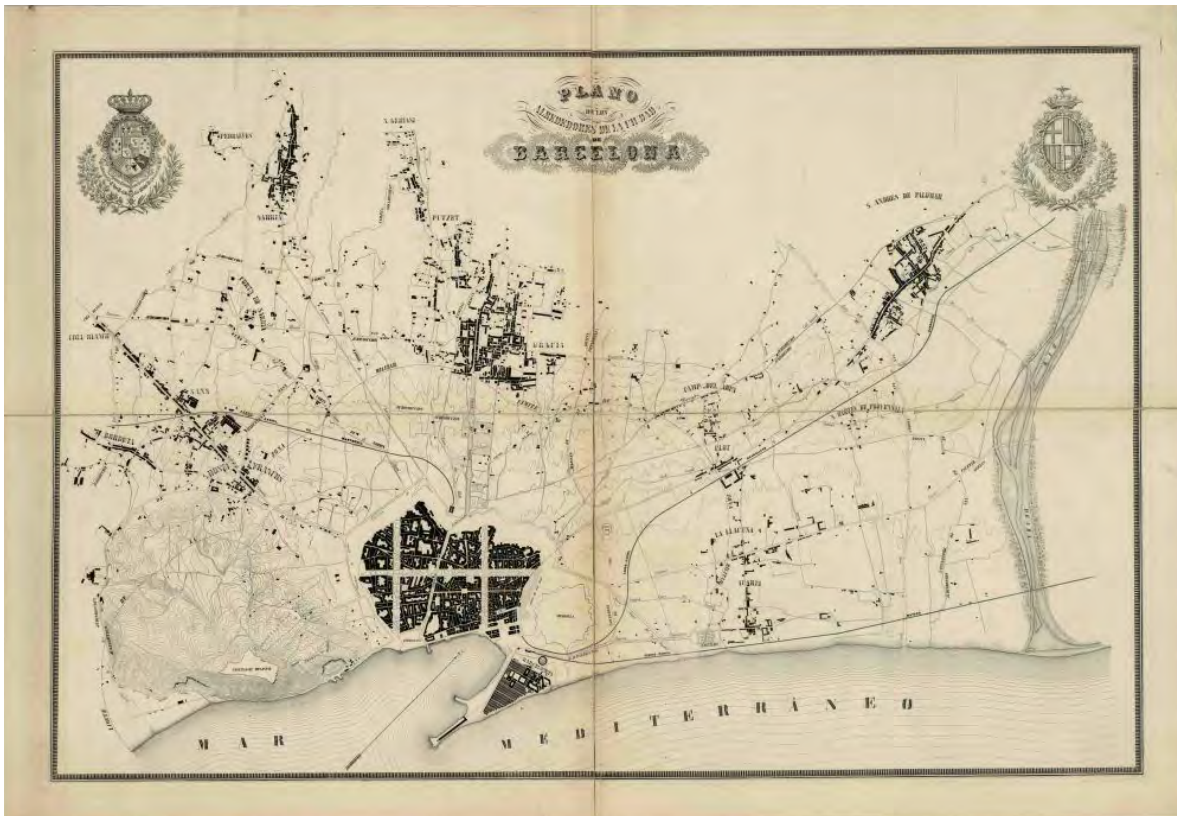


Figure 1.1: Topographic map elaborated by Ildefons Cerdà in 1855.

Figure 1.2: Original plan of the city designed by Ildefons Cerdà in 1859.

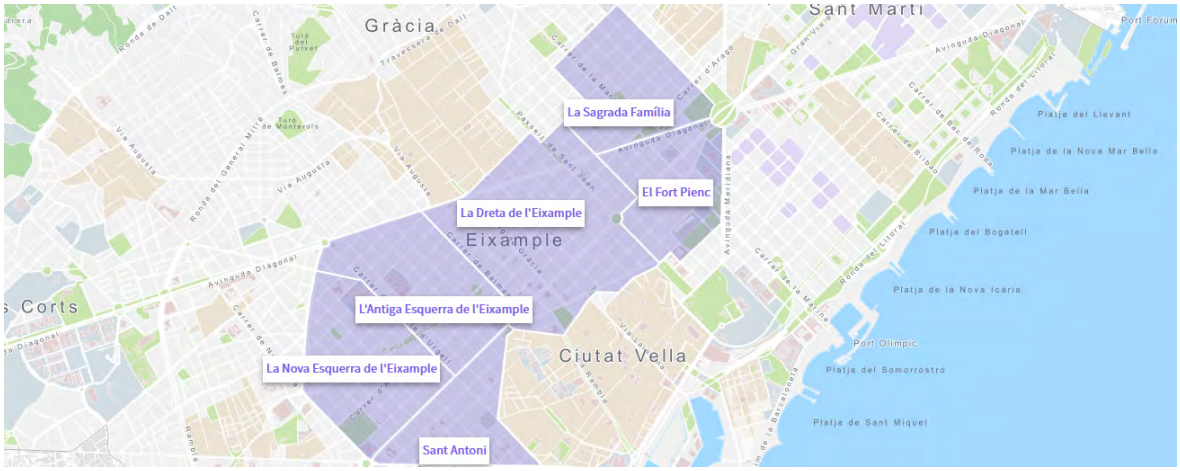


Figure 1.3: Neighbourhoods of the Eixample district of Barcelona (*El districte i els seus barris* 2021).

As it can be observed, the symmetry of the blocks eases the task of designing a highly scalable WASN, and hence it is ideal for the purpose of this thesis.

After analysing the different neighbourhoods of the city and together with the Environmental Quality Department of the City Council, *L'Antiga Esquerra de l'Eixample* has been defined as the area of interest for this thesis, as it is the zone that concentrates more noise-related complaints. The main reason for those complaints are the leisure locals occupying the area such as restaurants, bars or music venues (most of them containing terraces).

1.4 Research question and thesis objectives

This subsection exposes the research questions of the thesis and describes the scope of the thesis objectives of this dissertation considering the context and motivation exposed in the previous section.

- Research Question 1 o Pregunta de Recerca 1 (RQ1): Can we detect and identify acoustic events in a predefined universe using spectral and temporal information even if they occur simultaneously?
- Research Question 2 o Pregunta de Recerca 2 (RQ2): Is it possible to fit an audio classifier algorithm in a low-cost device so it outputs the classification results in real-time?
- Research Question 3 o Pregunta de Recerca 3 (RQ3): Up to what extent physical redundancy of sensors can help improving an acoustic classifier algorithm?

These research questions derive in the following thesis objectives:

- Thesis Objective 1 o Objectiu de Tesi 1 (TO1): **Develop an automatic classifier system capable of detecting acoustic events occurring in urban environments using spectral and temporal information.**

  This first objective aims to give an answer to RQ1. The purpose is to conceive a software algorithm capable of classifying those events that may typically occur in a urban environment. The scope of this objective is limited to classifying 10 different categories of sounds. The idea is to use Machine Learning o Aprenentatge Automàtic (ML) or Deep Learning o Aprenentatge Profund (DL) algorithms to perform the classification.

- Thesis Objective 2 o Objectiu de Tesi 2 (TO2): **Design a low-cost hardware platform capable of classifying acoustic events in real-time.**

  This second objective aims to give a preliminary answer to RQ2. The idea is to design and prototype a hardware architecture capable of hosting the classifier system developed in TO1. Also, the platform should be able to output a classification result in real-time. As real-time may have different interpretations depending on the reader, for the sake of this dissertation it will be considered that a system outputs a result in real-time if it is

able to supply the results in a period of time shorter than a predefined classification window. Concretely, in this thesis, the predefined window time will be 4-seconds. More details regarding the window selection will be further explained in Paper II.

- Thesis Objective 3 o Objectiu de Tesi 3 (TO3): **Use real-world data to train and evaluate the classification platform (hardware and software) in order to study the feasibility of a real-world deployment.**

  This third objective aims to evaluate the performance of the system using real-world data gathered in a use-case urban environment. Hence, as in urban environments the sounds typically occur simultaneously, this objective will give a final answer to RQ1.

  Prior to the development of the classification algorithm, it will be necessary to study the selected city be means of performing an exhaustive analysis to the sounds that occur in the city centre. To limit the scope of this objective, and due to human resources restrictions for gathering and annotating real-world data, only a few limited number of hour recordings will be used.

- Thesis Objective 4 o Objectiu de Tesi 4 (TO4): **Quantify up to what extent physical redundancy of sensors improves the accuracy of the classifier.**

  Usually, and due to the elevated price of high-quality acoustic sensors, to survey a wide urban area only a few acoustic points can be measured. This forth objective aims to check whether an (over)population of low-cost acoustic sensors may improve the classification results to answer RQ3. For this thesis, physical redundancy of sensors stands for using the sensors in a certain topology so the same acoustic event can be heard by multiple sensing nodes. Concretely, the scope of this objective is limited to check physical redundancy in a single and concrete set-up topology of sensors.

From these objective, it is expected to obtain a functional prototype as a proof of concept at Technology Readiness Level o Nivell de Maduresa Tecnològica (TRL) 6. That is, a "Representative model or prototype system, which is well beyond that of TRL 5, is tested in a relevant environment. Represents a major step up in a technology's demonstrated readiness. Examples include testing a prototype in a high-fidelity" (DoD 2011). In this sense, the real-world data gathered in the street will be used as a relevant environment. However, the prototype will still not be ready to be deployed in any part of the world, as to do so a higher level of TRL should be achieved by means of more tests using more data.

1.5 Thesis contributions

This subsection aims at giving an overview to the contributions presented in this dissertation together with an explanation of the workflow used for researching to accomplish the thesis objectives and answer the research questions.

According to the current Doctoral Regulations of the Ramon Llull University Doctoral program, this thesis is presented in the form of a compendium of publications. Three papers are used for the compendium: Paper I, Paper II and Paper III.

However, as supporting or complementary material, at the end of this dissertation four papers and two posters (Paper IV, Paper V, Paper VI, Paper VII, Paper VIII and Paper IX) are attached as well as they have been complementary steps for obtaining the three main contributions and may help the reader understanding the work-flow of the dissertation. Specifically, while the two first complementary papers are directly related to the firmware and software developed to achieve the thesis objectives and the two intermediate papers are posters presented at symposiums for dissemination purposes, the two last papers explore acoustic event detection and classification in a natural environments, using a taxonomy of sounds radically different to the ones that can be found in urban areas.

The work-flow of the dissertation and the relation of each paper to the research questions and objectives of the thesis is explained below.

1.5.1 Soundscape analysis and taxonomy definition

As the main topic of this thesis is acoustic event detection and classification in urban environments focused in the use-case scenario of the city centre of Barcelona, the first research carried out during this thesis was a study and analysis of the soundscape of the city and its impact to the population. One problem found at this stage was the unavailability of high-quality, open access and accurately labelled acoustic data on the area of study (city centre of Barcelona). For this purpose, a 6-hours length dataset gathered in a balcony of a street located in *l' Antiga Esquerra de l'Eixample* was manually annotated and analysed. The recording campaigns were carried out by other researchers of university in the frame of a Master's Degree Final Project, but the author of the dissertation conducted the annotation and analysis stages. One difficulty encountered when performing the tagging part was that, in some occasions, multiple events —including both traffic and leisure sounds— could be heard simultaneously. At this first stage, those events were considered as rare or strange events and distinctions were not made between them.

The exact location for gathering the data was chosen in a street surrounded by both traffic and leisure sounds, and the recordings were carried out at night time (between 22:00 and 03:00) and on Saturday to maximize the presence of people in the leisure establishments. This way, the analysis comprises both types of activities (traffic and leisure) and enables to quantify which activities may impact the most to the neighbouring population (considering the intensity of the sounds and also their duration) and the time distribution of the events (that is, at what moment did each event occur).

This analysis enabled to have an overview of the taxonomy of the area of study (the city centre of Barcelona), which constituted the first step towards accomplishing the thesis objectives TO1 and TO3.

It must be highlighted that this first analysis of the soundscape includes only objective parameters: that is, no perceptual or subjective tests were carried out to the neighbors of the city except for the owner of the balcony in which the recording campaign took place, who complained about the area being too noisy for proper sleep at night. An exhaustive perceptual analysis of perception of neighbors living in the Eixample is out of the scope of

this dissertation. It must also be considered that this first work that evaluates the soundscape of the use-case scenario serves for the purpose of defining a taxonomy that enable us to start defining the classification algorithm of this thesis. A complete and exhaustive study of the soundscape of the full city of Barcelona is out of the scope of this dissertation, as doing so would require more hours of recordings, gathered in distributed spots of the city and in different schedules.

As a contribution to the scientific community, the results of this analysis, together with the annotated dataset (that was given the name BCNDataset), were published in Paper I in open-access modality. Hence, Paper I constitutes the first paper of the compendium of this dissertation. The thesis author contribution in this first compendium work consisted in the annotation and analysis (metrics calculations and taxonomy definition) of the dataset and the writing of the paper. The recording campaigns were carried out by other authors of the paper.

1.5.2 Design of the WASN and development of a single-class algorithm

Once the geographic area to be surveyed was analyzed and the first taxonomy was defined, the next step consisted on designing a low-cost distributed platform capable of recognizing acoustic events. Concrete details regarding the selected sensing nodes are explained in Paper IV. The main selection criteria for the sensing nodes consisted on the following premises:

- Computing units should be able to perform real-time classification.
- The total cost of the sensors should be under 100€ each to be categorized as low-cost sensors.
- The frequency response of the microphones should be as flat as possible while maintaining the low-cost premise.
- The low-cost premise is more important than the precision in the measurements (that is, the sensing nodes do not need to be categorized as class A sensors).
- Using a low-cost comercial generic computing unit that supports the software libraries used to program a classifier system eases the task of deploying the algorithms for inference. Also, if the generic computing unit has big a supporting online community, troubleshooting for potential problems will be easier.

After an exhaustive analysis and different test, Raspberry Pi was selected as the computing unit of each of the sensors, and a plug-and-play USB was connected to it to gather acoustic information in real-time.

Once the computing units were selected, the second main contribution of the thesis, that further develops and uses the same sensing nodes of Paper IV, is explained in Paper II. More concretely, in this second contribution, a distributed computing architecture was proposed and tested using an online dataset. In the distributed computing architecture, the sensing nodes were arranged in a topology that allows to take advantage of physical redundancy. Take advantage of physical redundancy means that an acoustic event may be heard from different

sensors. This work aims at evaluating if physical redundancy is helpful for the classifier, and, hence, is the first step towards achieving TO4. One problem found when trying to evaluate this idea was that, usually, in online datasets, data has been gathered in a single spot. In fact, to the best of the author thesis knowledge, there exists no acoustic dataset gathered in close simultaneous spots that enables to evaluate the hypothesis and validate the proposed topology. For this reason, for this work, once the sensor was designed, a DL algorithm was trained to classify 10 different categories for urban sounds. More concretely, the 10 sounds that were evaluated were the ones present in the UrbanSound8K dataset (Salamon et al. 2014): air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music. This online dataset is considered a baseline for urban sound classification. Also, this dataset (UrbanSound) supplies the audio files in windows of up to 4-seconds. This means that, when the dataset is downloaded, there are about 8 000 .wav files, and the duration of each file is equal (and in some audio files smaller) than 4-seconds.

In the context of this thesis, after some evaluation tests in which the window size was varied conducting a grid search (ranging from about 100 milliseconds to 4-seconds), 4-seconds is the windows size that resulted in better classification results taking into account a trade-off between accuracy and classification speed. Hence, 4-seconds is the window size used in further experiments on this thesis.

The classification algorithm aims at accomplishing TO1. However, as the dataset consists only on single sensor data from a synthetic environment and does not contain the typical background traffic noise that can be heard in cities, once the classifier system was evaluated in a single sensor, an emulation of a real environment was carried out by means of mixing the clean audio files from the dataset with real-world traffic urban data . Specifically, the selected traffic noise belonged to the category *road traffic noise*, which mainly consists on the by-pass of cars. For the mixing procedure, and to be able to evaluate if the physical redundancy of the proposed topology improves the classification results, audios were mixed by imitating the sound propagation of an acoustic event to four close neighbor sensors according to the sensors architecture designed in the dissertation. For the imitation, we took into consideration only the distance between the acoustic event and the measuring point. The author of the thesis is aware that other factors—besides from the distance— may affect the propagation of the sound, such as the reflections in buildings or moving objects from the soundscape such as vehicles, trees or even pedestrians passing by. However, due to difficulties when trying to characterize the impulse response in the street (such as background noise occurring constantly, difficulties in trying to stand in the site in which the event would occur due to vehicles passing by, the transitory event that we would have to generate would be very loud, etc.), these variables were omitted. Nonetheless, to guarantee high fidelity of the sounds, the modified dataset was carefully listened.

To make sure that the selected algorithm would work in the selected hardware nodes, even though the training of the sensors was carried out in a computer with a powerful GPU, the testing was evaluated in the sensor and using all the pipeline of data acquisition, data processing, spectrogram transformation and classification. Also, for the communication

between sensors, a custom bespoke antenna was proposed to be attached to each computing node.

The thesis author contribution in this second compendium work has mainly consisted in the design of low-cost sensing nodes composed of commercial materials (i.e., Raspberry Pi and USB microphone) and the development and testing of the classifier software. The idea of the distributed protocol to send bytes between sensors and the design of the custom bespoke antenna presented on the work were carried out by other authors. Actually, the custom bespoke antenna has not been implemented in real-life, it has just been simulated. The physical implementation of the antenna is out of the scope of this dissertation. The communication between nodes in this thesis has been carried out via Internet (Ethernet or Wi-Fi).

To visually check if the classification was happening in real-time over the sensing nodes, a Printed Circuit Board o Placa de Circuit Imprès (PCB) to be plugged on top of the Raspberry Pi was designed containing 10 LEDs (one per category) plus a silk-screen layer with the name of an acoustic event and a tiny LCD display. Each time that a sensor detects an acoustic event, the LED with the label matching the classification output turns on. Moreover, the LCD display shows the probability of that event being true. Hence, a high probability value shown on the screen indicates that the classifier is pretty confident about its prediction. On the contrary, a low probability value indicates that the algorithm is unsure about its decision. To test the overall system, we provided several environmental sounds from specific categories such as sirens, car horns or drilling sounds coming from videos. A picture of the sensing node with the PCB is shown in Figure 1.4. To see a demo video of a node of the WASN classifying events in real-time, please check https://youtu.be/NQiwXDrfyUc .

1.5.3 Real-world data and polyphonic classification

After the first evaluation of the proposed architecture using the 10-categories online dataset and after validating that the proposed topology enables us to check whether physical redundancy improves the classification accuracy, the next step of the dissertation consisted of collecting real-world audio recordings in the selected use-case scenario. For this purpose, two recording campaigns were carried out in four simultaneous locations with physical redundancy in the streets of *L'Antiga Esquerra de l'Eixample*. The four simultaneous locations match the requirements stablished in the proposed topology: each node is a corner of a building in a traffic intersection.

The reason for conducting two recording campaigns instead of one was that, during the first recording campaign, there were mobility restriction due to COVID-19 pandemic that may bias the acoustic events happening in the streets (people were allowed to move for justified reasons such as going to work if telework was not possible, but they were not in the street for leisure purposes). The second recording campaign took place a few months after. During the second recording campaign, the restrictions were softened. The daytime and season of the two recording campaigns were also different, pursuing a richer variety of soundscape in the street (Autumn season in the afternoon and Spring season in the morning). In total, about 20

Figure 1.4: Sensing node and PCB showing how a dog barking sound is classified.

hours of acoustic data were obtained (5 hours per sensor). Currently, this dataset is being processed and further analyzed with the aim of being published for the sake of the scientific community. However, as this work is not yet finished, the deep analysis of the dataset is not part of this thesis dissertation.

As a first testing stage, after the first recording campaign but before conducting the second one, 1 hour of one location of the collected audio files was evaluated using the automatic classifier system of Paper II without doing any modification. This means that the algorithm splits the audio in windows of 4-seconds length and assigns to each window a provisional label consisting on 1 of the 10 categories of the UrbanSound8K dataset. This label was manually revised by the thesis author and classification metrics were obtained and presented in the complementary conference paper Paper V. The process was also illustrated in the poster presented at Paper VI. The reason of doing this test before operating with all the available data was to check if the work conducted with synthetic data could be extrapolated to real-world data from the use-case scenario. However, results of that study indicated that there were several acoustic events that did not belong to any of the 10 predefined categories, and also mostly all the audio fragments contained more than one acoustic event. This made the author thesis conclude that it would be convenient to use a polyphonic or multilabel classifier instead of the one used until that moment.

Again, finding online data matching the desired characteristics was a problem. The author

had to manually label the audio files from the two recording campaigns using a multilabel approach (that is, labelling all the acoustic events that can be heard in each of the fragments, even if they occur simultaneously). As manually labelling audio files is an exhaustive, time consuming and error prompt task, a newly designed annotation process was implemented by means of a Python script. The idea was to minimize the amount of time spent on the task by minimizing the user interactions with the mouse or the keypad of the computer. Also, weak labels were used instead of strong labels aiming to reduce the time spent in the task. That is, in a given 4-seconds fragment (that would be the input of the algorithm), all the labels were tagged, independently of the exact segmentation of the acoustic event. This means that if a short event lasted a few milliseconds only, the label would be assigned to the full 4-seconds fragment anyway. The logic behind this idea is that, even though labelling precision is lost, as the system would output a classification result for each 4-seconds fragment, knowing the events occurring in that window of time would be enough. Note that the window size was maintained from the previous work (i.e., 4-seconds).

Once the dataset was labelled, the polyphonic classifier was implemented using a DNN on each of the sensors. This classifier was designed so it could classify all the events labelled from the dataset (and not only the 10 categories from the UrbanSound dataset). A remarkable problem found when training the DL-based classifier was that real-world data presents a high class-imbalance (i.e., not all the events appear the same amount of times in the dataset. Some events are repeated almost constantly, while some others appear only in a very few occasions). This caused the classifier to struggle when assessing the classification of data poorly represented in the dataset. To mitigate this problem, data augmentation techniques were evaluated and applied. The specific technique that was selected was mix-up augmentation, which consists on mixing (by means of a weighted sum) different audio fragments and then combining their labels as well. This mix-up process was done using the BCNDataset, which is the one analyzed and published in Paper I , and the UrbanSound8k dataset. The results obtained in this process are exposed in the poster presented in Paper VII.

Once the polyphonic system was working on each of the sensors, to assess physical redundancy of sensors, an intelligent ML-based layer was added to the system. This layer gathers the classification results of the neural network running on each neighboring node and gives a final classification output. Several metrics were calculated over the classification results to be able to discuss if TO3 and TO4 are achieved. Also, the system was tested on three different computation units (different Raspberry Pi models) to validate also if the system is able to supply a result in real-time and, hence, if TO2 is achieved as well. All this polyphonic system, together with the ML-based intelligent system, is detailed in the third contribution of the compendium, which is Paper III.

The thesis author contribution in this third compendium work has mainly consisted in the development of the multilabel classifier software and the tests over physical sensors. Also, the author has organized the data gathering campaigns and carried out the data annotation.

1.5.4 Complementary work in bioacoustics

Independently from the three main works that compose the compendium, during the thesis, two works related to bioacoustic monitoring have been carried out.

First, Paper VIII explores and proposes a software system that distinguishes vocalizations and sounds from woodpeckers inhabiting Iberian Peninsula. The thesis author contribution for this work has been the design of a two-layers classifier system, the feature engineering and feature extraction and the programming of the classifier.

Second, Paper IX explores the soundscape of a natural environment close to Barcelona Airport and aims at proposing a classifier system from the detected sounds. In this work, the thesis author has contributed in the data gathering campaign, the experimental evaluation design and the paper writing. The software programming has been carried out by another author.

1.6 Dissertation roadmap

The dissertation is arranged in the following chapters:

Chapter 2 : explores the state of the art in the field of acoustic event detection and classification for urban environments by means of a systematic literature review. The most relevant publications are studied to know what have the researchers around the world developed in the same field in the latest years.

Chapter 3 : includes the three main papers of the thesis compendium:

Paper I : 'BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset', reports on the work conducted to conceive and analyse a 6 hours of recordings dataset obtained in a lively area of Barcelona city center. The analysis includes the duration of the events detected on the dataset, the signal-to-noise ratio, the number of occurrences, the impact of each occurrence on the background noise L_{Aeq}, and the Intermittency Ratio o Ratio d'Intermitència (IR) of the entire data samples, which are metrics that can be correlated to health effects of noise in population.

Paper II : 'Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring', presents a highly scalable low-cost distributed infrastructure that features an ubiquitous acoustic sensor network to monitor urban sounds. To analytically validate the feasibility of the proposed hardware architecture, classification experiments are conducted using a dataset containing 10 different audio categories. Moreover, to check up to what extend physical redundancy may help to improve the classification results, the audio files are synthetically adapted to imitate the sound propagation in a certain location on the city center of Barcelona.

Paper III : 'Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors'. Using real-world recorded and manually labelled urban data

gathered in four simultaneous spots in a concrete location of the city centre of Barcelona, this work tests the system architecture proposed on [Paper II] using a two-stage multilabel classifier. It shows how data augmentation techniques help the system obtaining higher classification metrics and the amount of time that it takes to the system to test one sample when using three different computation units. Moreover, this paper explains a new labelling methodology to speed up the annotation process.

Chapter 4 : concludes the dissertation of the thesis by *i)* summarizing the main contributions of the work, *ii)* discussing on the achieved results and *iii)* explaining the potential future research directions. Also, it links the obtained results to the thesis objectives and answers the research questions that have been proposed in Section 1.4.

Chapter 5 : includes some complementary papers to the thesis compendium:

Paper IV : 'Low-Cost WASN for Real-Time Soundmap Generation', presents a low-cost hardware architecture conceived to gather acoustic data to build a 24/7 real-time soundmap. Each node of the network is composed of an omnidirectional microphone and a computation unit (Raspberry Pi), which processes acoustic information locally to obtain non-sensitive data (i.e., equivalent continuous loudness levels or acoustic event labels) that are later sent to a cloud server. The ultimate goal of the system is to enable the following functions: *i)* to measure the L_{eq} or other similar parameters in real-time in a predefined window, *ii)* to identify changing patterns in the previous measurements so that anomalous situations can be detected and *iii)* to prevent and attend potential irregular situations.

Paper V : 'Improving classification accuracy of acoustic real-world urban data using sensors physical redundancy', assesses the performance of the approach composed of a low-cost acoustic wireless sensor network that takes advantage of physical redundancy presented in Paper II. To do it, the work evaluates over 1-hour of real-world acoustic data gathered in the city centre of Barcelona if physical redundancy helps obtaining more robust classification results. The evaluated system incorporates a deep neural network running in each sensor node and a distributed consensus protocol that implements a set of heuristics to benefit from the classification results of neighboring nodes surveying the same area (i.e., physical redundancy).

Paper VI : 'Prototyping a low-cost Wireless Acoustic Sensor Network with physical redundancy to automatically classify acoustic events in urban environments', depicts a poster presented at an international symposium with the topic of urban sounds. This work was the intermediate software step between Paper II and Paper III. Analyzing one hour of real-world recordings, the DNN of Paper II was tested and their weaknesses were analyzed.

Paper VII : 'Multilabel acoustic event classification for urban sound monitoring at a traffic intersection', shows a poster presented at a local symposium in Barcelona with the topic

of DL. This poster summarizes the results of the first classification layer obtained in Paper III for dissemination and promotion purposes.

Paper VIII : 'A Two-Stage Approach To Automatically Detect and Classify Woodpecker (Fam. *Picidae*) Sounds', proposes a two-layers classifier system to classify sounds from woodpeckers inhabiting the Iberian Peninsula. More specifically, the proposed architecture features a two-stage Learning Classifier System that uses *i)* Mel Frequency Cepstral Coefficients and Zero Crossing Rate to detect bird sounds over environmental noise, and *ii)* Linear Predictive Cepstral Coefficients, Perceptual Linear Predictive Coefficients and Mel Frequency Cepstral Coefficients to identify the bird species and sound type (i.e., vocal sounds such as advertising calls, excitement calls, call notes and drumming events) associated to that bird sound.

Paper IX : 'Analysis of the Noise Impact of the Airport of Barcelona to the Llobregat Delta Natural Environment during the 2021 Lockdown period', explores the acoustic soundscape of a natural park near the airport from Barcelona and applies machine learning techniques to classify the acoustic events produced by both airport activity and wildlife. For the analysis, data recorded in three simultaneous spots of biological interest (according to the park's curators) near the airport is used. The recordings and posterior analysis were made on March 5, 2021, when airport activity was still greatly diminished by the mobility restrictions.

References

Abbaspour, Majid, Karimi, Elham, Nassiri, Parvin, Monazzam, Mohammad Reza and Taghavi, Lobat (2015). 'Hierarchal assessment of noise pollution in urban areas–A case study'. In: *Transportation Research Part D: Transport and Environment* vol. 34, pp. 95–103.

Bello, Juan P, Silva, Claudio, Nov, Oded, Dubois, R Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles and Doraiswamy, Harish (2019). 'Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution'. In: *Communications of the ACM* vol. 62, no. 2, pp. 68–77.

DoD, US (2011). 'Technology readiness assessment (TRA) guidance'. In: *Revision posted* vol. 13.

El districte i els seus barris (2021). URL: https://ajuntament.barcelona.cat/eixample/ca/el-districte-i-els-seus-barris/el-districte-i-els-seus-barris. (accessed: 18.11.2021).

Parliament, European (2002). *Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise*. European Parliament.

Permanyer, Lluís (2008). *L'Eixample : 150 anys d'història / Lluís Permanyer ; selecció fotogràfica de Daniel Venteo*. cat. Barcelona.

Radle, Autumn Lyn (2007). 'The effect of noise on wildlife: a literature review'. In: *World Forum for Acoustic Ecology Online Reader*, pp. 1–16.

Salamon, J., Jacoby, C. and Bello, J. P. (Nov. 2014). 'A Dataset and Taxonomy for Urban Sound Research'. In: *22nd ACM International Conference on Multimedia (ACM-MM'14)*. Orlando, FL, USA, pp. 1041–1044.

University, Cambridge (2021). 'Cambridge dictionary'. In.

WHO (2011). *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*. World Health Organization. Regional Office for Europe.

Worldwide Hearing Index (2017). URL: https://www.mimi.io/en/blog/2017/3/8/worldwide-hearing-index-2017. (accessed: 22.11.2021).

Capítol 2

Estat de l'art

Per dur a terme un estat de l'art complert en el camp de la detecció d'esdeveniments acústics en entorns urbans, s'ha dut a terme una revisió sistemàtica de la literatura. El propòsit principal d'una revisió sistemàtica de la literatura és recollir les proves i les troballes més rellevants d'un tema de recerca concret a través d'un procés sistemàtic (Amo Filvà et al. 2020). Per tant, el treball que es duu a terme en aquest capítol té com a objectiu fer una investigació exhaustiva sobre el tema que s'està debatent en aquesta dissertació (és a dir, la classificació d'esdeveniments acústics en entorns urbans).

2.1 Metodologia

Per realitzar una revisió sistemàtica de la literatura satisfactòria, el primer pas és seleccionar una estratègia que permeti escollir les publicacions i la informació en línia més rellevants del camp d'estudi (és a dir, per a aquest treball, detecció i classificació d'esdeveniments acústics). Per a aquest propòsit, s'ha seguit la metodologia explicada a (Khan et al. 2003), que consisteix en cinc passos:

1. **Escollir les preguntes per a la revisió**: En aquest sentit, les preguntes que s'han de resoldre són les preguntes d'investigació que s'indiquen al Capítol 1 d'aquest document: RQ1, RQ2 i RQ3. A partir d'aquestes qüestions de recerca s'han definit els criteris d'inclusió i exclusió per a seleccionar les obres més rellevants.

2. **Identificar treballs rellevants**: Per aquest propòsit, la base de dades principal que s'ha escollit per cercar informació és el Web Of Science (WOS). El procés de trobar treballs rellevants es duu a terme de manera iterativa, i els criteris d'inclusió i exclusió s'apliquen per seleccionar només les obres adequades per a la revisió. A més, aquesta etapa considera els criteris de Population, Intervention, Comparison, Outcome, Context o Població, Intervenció, Comparació, Resultats, Context (PICOC) per donar una resposta apropiada a les preguntes de recerca, limitant l'abast de la revisió.

3. **Avaluació de la qualitat dels estudis**: Per fer-ho, s'ha tingut en compte el nombre de cites i mitjans en els quals es va publicar la informació. Concretament, s'han considerat fonts de dades fiables les conferències revisades per parells, les revistes i els informes tècnics.

4. **Resumir les evidències**: Obtenir un resum de la metodologia i les conclusions principals de cada treball.

5. **Interpretar els resultats**: S'ha de comprovar l'heterogeneïtat de les dades i decidir si es poden confiar en els resultats obtinguts a partir del resum de les obres.

2.2 Criteris d'inclusió i d'exclusió

Els criteris d'inclusió concrets que s'han seguit per a incloure articles a la revisió són:

Inclusion Criteria 1 o Criteris d'Inclusió 1 (IC1): Les obres recollides estan dins del camp de la detecció d'esdeveniments acústics **AND**

Inclusion Criteria 2 o Criteris d'Inclusió 2 (IC2): Les obres recollides contenen informació sobre el processament de senyals acústics en temps real OR informació sobre arquitectures de xarxes de sensors acústics sense fils de baix cost **AND**

Inclusion Criteria 3 o Criteris d'Inclusió 3 (IC3): Les obres estan escrites en anglès **AND**

Inclusion Criteria 4 o Criteris d'Inclusió 4 (IC4): Les obres estan publicades en conferències revisades per parells, revistes o informes tècnics.

I els criteris d'exclusió concrets que s'han escollit per incloure articles a la revisió són:

Criteris d'Exclusió 1 (CE1): Les obres recollides no estan dins del camp de la detecció d'esdeveniments acústics **OR**

Criteris d'Exclusió 2 (CE2): Les obres recollides no contenen informació sobre el processament de senyals acústics en temps real OR informació sobre arquitectures de xarxes de sensors acústics sense fils de baix cost **OR**

Criteris d'Exclusió 3 (CE3): Les obres no estan escrites en anglès **OR**

Criteris d'Exclusió 4 (CE4): Les obres no estan publicades en conferències revisades per parells, revistes o informes tècnics.

2.3 Consultes

Per cercar la informació, es van formular diverses consultes en el WOS, que van proporcionar els resultats mostrats a la Taula 2.1:

Consultes	Nombre de resultats
TS=(acoustic classification AND urban)	299
TS=(low cost device* AND acoustic event classification)	22
TS=(real-time classification AND acoustic events AND urban)	9
TS=(urban sound AND acoustic dataset)	51
TS=(multilabel classification AND acoustic event)	16
TS=(event classification AND urban soundscape)	12

Taula 2.1: Consultes i nombre de resultats fetes al WOS per obtenir informació.

Això significa que, en total, les cerques van resultar en 409 resultats en el WOS. Les consultes es van executar a la data de desembre de 2021, per la qual cosa les obres publicades més tard no s'han inclòs. No obstant això, aquests resultats es van haver de filtrar abans del seu processament, ja que alguns resultats contenien treballs duplicats de diferents consultes o no complien amb els criteris d'inclusió i exclusió descrits anteriorment.

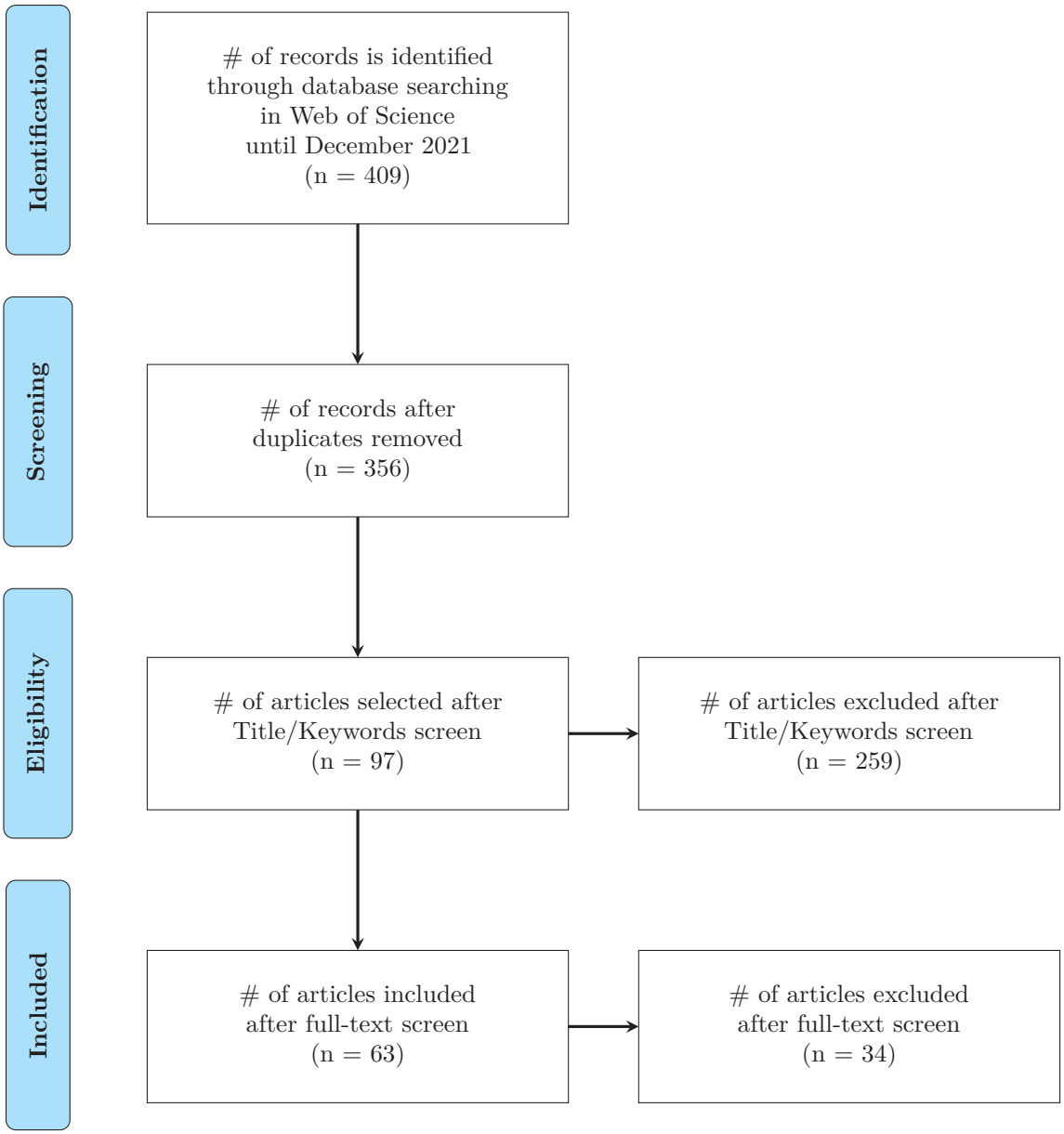


Figura 2.1: Diagrama de flux del procés de selecció d'articles seguint la metodologia PRISMA.

2.4 Procés de selecció

Dels 409 resultats, 53 obres van ser eliminades a causa que eren resultats duplicats obtinguts en diferents consultes. Després d'això, els criteris d'inclusió i exclusió es van aplicar llegint manualment el títol i les paraules clau de cada article. Això va donar lloc a 97 obres seleccionades. D'aquestes obres, se'n van analitzar els seus resums per confirmar que eren adequades per a aquest estat de l'art. Després de la lectura dels resums, 34 obres van ser descartades, cosa que significa que 63 obres van ser finalment seleccionades per a una anàlisi completa. D'aquestes 63 obres, 19 són publicacions a conferències i 44 són publicacions a revistes. La Figura 2.1 mostra un diagrama de flux segons la metodologia PRISMA que il · lustra el nombre d'articles que s'han tingut en consideració per a construir aquest estat de l'art.

Els articles seleccionats van ser categoritzats per tema per facilitar el procés de revisió. Concretament, es van considerar les següents categories:

- Obres que fan servir tècniques de DL per a la classificació acústica: sobre un 45%.
- Obres que fan servir tècniques de ML per a la classificació acústica: sobre un 27%.
- Obres que fan servir classificació multi-etiqueta (polifònica): 8%
- Obres que estudien paisatges sonors urbans o la percepció de so: sobre un 20%.
- Obres que executen els seus algoritmes en WASN o en proposen el seu ús: sobre un 17%.

Tingueu en compte que l'addició de tots els percentatges no és del 100% perquè algunes de les obres van ser assignades a més d'una categoria.

2.5 Anàlisi dels resultats i estat de l'art

Aquesta subsecció presenta una anàlisi de les obres que han estat seleccionades, avaluades i classificades com a estat de l'art de la tesi.

El *pipeline* típic d'un sistema classificador automàtic d'esdeveniments acústics conté els mòduls i aplica les tècniques de classificació detallades en el treball de Mesaros et al. publicat a (Mesaros et al. 2021). La manera més comuna d'aproximar-se a un problema de classificació o de detecció d'esdeveniments sonors és aplicar *aprenentatge supervisat* (Mesaros et al. 2021), que té com a objectiu classificar els esdeveniments acústics mitjançant la creació d'un model a partir de mostres acústiques anotades. L'etapa de classificació es pot dur a terme mitjançt tècniques tradicionals ML (com ara Gaussian Model Mixture (GMM), Hidden Markov Models o Models Ocults de Markov (HMM) o Support Vector Machine o Màquina de Vectors de Suport (SVM)) o mitjançant tècniques de DL.

2.5.1 Història del Deep Learning

Els orígens de DL es remunten als anys 1940 i 1950, quan el perceptró va ser introduït per primera vegada per Frank Rosenblatt el 1958 (Rosenblatt 1957) partint del treball de Warren McCulloch i Walter Pitts (Fitch 1944). No obstant això, un temps més tard, aquest algoritme va ser criticat: Marvin Minsky i Seymour Papert van publicar el 1969 un llibre explicant les limitacions del perceptró titulat "Perceptrons: una introducció a la geometria computacional"(Minsky i Papert 2017). Un dels temes més debatuts que els autors van presentar en el llibre va ser la dificultat que una xarxa neuronal tindria per calcular una simple operació XOR (OR-exclusiva). Van afirmar que utilitzant l'algorisme de Rosenblatt, l'operació no es podia resoldre, ja que requeriria múltiples capes de perceptrons.

El 1974, Paul Werbos va estudiar en el seu doctorat l'aplicació de l'algorisme de retropropagació en les xarxes neuronals (Werbos 1974), fent possible la creació de xarxes neuronals multicapa. Aquesta tècnica no va guanyar popularitat fins al 1986, quan David

Rumelhart, Geoffrey Hinton i Ronald Williams van publicar una obra descrivint la metodologia de l'algorisme i abordant els problemes debatuts per l'obra de Minsky (Rumelhart et al. 1986; Rumelhart et al. 1985).

Una vegada que la clau de com entrenar les xarxes neuronals multicapa es va fer pública, les primeres aplicacions d'aprenentatge profund van començar a aparèixer. Una de les contribucions més rellevants d'aquell temps es va produir el 1989, quan Yann LeCun et. va aplicar una Convolutional Neural Network o Xarxa Neuronal Convolucional (CNN) per a la classificació de dígits escrits a mà (LeCun et al. 1989). En aquella època, tot i que no eren tan populars, altres aplicacions van començar a utilitzar tècniques de DL. Per exemple, el 1988, investigadors com Lewis (Lewis 1988) i Todd (Todd 1988) van proposar l'ús de xarxes neuronals per a la composició automàtica de música. El principal inconvenient del DL en aquell moment era la quantitat de temps que es trigava a l'hora d'entrenar els models, ja que la tecnologia que estava disponible tenia grans limitacions —en comparació amb la capacitat de computació de les Graphics Processing Unit o Unitat de Processament de Gràfics (GPU) que hi ha disponibles actualment. Com que la tecnologia ha evolucionat molt en els últims anys, avui en dia, i malgrat els inconvenients que pot tenir el DL (requereix grans quantitats de dades i capacitats de computació), aquest s'ha convertit en una tendència popular a la investigació a causa dels resultats excepcionals de taxa d'encert que aconsegueix.

2.5.2 *Baseline* per a la classificació acústica

En general, els fluxos de treball més comuns per abordar un problema de ML o DL en qualsevol camp (no necessàriament en el domini d'àudio) són els que es mostren a la Figura 2.2. Com es pot observar, mentre que els treballs de ML requereixen un gran esforç en el procés de *feature engineering* (és a dir, seleccionar les característiques més convenients per a cada problema en particular), els treballs de DL ténen com a objectiu saltar-se aquesta part i utilitzar les dades en cru com a entrada del model, esperant que aquest extregui automàticament les característiques més convenients per a l'aprenentatge de la xarxa neuronal. Tanmateix, aquest no és necessàriament el cas (o, almenys, encara) en el domini de l'àudio. En l'estat de l'art de l'actualitat, una gran quantitat de problemes de detecció i classificació d'àudio es resolen mitjançant l'ús d'una xarxa neuronal que pren com a entrada els espectrogrames de l'àudio. D'aquesta manera, en alguns problemes de classificació del domini de l'àudio, el DL encara pot requerir d'un procés de selecció de característiques per entrenar xarxes neuronals. Els detalls sobre com es duu a terme aquest procés i algunes contribucions rellevants en el camp s'expliquen a laSecció 2.5.3 i la Secció 2.5.4.

2.5.3 Machine learning per a la classificació acústica

Un procés adequat de selecció de característiques és crucial quan s'utilitzen algoritmes tradicionals ML. Aquestes característiques poden contenir informació sobre el contingut en freqüència dels fitxers d'àudio, la seva evolució temporal o una barreja entre ambdós.

Les característiques bàsiques o més comunament utilitzades en aquest tipus de problemes són els Mel Frequency Cepstral Coefficients o Coeficients Cepstrals de Freqüència Mel (MFCC),

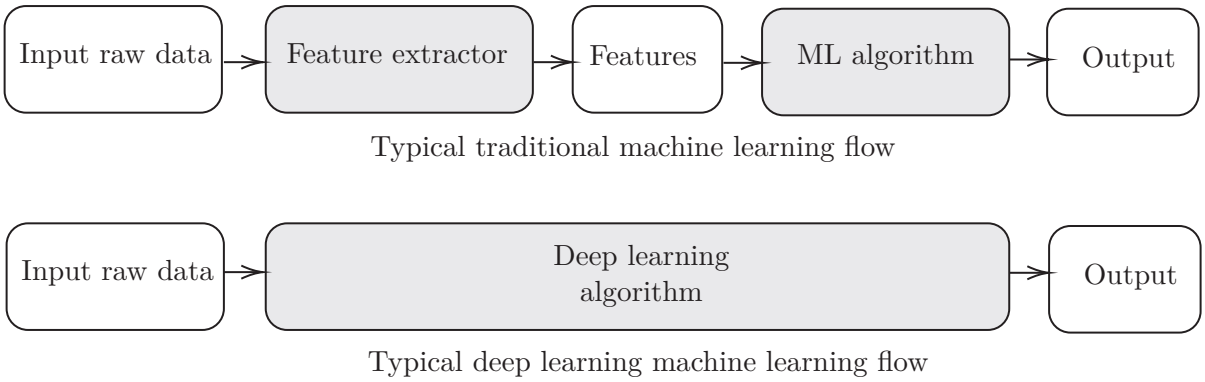


Figura 2.2: Fluxos de treball típics per a l'aprenentatge automàtic i l'aprenentatge profund . Inspirat en la presentació de Vivek Kumar titulada *Artificial Intelligence in Audio Event* (https://www.youtube.com/watch?v=388AZ2ujM9w&t=208s, accessed on 29 December 2021).

que ténen com a objectiu caracteritzar els arxius d'àudio emulant la percepció auditiva humana. Normalment, aquestes característiques es combinen amb altres com Linear Prediction Cepstrum Coefficients o Coeficients Cepstrals de Predicció Lineal (LPCC) o el Zero Crossing Rate o Taxa de Creuament per Zero (ZCR) entre altres (Dave 2013; Ito i Donaldson 1971).

Per exemple, a (Giannakopoulos et al. 2015), els autors presenten un sistema que pren com a objectiu estimar la qualitat del paisatge sonor (tant en entorns naturals com urbans) mitjançant l'anàlisi de l'àudio. Concretament, per dur a terme l'estimació de la qualitat del paisatge sonor, el sistema utilitza 68 característiques per cada fragment de 15 segons d'àudio. Aquestes característiques inclouen les *features* abans citades (MFCC i ZCR) combinades amb l'energia a curt termini del fragment, l'entropia de l'energia, el centroide espectral, l'entropia espectral, el flux i *roll-off* i les característiques de la crominància. Una vegada que les característiques es calculen, quatre regressors SVM diferents estimen tres nivells de context i el nivell de qualitat del paisatge sonor.

De la mateixa manera, a (Noviyanti et al. 2019), els autors pretenen predir el paisatge sonor urbà fent servir un conjunt de característiques acústiques. Específicament, la predicció es duu a terme utilitzant coeficients MFCC juntament amb paràmetres d'ecologia acústica. Amb tots aquests paràmetres, el treball prediu la percepció de relaxació, dinàmica i comunicació en un paisatge sonor donat. Els autors conclouen dels resultats obtinguts que els MFCC són millors característiques que la mètrica d'ecologia acústica per a la interpretació del seu model, que es basa en un regressor logístic binari.

Es poden trobar altres exemples a (Tsalera et al. 2020) o a (Lojka et al. 2014). En el primer cas, els autors també combinen característiques per a la classificació. Utilitzen 8 característiques temporals (incloent-hi ZCR, 11 característiques espectrals i 4 característiques perceptives com els MFCC). En aquest cas, l'algorisme de classificació seleccionat és un simple K-Nearest Neighbors o K-Veïns Propers (KNN), i els resultats es discuteixen quan s'utilitzen diferents nombres de veïns (1 a 3) i diferents mètriques de distància (Euclídea, Chebyshev i Cosinus). L'objectiu d'aquest treball és classificar esdeveniments de 9 categories diferents que es produeixen en entorns urbans. En el segon cas, les característiques utilitzades són

MFCC juntament amb Frequency Bank Coefficients o Coeficients de Bancs de Freqüències (FBC) i Mel-Spectral Coefficients o Coefficients Espectrals Mel (MSC). Per a la classificació, utilitzen un procés de descodificació Viterbi modificat juntament amb Weighted Finite-State Transducers (WFST) i un HMM.

Les tècniques de clustering també s'utilitzen àmpliament en el domini d'àudio. Per exemple, a (Pita et al. 2021), els autors utilitzen un algorisme ML no supervisat (és a dir, no requereix dades etiquetades per entrenar el model) per a crear clústers de la ciutat de Barcelona i, així, detectar zones properes a carreteres urbanes, àrees residencials i àrees d'oci. Tot i que aquest treball no realitza la classificació d'esdeveniments acústics, mostra una metodologia interessant per segmentar la ciutat segons els nivells de soroll presents en cada àrea. En aquest cas, només s'utilitzen com a característiques els nivells equivalents de soroll. L'algorisme d'agrupació ML utilitzat en l'obra és el K-means.

Un altre treball que utilitza el nivell de soroll (processat en diverses característiques) com a entrada per a un model per avaluar un cert paisatge sonor és el presentat a (Torija et al. 2014). En aquest treball, els algoritmes utilitzats per fer l'avaluació del paisatge sonor són una SVM i un Sequential Minimal Optimization o Optimització Mínima Seqüencial (SMO). Els resultats mostren que SMO supera a SVM quan es realitza la tasca de la classificació de paissatge sonor.

Com es pot veure, l'estudi dels paisatges sonors per mitjà de tècniques de ML es pot veure des de diferents punts de vista: agrupant diferents àrees d'una ciutat depenent dels seus nivells de soroll, correlacionant diferents esdeveniments de soroll amb la percepció de la població, o detectant automàticament esdeveniments acústics que ocorren en un entorn urbà per saber quines són les àrees més contaminades acústicament. Un gran projecte que va considerar diversos punts de vista en dues àrees diferents (una àrea urbana i una zona suburbana) va ser el projecte LIFE+ DYNAMAP. Concretament, en aquest projecte, els investigadors van desenvolupar una WASN de baix cost per supervisar dues àrees a gran escala a les ciutats de Milà i Roma utilitzant un mapa acústic dinàmic actualitzat en temps real. Durant el projecte, una vegada que es va desplegar la WASN, es va dur a terme una llarga campanya d'enregistrament que va permetre una recopilació massiva de dades (Alsina-Pagès et al. 2019). Algunes d'aquestes dades van ser posteriorment etiquetades manualment per experts, que van permetre (1) estudiar l'impacte dels diferents esdeveniments acústics a la població (Alías et al. 2020) i (2) el desenvolupament d'un algorisme que detecta i diferencia esdeveniments acústics anòmals respecte al soroll de trànsit de carretera (Socoró et al. 2017; Alías et al. 2018). Per a la classificació, els autors utilitzen MFCC com a característiques i GMM com a model de classificació.

Una altra tècnica que s'ha utilitzat en el camp és la *baf-of-features* (Grzeszick et al. 2017). Per exemple, a (Grzeszick et al. 2017), els autors utilitzen MFCC i GammaTone Cepstral Coefficients o Coeficients Cepstrals de Tons Gamma (GTCC) com a conjunt subjacent de característiques i els quantifiquen respecte a un determinat *codebook* per generar una *baf-of-features*. La obra també mostra com es pot millorar la robustesa dels models fusionant dades acústiques de múltiples canals. Un avantatge d'aquest enfocament és que no requereix altes

capacitats computacionals, cosa que significa que l'algorisme pot córrer sobre maquinari de baix cost (en termes de computació) o sobre entorns en línia.

En algunes altres obres, les tècniques de ML es barregen amb tècniques de processament de senyals i la classificació d'esdeveniments acústics es divideix en diferents capes de classificació. Per exemple, a (Luitel et al. 2016), els autors conceben un classificador de dues capes per categoritzar els esdeveniments urbans procedents principalment de vehicles (per exemple, motor d'autobús, clàxon d'autobús, clàxon d'automòbil i xiulet). Una primera capa divideix els esdeveniments acústics en dues classes amb tècniques de processament de senyals (mirant l'espectre de freqüències i filtrant a les freqüències desitjades) i una segona capa finalment classifica l'esdeveniment acústic. De nou, el classificador utilitza les característiques de base MFCC com a entrades per als models. També avalúen diversos classificadors, per exemple una Artificial Neural Network o Xarxa Neuronal Artificial (ANN), un Naive Bayes (NB), un Decision Tree o Arbre de Decisió (DT) i un Random Forest o Bosc Aleatòri (RF). Els seus resultats mostren que l'ús de dues capes millora els resultats de la classificació respecte a un sistema bàsic d'una capa.

No obstant això, no tots els problemes de classificació fan servir les característiques MFCC per a la classificació. A (Salamon i Juan Pablo Bello 2015a) i (Salamon i Juan Pablo Bello 2015b), els autors demostren que és possible obtenir millors resultats de classificació (només en termes de taxa d'encert) quan s'utilitza un aprenentatge de funcions no supervisades a partir d'espectrogrames-mel 2D i una tècnica de dispersió. L'aplicació d'una transformada de dispersió permet caracteritzar la dinàmica temporal a curt termini capturada per trossos d'espectrogrames 2D amb l'avantatga afegida de ser invariant en quant a fase. Això és un avantatge, ja que són capaços de caracteritzar senyals que varien en el temps sobre les finestres relativament llargues en comparació amb altres mètodes com el càlcul de MFCC.

De la mateixa manera, a (Waldekar i Saha 2020), els autors proposen l'ús de la transformada *wavelet*, atès que pot variar en longitud i, per tant, és convenient a l'hora d'extreure les característiques d'àudio ambiental. Com que el soroll ambiental pot tenir contingut de freqüència superposat i també un rang de freqüència més ampli (en comparació amb altres camps com el reconeixement de la parla), la transformada *wavelet* a escala mel presentada en el treball supera un sistema de classificació basat en els clàssics MFCC.

En resum, en l'estat de l'art de la detecció d'esdeveniments acústics i l'avaluacio del paissatge sonor s'han utilitzat ampliament diferents tècniques de ML. A més, quan s'utilitzen aquests algorismes, els MFCC són les característiques clàssiques, però normalment es combinen amb altres paràmetres acústics per aconseguir millors resultats de classificació.

2.5.4 Deep learning per a la classificació acústica

En els darrers anys, i a causa del ràpid desenvolupament de la tecnologia, el DL ha guanyat popularitat en el camp de la classificació d'esdeveniments acústics. En la seva majoria, les grans xarxes de DL han estat entrenades per classificar esdeveniments acústics en màquines amb altes capacitats de computació. Tanmateix, aquest no és el cas de tots els problemes de classificació de DL. Algunes xarxes petites han estat concebudes per classificar esdeveniments

acústics fins i tot en dispositius de baix cost (en termes de capacitats computacionals) com per exemple telèfons mòbils de manera eficient (Stowell 2021).

A més a més, en comptes de classificar els esdeveniments acústics, algunes obres se centren en classificar les escenes acústiques (Acoustic Scene Classification o Classificació d'Escena Acústica (ASC)). La diferència entre la detecció acústica d'esdeveniments i la classificació acústica d'escenes és que, mentre que la primera se centra en l'assignació d'una etiqueta semàntica a un esdeveniment acústic concret procedent d'una font de soroll, la segona assigna una etiqueta que es refereix a l'entorn en el qual es va enregistrar un audio. En aquest camp, és molt comú utilitzar també Deep Neural Network o Xarxa Neuronal Profunda (DNN). Tanmateix, no només les CNN són populars en aquest camp. Per exemple, a l'obra de (Singh et al. 2021), els autors utilitzen una xarxa neuronal prototípica per obtenir un *embedding space* per a la tasca d' ASC. La hipòtesi darrere de les xarxes prototípices és que existeix un *embedding space* en el qual els punts s'agrupen al voltant d'un únic prototip de representació per a cada classe (Snell et al. 2017). No obstant això, tot i que la hipòtesi és prometedora, aquest tipus de xarxes normalment obtenen valors moderats de taxa d'encert i haurien de ser estudiades i més desenvolupades en el futur.

A continuació, s'expliquen alguns projectes de l'estat d'art sobre classificació d'esdeveniments acústics per a diferents aplicacions amb DL.

En el treball (Genaro et al. 2010), els autors utilitzen 25 característiques per entrenar una ANN per predir el nivell de soroll en un entorn urbà. En el seu estudi, els autors comparen els resultats predits per la ANN i els resultats obtinguts quan s'aplica Principal Component Analysis o Anàlisi de Components Principals (PCA) amb l'objectiu de simplificar el model. Tot i que els resultats són pitjors després d'aplicar PCA, els autors afirmen que són acceptables.

Una obra notable és la publicada a (Lopez-Ballester et al. 2019; Lopez-Ballester et al. 2020). La seva aplicació se centra en l'avaluació de la molèstia que produeixen els sons utilitzant DL. Concretament, fan servir una CNN que és capaç de predir la molèstia psicoacústica utilitzant com a entrades senyals d'àudio crues. Les conclusions dels resultats que obtenen són que la seva xarxa és capaç de predir més ràpid que els sistemes convencionals la molèstia psicoacústica, tot i mantenint una alta precisió. Així, el desplegament de la seva xarxa és adequat per a dispositius IoT.

Les CNNs també s'han utilitzat àmpliament per classificar esdeveniments acústics per monitoritzar o supervisar l'estat de la biodiversitat. Per exemple, a (Morgan i Braasch 2021), s'han recopilat, analitzat i classificat dades de l'estat de Nova York (Estats Units) per calcular la riquesa i distribució de les espècies a partir d'algunes pseudoespècies. Les entrades de la xarxa neuronal són els espectrogrames dels senyals d'àudio. A més, el treball estudia la correlació entre els esdeveniments acústics i altres paràmetres abiòtics com la temperatura o les condicions meteorològiques. Un altre exemple es pot trobar a (Nanni et al. 2021). En aquest cas, els autors utilitzen un conjunt de CNNs per classificar esdeveniments acústics de diferents conjunts de dades (vocalitzacions d'ocells, sons de gat o sons ambientals). Els autors afirmen que el seu conjunt de classificadors es pot entrenar amb diferents conjunts de dades i arribar a les taxes d'encert de l'estat de l'art.

Al treball (Mushtaq i Su 2020), els autors també utilitzen CNNs per a la classificació acústica, en aquest cas per a la classificació de so ambiental. Com a característiques, els autors utilitzen l'espectrograma mel juntament amb els paràmetres MFCC. En aquest treball, els autors destaquen la importància d'utilitzar tècniques d'augment de dades per millorar els resultats de classificació. L'augment de dades, en aquest context, es refereix a l'ús de tècniques per augmentar la mida de les dades d'entrenament, afegint més mostres generades o bé modificant les dades del conjunt de dades o bé creant noves dades a partir de les mostres existents.

La importància de la selecció de les dades d'entrada a la CNN s'ha avaluat en algunes obres del camp. Per exemple, a (Zhou et al. 2017). Els seus experiments conclouen que la millor classificació dels sons urbans s'aconsegueix normalment quan els espectrogrames d'entrada tenen una resolució de temps moderada. A més, la normalització de les dades d'entrada quan s'utilitzen espectrogrames és un tema que encara està obert a discussió. Per exemple, al treball (Ick i McFee 2021), els autors exploren els diferents paràmetres de Per-Channel Energy Normalization o Normalització d'Energia Per Canal (PCEN) (que és un procediment adaptatiu que s'ha demostrat que és útil en alguns problemes de classificació d'àudio (Lostanlen et al. 2018)) i proposen un enfocament multi-rate PCEN per millorar els resultats de la classificació.

A causa de la gran quantitat de dades necessàries per entrenar els models DL, l'augment de dades s'ha utilitzat en la majoria de les obres DL en el camp. A part del treball de (Mushtaq i Su 2020), aquesta tècnica ha estat àmpliament utilitzada per la comunitat en els últims anys (Davis i Suresh 2018; Shah et al. 2019; Shen et al. 2020; Nanni et al. 2021; Dinkel et al. 2021).

Es poden trobar més obres que apliquen la classificació d'àudio utilitzant CNNs a (Sang et al. 2018; AbeBer et al. 2018; Bai et al. 2019; Phan et al. 2019; Fairbrass et al. 2019; Cao et al. 2019; Shen et al. 2020; Ciaburro 2020; Ciaburro i Iannace 2020). Mentre que algunes de les obres utilitzen els espectrogrames (o una fusió entre espectrogrames i altres característiques acústiques) com a entrades per a les seves xarxes, les altres utilitzen dades crues d'àudio.

A més, mentre que la majoria dels models DL són massa grans per ser desplegats en dispositius de baix cost, s'han fet alguns esforços en el camp per estudiar estratègies de desplegament de models sobre aquest tipus de dispositius. Aquest és el cas, per exemple, del projecte presentat en (Arce et al. 2021). L'obra presenta una WASN que supervisa àrees urbanes i reconeix un grup concret d'esdeveniments acústics. Els nodes de la xarxa estan formats per una Rasperry Pi com a unitat de computació, i l'algorisme de classificació és una CNN juntament amb una fase de predetecció que és capaç de diferir tres esdeveniments rellevants de trànsit i activa la CNN només quan es produeix un dels tres esdeveniments rellevants. D'aquesta manera, utilitzant l'etapa de predetecció, els autors són capaços de reduir l'ús de la Central Processing Unit o Unitat de Processament Central (CPU) del seu dispositiu de computació per un factor de 6.

Finalment, cal tenir en compte que, en ambients urbans, és comú trobar esdeveniments acústics que ocorren simultàniament (també coneguts com a esdeveniments polifònics). Reconeixe'ls és una tasca difícil, ja que alguns esdeveniments tenen un nivell acústic més alt que els altres, i a més a més tenen una durada i estructura diferents. Diverses obres del camp

tenen com a objectiu reconèixer diversos esdeveniments al mateix temps. Normalment, aquesta tasca es fa aplicant un llindar manual a l'última capa de la DNN per decidir si l'esdeveniment està present en el fragment acústic que s'està classificant o no. No obstant això, hi ha obres que elaboren diferents estratègies per aconseguir la classificació multietiqueta. Per exemple, en (Xia et al. 2018), els autors utilitzen un model de regressió multivariable i donen un nivell de confiança a cada segment d'àudio. Els seus resultats mostren que, d'aquesta manera, la taxa d'encert de classificiació és més alta.

En una obra diferent (Pankajakshan et al. 2019), els autors proposen un model que té com a objectiu millorar la localització temporal dels esdeveniments sonors utilitzant una combinació de dos models. El primer model prediu quins esdeveniments de so són presents en cada fotograma, i el segon prediu si un esdeveniment de so és present o no en un marc acústic. Els models conjunts donen lloc a taxes d'encert de classificació més altes que una implementació separada de cadascun d'ells.

Un altre treball que gestiona les dades multi-etiqueta és el presentat a (He et al. 2020). Per avaluar el procés de classificació multi-etiqueta, els autors utilitzen una estructura de multi-activació sigmoide-sparsemax.

Altres estudis que tracten les dades polifòniques o multi-etiqueta es poden trobar a (Xia et al. 2020; Gontier et al. 2021; Luo et al. 2021). En el primer treball (Xia et al. 2020), els autors utilitzen la posició de l'esdeveniment dins d'un segment d'àudio complet (tasca 1) i la posició d'un fragment concret dins d'un esdeveniment d'àudio (tasca 2) per al desenvolupament d'un enfocament d'aprenentatge multitasca. Els resultats d'un conjunt de dades monofòniques i un conjunt de dades polifòniques confirmen que el seu enfocament aconsegueix millors resultats de classificació en comparació amb el *baseline* d'aquests conjunts de dades respectius. En el segon treball (Gontier et al. 2021), els autors utilitzen la síntesi de conjunts d'entrenament polifònics per millorar els resultats de classificació en comparació amb un mètode d'aprenentatge autosupervisat. La síntesi de conjunts d'entrenament polifònics consisteix en anotar un petit corpus d'esdeveniments acústics d'interès, que després es barregen automàticament a l'atzar per formar un corpus més gran d'escenes polifòniques. En el treball, els autors afirmen que l'origen geogràfic dels esdeveniments acústics en la síntesi de conjunts d'entrenament té un gran impacte en els resultats de la classificació. Finalment, a (Luo et al. 2021), els autors utilitzen un model que combina una xarxa neuronal de càpsula (CapsNet) i una xarxa neuronal recurrent. Com a entrades al seu model, els autors utilitzen un mètode d'agregació de característiques incloent-hi MFCC i les característiques log-mel. També han implementat un sistema al món real capaç de detectar els esdeveniments acústics en ambients urbans.

2.5.5 Xarxes de Sensors Acústics Sense Fils desplegades en la societat moderna

Aquesta subsecció ofereix una visió general d'algunes WASN que actualment estan desplegades en diferents entorns: àrees urbanes o suburbanes. A part del ja esmentat projecte DYNAMAP, (Alsina-Pagès et al. 2019), durant el qual es van desplegar sensors acústics en diferents àrees

de Roma i Milà, altres obres de tot el món han dut a terme enfocaments similars.

Per exemple, al Regne Unit, el projecte DREAMsys (Distributed Remote Environmental Array & Monitoring System) (Barham et al. 2010) ha desenvolupat un sistema basat en fer mesures de soroll per a l'estudi del paissatge sonor i realitzar mapes acústics utilitzant sensors distribuïts. El maquinari dels seus sensors inclou un micròfon MEMS protegit amb una capa impermeable, una unitat de computació que calcula el nivell equivalent de soroll, un mòdem GSM, bateries que poden durar fins a 15 dies i un trípode que permet la mobilitat dels sensors.

A Itàlia, en un projecte a curt termini anomenat SENSEable (Nencini et al. 2012) va desplegar una WASN a la ciutat de Pisa per mesurar els nivells de soroll en temps real en diferents llocs de la ciutat.

En l'escenari d'ús d'aquesta Tesi, Barcelona, també s'han desplegat diversos nodes de detecció d'alta qualitat (classe I) (Farrés 2015). L'objectiu de la WASN desplegada a Barcelona és avaluar els nivells de soroll en zones sorolloses, quantificar la reducció del soroll quan s'apliquen determinats plans d'acció, (3) actualitzar un mapa de soroll en temps real i (4) identificar fonts de soroll i avaluar-les. No obstant això, aquesta avaluació encara es fa manualment, no s'ha desplegat cap sistema de classificació automàtic. Els nivells de soroll es poden veure en temps real a la plataforma SENTILO, que subministra informació sobre la situació acústica de la ciutat, però també mostra valors recollits per diferents tipus de sensors (per exemple, mostra informació meteorològica). El mapa en temps real es pot veure a https://connecta.bcn.cat/connecta-catalog-web/component/map (últim accés el 30 de desembre de 2021).

Una altra WASN es pot trobar al Canadà, en el marc del projecte UrbanSense (Rainham 2016). En aquest cas, la xarxa no és només responsable de controlar les dades acústiques (LAeq), sinó que també té en compte altres paràmetres d'interès com la quantitat de diòxid de carboni o monòxid de carboni a l'aire, velocitat i direcció del vent, la temperatura, la humitat relativa i els nivells de les precipitacions.

A París (França), l'organització BUITPARIF ha dut a terme un projecte que ha donat lloc al disseny i la patent d'un dispositiu de control de soroll anomenat MEDUSA (C. Mietlicki i F. Mietlicki 2018), que combina quatre micròfons i un sistema òptic, de manera que és possible representar nivells de soroll en 360º. Aquestes dades es projecten sobre un mapa geogràfic que crea hexàgons de colors que permeten veure els nivells de soroll de la zona.

El projecte LIFE Monza (Bartalucci et al. 2018), un projecte LIFE que va durar fins a 2020, va desenvolupar un mètode per a la identificació i la gestió de les zona de baixes emissions de soroll de la ciutat, que són zones urbanes subjectes a restriccions de trànsit per mitigar l'impacte del soroll a la població. Concretament, el projecte va desplegar una prova pilot a la ciutat de Monza (al nord d'Itàlia). La implementació física del projecte incloïa l'ús de sensors de baix cost i una interfície mitjançant una pàgina web.

Fora d'Europa, el control del soroll també es considera una qüestió important que ha de tenir-se en compte. Per exemple, el projecte SONYC (Sound Of New York City) (Juan P Bello et al. 2019), ha desplegat 55 sensors acústics de baix cost a la ciutat de Nova York. Cada sensor està compost per un nucli de sensor (Raspberri Pi + antena WiFi) i un mòdul de

detecció acústica (basat en un micròfon MEMS i un microcontrolador) (Mydlarz et al. 2019). A més de controlar el nivell de soroll, aquests sensors són capaços de classificar esdeveniments en temps real dins d'una taxonomia limitada.

Finalment, a Espanya, a més de la xarxa ja esmentada desplegada a Barcelona, diversos projectes en diferents ciutats han abordat el repte del control del soroll mitjançant sensors acústics. Per exemple, a Màlaga (López et al. 2020), per avaluar la qualitat de vida dels ciutadans, i com en algunes àrees el soroll d'oci excedeix els límits permesos per les regulacions actuals, es va desplegar un sub-conjunt de 8 sensors acústics com a part d'una subxarxa. Aquests sensors tenen com a objectiu obtenir diversos (86) paràmetres acústics en temps real per controlar el nivell de soroll en àrees problemàtiques. Per altra banda, a la capital del país, Madrid, s'han instal · lat 31 sonòmetres premium (Brüel & Kjaer, classe 1) amb un micròfon que permet desplegaments a l'aire lliure (Asensio et al. 2020). Per garantir resultats fiables, els sensors es calibren cada any d'acord amb la normativa. Aquests sensors s'instal · len en el sostre de les cabines de mesures de condicions ambientals, i s'encarreguen de controlar els nivells de soroll de les ubicacions seleccionades.

2.5.6 Oportunitats de recerca i comunitat

Per concloure aquest capítol, aquesta secció ofereix una visió general dels reptes actuals en el camp de la classificació d'esdeveniments acústics i els esforços que està fent la comunitat per aconseguir-los.

Com es pot veure en aquest estat de l'art, els investigadors de tot el món han estat esforçant-se per aconseguir millors resultats any rere any mitjançant l'ús de nous algoritmes ML, algorismes d'extracció de característiques, o una combinació d'ambdós. No obstant això, la majoria d'aquestes obres es centren en aconseguir la millor taxa de classificació possible, sense preocupar-se per la mida dels seus models o el maquinari necessari per realitzar la classificació. En realitat, en diferents treballs, és comú veure investigadors preocupats pel temps d'entrenament dels seus models, subestimant el temps d'inferència. Això es deu a les grans capacitats computacionals necessàries per a l'entrenament de models ML o, especialment, l'entrenament de models DL, que pot requerir fins i tot de mesos d'entrenament per obtenir un estat estable (depenent del maquinari que s'utilitza per a l'entrenament, la mida dels models i la quantitat de dades disponibles). No obstant això, per a les aplicacions del món real, el temps d'inferència és més important que el temps d'entrenament si l'objectiu final és proporcionar els resultats de classificació en temps real. Per aquesta raó, aquesta dissertació té com a objectiu contribuir a aquesta oportunitat de recerca en lloc de construir un model que aconsegueixi la millor taxa de classificació possible. En aquest sentit, aquesta dissertació equilibrarà el temps i la memòria d'inferència en el maquinari de baix cost proposat i les taxes d'encert de classificació.

En quant a la comunitat, per a impulsar l'estat de l'art, cada any s'organitza una competició d'investigació en la qual es proposen diversos reptes (generalment 5 o 6 tasques) que aborden diferents àmbits de la classificació acústica d'esdeveniments. La competició s'anomena Detection and Classification of Acoustic Scenes and Events o Detecció i Classificació

d'Escenes i Esdeveniments Acústics (DCASE) (http://dcase.community, accedit el 29 de desembre de 2021) i té com a objectiu promoure diversos conjunts de dades i sistemes de classificació, animant als investigadors a aplicar noves tècniques de classificació per guanyar la competició. Les diferents tasques poden incloure dades acústiques de diferents dominis com ara esdeveniments urbans, vocalitzacions d'animals o classificació d'escenes acústiques.

Referències

AbeBer, Jakob, Gotze, Marco, Kuhnlenz, Stephanie, Grafe, Robert, Kuhn, Christian, ClauB, Tobias, and Lukashevich, Hanna (Aug. 2018). 'A Distributed Sensor Network for Monitoring Noise Level and Noise Sources in Urban Environments'. In: *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*. Barcelona, Spain: IEEE, pp. 318–324.

Alías, Francesc, Alsina-Pagès, Rosa Ma, Orga, Ferran, and Socoró, Joan Claudi (July 2018). 'Detection of Anomalous Noise Events for Real-Time Road-Traffic Noise Mapping: The Dynamap's project case study'. In: *Noise Mapping* vol. 5, no. 1, pp. 71–85.

Alías, Francesc, Orga, Ferran, Alsina-Pagès, Rosa Ma, and Socoró, Joan Claudi (Jan. 2020). 'Aggregate Impact of Anomalous Noise Events on the WASN-Based Computation of Road Traffic Noise Levels in Urban and Suburban Environments'. en. In: *Sensors* vol. 20, no. 3, p. 609.

Alsina-Pagès, Rosa Ma, Orga, Ferran, Alías, Francesc, and Socoró, Joan Claudi (May 2019). 'A WASN-Based Suburban Dataset for Anomalous Noise Event Detection on Dynamic Road-Traffic Noise Mapping'. en. In: *Sensors* vol. 19, no. 11, p. 2480.

Amo Filvà, Daniel, Alier Forment, Marc, García Peñalvo, Francisco Javier, Fonseca Escudero, David, and Casany Guerrero, María José (2020). 'Privacidad, seguridad y legalidad en soluciones educativas basadas en Blockchain: Una Revisión Sistemática de la Literatura'. In: *RIED. Revista iberoamericana de educación a distancia* vol. 23, no. 2, pp. 213–236.

Arce, Pau, Salvo, David, Piñero, Gema, and Gonzalez, Alberto (Sept. 2021). 'FIWARE based low-cost wireless acoustic sensor network for monitoring and classification of urban soundscape'. en. In: *Computer Networks* vol. 196, p. 108199.

Asensio, César, Pavón, Ignacio, and De Arcas, Guillermo (2020). 'Changes in noise levels in the city of Madrid during COVID-19 lockdown in 2020'. In: *The Journal of the Acoustical Society of America* vol. 148, no. 3, pp. 1748–1755.

Bai, Jisheng, Chen, Chen, and Chen, Jianfeng (Nov. 2019). 'A Multi-feature Fusion Based Method For Urban Sound Tagging'. In: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Lanzhou, China: IEEE, pp. 1313–1317.

Barham, Richard, Chan, Martin, and Cand, Matthew (2010). 'Practical experience in noise mapping with a MEMS microphone based distributed noise measurement system'. In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Vol. 2010. 6. Institute of Noise Control Engineering, pp. 4725–4733.

Bartalucci, Chiara, Borchi, Francesco, Carfagni, Monica, Furferi, Rocco, Governi, Lapo, Lapini, Alessandro, Bellomini, Raffaella, Luzzi, Sergio, and Nencini, Luca (2018). 'The smart noise monitoring system implemented in the frame of the Life MONZA project'. In: *Proceedings of the EuroNoise*, pp. 783–788.

Bello, Juan P, Silva, Claudio, Nov, Oded, Dubois, R Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles, and Doraiswamy, Harish (2019). 'Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution'. In: *Communications of the ACM* vol. 62, no. 2, pp. 68–77.

Cao, Jiuwen, Cao, Min, Wang, Jianzhong, Yin, Chun, Wang, Danping, and Vidal, Pierre-Paul (Oct. 2019). 'Urban noise recognition with convolutional neural network'. en. In: *Multimedia Tools and Applications* vol. 78, no. 20, pp. 29021–29041.

Ciaburro, Giuseppe (Aug. 2020). 'Sound Event Detection in Underground Parking Garage Using Convolutional Neural Network'. en. In: *Big Data and Cognitive Computing* vol. 4, no. 3, p. 20.

Ciaburro, Giuseppe and Iannace, Gino (July 2020). 'Improving Smart Cities Safety Using Sound Events Detection Based on Deep Neural Network Algorithms'. en. In: *Informatics* vol. 7, no. 3, p. 23.

Dave, Namrata (2013). 'Feature extraction methods LPC, PLP and MFCC in speech recognition'. In: *International journal for advance research in engineering and technology* vol. 1, no. 6, pp. 1–4.

Davis, Nithya and Suresh, K (Dec. 2018). 'Environmental Sound Classification Using Deep Convolutional Neural Networks and Data Augmentation'. In: *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. Thiruvananthapuram, India: IEEE, pp. 41–45.

Dinkel, Heinrich, Wu, Mengyue, and Yu, Kai (2021). 'Towards Duration Robust Weakly Supervised Sound Event Detection'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 29, pp. 887–900.

Fairbrass, Alison J., Firman, Michael, Williams, Carol, Brostow, Gabriel J., Titheridge, Helena, and Jones, Kate E. (Feb. 2019). 'CityNet—Deep learning tools for urban ecoacoustic assessment'. en. In: *Methods in Ecology and Evolution* vol. 10, no. 2. Ed. by Isaac, Nick, pp. 186–197.

Farrés, Júlia Camps (2015). 'Barcelona noise monitoring network'. In: *Proceedings of the Euronoise*, pp. 218–220.

Fitch, Frederic B (1944). 'McCulloch Warren S. and Pitts Walter. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics, vol. 5, pp. 115–133'. In: *Journal of Symbolic Logic* vol. 9, no. 2.

Genaro, N., Torija, A., Ramos-Ridao, A., Requena, I., Ruiz, D. P., and Zamorano, M. (Oct. 2010). 'A neural network based model for urban noise prediction'. en. In: *The Journal of the Acoustical Society of America* vol. 128, no. 4, pp. 1738–1746.

Giannakopoulos, Theodoros, Siantikos, Georgios, Perantonis, Stavros, Votsi, Nefta-Eleftheria, and Pantis, John (July 2015). 'Automatic soundscape quality estimation using audio

analysis'. en. In: *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. Corfu Greece: ACM, pp. 1–9.

Gontier, Félix, Lostanlen, Vincent, Lagrange, Mathieu, Fortin, Nicolas, Lavandier, Catherine, and Petiot, Jean-Francois (2021). 'Polyphonic training set synthesis improves self-supervised urban sound classification'. In: *The Journal of the Acoustical Society of America* vol. 149, no. 6, pp. 4309–4326.

Grzeszick, Rene, Plinge, Axel, and Fink, Gernot A. (June 2017). 'Bag-of-Features Methods for Acoustic Event Detection and Classification'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 25, no. 6, pp. 1242–1252.

He, Kexin, Shen, Yuhan, Zhang, Wei-Qiang, and Liu, Jia (May 2020). 'Staged Training Strategy and Multi-Activation for Audio Tagging with Noisy and Sparse Multi-Label Data'. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, pp. 631–635.

Ick, Christopher and McFee, Brian (June 2021). 'Sound Event Detection in Urban Audio with Single and Multi-Rate Pcen'. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, pp. 880–884.

Ito, M and Donaldson, R (1971). 'Zero-crossing measurements for analysis and recognition of speech sounds'. In: *IEEE Transactions on Audio and Electroacoustics* vol. 19, no. 3, pp. 235–242.

Khan, Khalid S, Kunz, Regina, Kleijnen, Jos, and Antes, Gerd (2003). 'Five steps to conducting a systematic review'. In: *Journal of the royal society of medicine* vol. 96, no. 3, pp. 118–121.

LeCun, Yann, Boser, Bernhard, Denker, John S, Henderson, Donnie, Howard, Richard E, Hubbard, Wayne, and Jackel, Lawrence D (1989). 'Backpropagation applied to handwritten zip code recognition'. In: *Neural computation* vol. 1, no. 4, pp. 541–551.

Lewis, John Peter (1988). 'Creation by refinement: a creativity paradigm for gradient descent learning networks.' In: *ICNN*, pp. 229–233.

Lojka, Martin, Pleva, Matúš, Kiktová, Eva, Juhár, Jozef, and Čižmár, Anton (2014). 'EARTUKE: The Acoustic Event Detection System'. In: *Multimedia Communications, Services and Security*. Ed. by Junqueira Barbosa, Simone Diniz et al. Vol. 429. Cham: Springer International Publishing, pp. 137–148.

López, Juan Manuel, Alonso, Jesús, Asensio, César, Pavón, Ignacio, Gascó, Luis, and Arcas, Guillermo de (2020). 'A Digital Signal Processor Based Acoustic Sensor for Outdoor Noise Monitoring in Smart Cities'. In: *Sensors* vol. 20, no. 3, p. 605.

Lopez-Ballester, Jesus, Pastor-Aparicio, Adolfo, Felici-Castell, Santiago, Segura-Garcia, Jaume, and Cobos, Maximo (Oct. 2020). 'Enabling Real-Time Computation of Psycho-Acoustic Parameters in Acoustic Sensors Using Convolutional Neural Networks'. In: *IEEE Sensors Journal* vol. 20, no. 19, pp. 11429–11438.

Lopez-Ballester, Jesus, Pastor-Aparicio, Adolfo, Segura-Garcia, Jaume, Felici-Castell, Santiago, and Cobos, Maximo (Aug. 2019). 'Computation of Psycho-Acoustic Annoyance Using Deep Neural Networks'. en. In: *Applied Sciences* vol. 9, no. 15, p. 3136.

Lostanlen, Vincent, Salamon, Justin, Cartwright, Mark, McFee, Brian, Farnsworth, Andrew, Kelling, Steve, and Bello, Juan Pablo (2018). 'Per-channel energy normalization: Why and how'. In: *IEEE Signal Processing Letters* vol. 26, no. 1, pp. 39–43.

Luitel, Bibek, Murthy, Y. V. Srinivasa, and Koolagudi, Shashidhar G. (Aug. 2016). 'Sound event detection in urban soundscape using two-level classification'. In: *2016 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. Mangalore, India: IEEE, pp. 259–263.

Luo, Liyan, Zhang, Liujun, Wang, Mei, Liu, Zhenghong, Liu, Xin, He, Ruibin, and Jin, Ye (2021). 'A System for the Detection of Polyphonic Sound on a University Campus Based on CapsNet-RNN'. In: *IEEE Access* vol. 9, pp. 147900–147913.

Mesaros, Annamaria, Heittola, Toni, Virtanen, Tuomas, and Plumbley, Mark D (2021). 'Sound event detection: A tutorial'. In: *IEEE Signal Processing Magazine* vol. 38, no. 5, pp. 67–83.

Mietlicki, Christophe and Mietlicki, Fanny (2018). 'Medusa: a new approach for noise management and control in urban environment'. In: *Proceedings of the 11th European Congress and Exposition on Noise Control Engineering (Euronoise2018), Crete, Greece*, pp. 27–31.

Minsky, Marvin and Papert, Seymour A (2017). *Perceptrons: An introduction to computational geometry*. MIT press.

Morgan, M.M. and Braasch, J. (Mar. 2021). 'Long-term deep learning-facilitated environmental acoustic monitoring in the Capital Region of New York State'. en. In: *Ecological Informatics* vol. 61, p. 101242.

Mushtaq, Zohaib and Su, Shun-Feng (Oct. 2020). 'Environmental sound classification using a regularized deep convolutional neural network with data augmentation'. en. In: *Applied Acoustics* vol. 167, p. 107389.

Mydlarz, Charlie, Sharma, Mohit, Lockerman, Yitzchak, Steers, Ben, Silva, Claudio, and Bello, Juan Pablo (2019). 'The life of a New York City noise sensor network'. In: *Sensors* vol. 19, no. 6, p. 1415.

Nanni, Loris, Maguolo, Gianluca, Brahnam, Sheryl, and Paci, Michelangelo (June 2021). 'An Ensemble of Convolutional Neural Networks for Audio Classification'. en. In: *Applied Sciences* vol. 11, no. 13, p. 5796.

Nencini, Luca, De Rosa, Paolo, Ascari, Elena, Vinci, Bruna, and Alexeeva, Natalia (2012). 'SENSEable Pisa: A wireless sensor network for real-time noise mapping'. In: *Proceedings of the EURONOISE, Prague, Czech Republic*, pp. 10–13.

Noviyanti, Anastasia, Sudarsono, Anugrah Sabdono, and Kusumaningrum, Dian (2019). 'Urban soundscape prediction based on acoustic ecology and MFCC parameters'. In: Padang, Indonesia, p. 050005.

Pankajakshan, Arjun, Bear, Helen L., and Benetos, Emmanouil (Aug. 2019). 'Polyphonic Sound Event and Sound Activity Detection: A Multi-task approach'. In: *arXiv:1907.05122 [cs, eess]*. arXiv: 1907.05122.

Phan, Huy, Chén, Oliver Y., Koch, Philipp, Pham, Lam, McLoughlin, Ian, Mertins, Alfred, and De Vos, Maarten (Feb. 2019). 'Unifying Isolated and Overlapping Audio Event

Detection with Multi-Label Multi-Task Convolutional Recurrent Neural Networks'. In: *arXiv:1811.01092 [cs, eess, stat]*. arXiv: 1811.01092.

Pita, Antonio, Rodriguez, Francisco J., and Navarro, Juan M. (Aug. 2021). 'Cluster Analysis of Urban Acoustic Environments on Barcelona Sensor Network Data'. en. In: *International Journal of Environmental Research and Public Health* vol. 18, no. 16, p. 8271.

Rainham, D (2016). 'A wireless sensor network for urban environmental health monitoring: UrbanSense'. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 34. 1. IOP Publishing, p. 012028.

Rosenblatt, Frank (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J (1985). *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J (1986). 'Learning representations by back-propagating errors'. In: *nature* vol. 323, no. 6088, pp. 533–536.

Salamon, Justin and Bello, Juan Pablo (Aug. 2015a). 'Feature learning with deep scattering for urban sound analysis'. In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. Nice: IEEE, pp. 724–728.

Salamon, Justin and Bello, Juan Pablo (Apr. 2015b). 'Unsupervised feature learning for urban sound classification'. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, pp. 171–175.

Sang, Jonghee, Park, Soomyung, and Lee, Junwoo (Sept. 2018). 'Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms'. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. Rome: IEEE, pp. 2444–2448.

Shah, Sayed Khushal, Tariq, Zeenat, and Lee, Yugyung (Dec. 2019). 'IoT based Urban Noise Monitoring in Deep Learning using Historical Reports'. In: *2019 IEEE International Conference on Big Data (Big Data)*. Los Angeles, CA, USA: IEEE, pp. 4179–4184.

Shen, Yexin, Cao, Jiuwen, Wang, Jianzhong, and Yang, Zhixin (Jan. 2020). 'Urban acoustic classification based on deep feature transfer learning'. en. In: *Journal of the Franklin Institute* vol. 357, no. 1, pp. 667–686.

Singh, Shubhr, Bear, Helen L., and Benetos, Emmanouil (June 2021). 'Prototypical Networks for Domain Adaptation in Acoustic Scene Classification'. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, pp. 346–350.

Snell, Jake, Swersky, Kevin, and Zemel, Richard S (2017). 'Prototypical networks for few-shot learning'. In: *arXiv preprint arXiv:1703.05175*.

Socoró, Joan, Alías, Francesc, and Alsina-Pagès, Rosa (Oct. 2017). 'An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments'. en. In: *Sensors* vol. 17, no. 10, p. 2323.

Stowell, Dan (2021). *Computational bioacoustics with deep learning: a review and roadmap*. arXiv: 2112.06725 [cs.SD].

Todd, Peter (1988). 'A sequential network design for musical applications'. In: *Proceedings of the 1988 connectionist models summer school*, pp. 76–84.

Torija, Antonio J., Ruiz, Diego P., and Ramos-Ridao, Ángel F. (June 2014). 'A tool for urban soundscape evaluation applying Support Vector Machines for developing a soundscape classification model'. en. In: *Science of The Total Environment* vol. 482-483, pp. 440–451.

Tsalera, Eleni, Papadakis, Andreas, and Samarakou, Maria (Nov. 2020). 'Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm'. en. In: *Energy Reports* vol. 6, pp. 223–230.

Waldekar, Shefali and Saha, Goutam (Mar. 2020). 'Analysis and classification of acoustic scenes with wavelet transform-based mel-scaled features'. en. In: *Multimedia Tools and Applications* vol. 79, no. 11-12, pp. 7911–7926.

Werbos, Paul (1974). 'Beyond regression:" new tools for prediction and analysis in the behavioral sciences'. In: *Ph. D. dissertation, Harvard University*.

Xia, Xianjun, Togneri, Roberto, Sohel, Ferdous, and Huang, David (Apr. 2018). 'Confidence Based Acoustic Event Detection'. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB: IEEE, pp. 306–310.

Xia, Xianjun, Togneri, Roberto, Sohel, Ferdous, Zhao, Yuanjun, and Huang, Defeng (Mar. 2020). 'Multi-Task Learning for Acoustic Event Detection Using Event and Frame Position Information'. In: *IEEE Transactions on Multimedia* vol. 22, no. 3, pp. 569–578.

Zhou, Huan, Song, Ying, and Shu, Haiyan (Nov. 2017). 'Using deep convolutional neural network to classify urban sounds'. In: *TENCON 2017 - 2017 IEEE Region 10 Conference*. Penang: IEEE, pp. 3089–3092.

Chapter 2

State of the art

To conduct a complete state of the art for acoustic event detection in urban environments, a systematic literature review has been carried out. The main purpose of a systematic literature review is to gather the evidences and most relevance findings of a concrete research topic through a systematic process (Amo Filvà et al. 2020). Hence, the work that is carried out in this chapter aims for an exhaustive research on the topic being discussed in this dissertation (that is, acoustic event classification in urban environments).

2.1 Methodology

To perform a successful systematic literature review, the first step is to come up with a strategy that allows to select the most relevant publications and online information from the field to be studied (i.e., for this work, acoustic event detection and classification). For this purpose, the methodology explained in (Khan et al. 2003), that consists of five steps, has been followed.

1. **Framing questions for a review**: In this sense, the questions to be solved are the research questions stated in Capítol 1 from this document: RQ1, RQ2 and RQ3. From these research questions, the inclusion and exclusion criteria to select the most relevant works have been defined.
2. **Identifying relevant work**: For this purpose, the main database used to search information is the Web Of Science (WOS). This process is carried out in an iterative way, and the inclusion and exclusion criteria are applied to select only the appropriate works for the review. Also, this stage considers the Population, Intervention, Comparison, Outcome, Context o Població, Intervenció, Comparació, Resultats, Context (PICOC) criteria to give an appropriate answer to the research questions by limiting the scope of the review.
3. **Assessing the quality of studies**: To do so, the number of citations and media in which the information was published have been taken into account. Concretely, peer reviewed conferences, journals and technical reports have been considered as reliable data sources.
4. **Summarizing the evidence**: Obtaining a summary of the methodology and the key main conclusions from the work.
5. **Interpreting the findings**: Checking the hetereogenity of data and decide whether the summaries obtained can be trusted.

2.2 Inclusion and exclusion criteria

The concrete inclusion criteria that have been followed to include works to the review are:

Inclusion Criteria 1 o Criteris d'Inclusió 1 (IC1): The gathered works are applied to the field of acoustic event detection **AND**

Inclusion Criteria 2 o Criteris d'Inclusió 2 (IC2): The gathered works contain information about real-time acoustic signal processing OR low-cost acoustic wireless sensor network architectures **AND**

Inclusion Criteria 3 o Criteris d'Inclusió 3 (IC3): The works are written in English language **AND**

Inclusion Criteria 4 o Criteris d'Inclusió 4 (IC4): The works are published in peer reviewed conferences, journals or technical reports.

And the concrete exclusion criteria that have been followed to exclude works to the review are:

Criteris d'Exclusió 1 (CE1): The gathered works are not applied to the field of acoustic event detection **OR**

Criteris d'Exclusió 2 (CE2): The gathered works do not contain information about real-time acoustic signal processing OR low-cost acoustic wireless sensor network architectures **OR**

Criteris d'Exclusió 3 (CE3): The works are not written in English language **OR**

Criteris d'Exclusió 4 (CE4): The works are not published in peer reviewed conferences, journals or technical reports.

2.3 Queries

To search the information, several queries were formulated on the WOS, which provided the results shown in Taula 2.1:

Queries	Number of results
TS=(acoustic classification AND urban)	299
TS=(low cost device* AND acoustic event classification)	22
TS=(real-time classification AND acoustic events AND urban)	9
TS=(urban sound AND acoustic dataset)	51
TS=(multilabel classification AND acoustic event)	16
TS=(event classification AND urban soundscape)	12

Table 2.1: Queries and number of results formulated in WOS to gather information.

This means that, in total, the searching queries resulted in 409 results in the WOS. The research queries were executed at date of December 2021, hence works published later have not been included. However, these results had to be filtered before their processing, as they might contain duplicated results from different queries or they might not satisfy the inclusion and exclusion criteria described above.

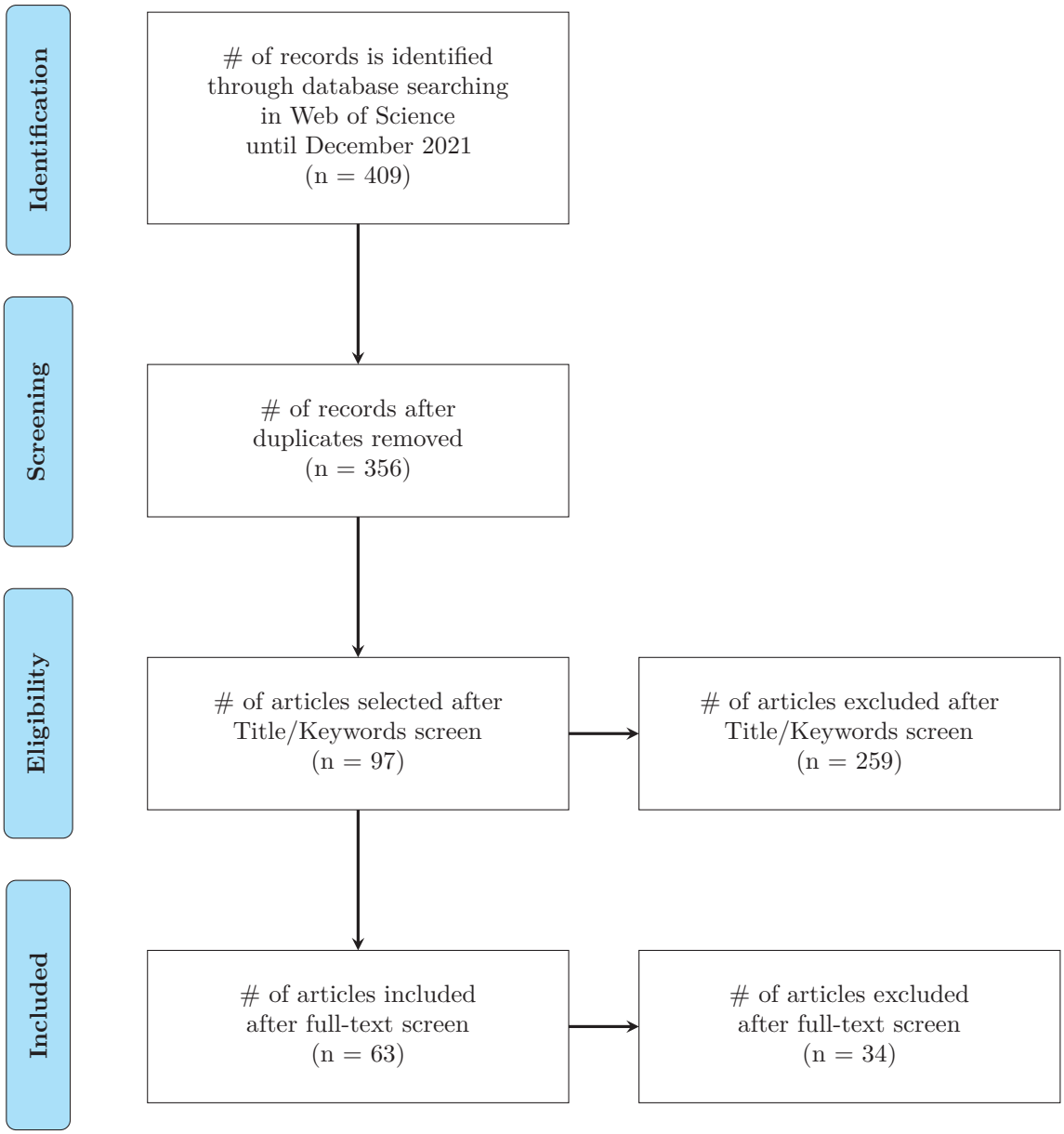


Figure 2.1: Flow diagram of the articles selection process following the PRSIMA methodology.

2.4 Selection process

From the 409 results, 53 works were removed due to that they were repeated results outputted from different queries. After that, the inclusion and exclusion criteria were applied by manually reading the title and keywords. This resulted in 97 selected works. From those works, the abstracts were fully analysed to confirm that they were suitable for this state of the art. After the abstract reading, 34 works were discarded, which means that 63 works were finally selected for a full analysis. From those 63 works, 19 are publications from conferences and 44 are publications from journals. Figura 2.1 shows a flow-diagram according to the PRISMA methodology that illustrates the amount of articles that have been taken into consideration to develop the state of the art.

The selected articles were later classified by topic to ease the reviewing process considering

the following categories:

- Works that use DL techniques for acoustic classification: about 45%.
- Works that use classical ML techniques for acoustic classification: about 27%.
- Works that consider multilabel (also known as polyphonic) classification: 8%
- Works that study urban soundscapes or sound perception: about 20%.
- Works that run their algorithms or propose the usage of WASN: about 17%.

Note that the addition of all the percentages is not 100% because some of the works were assigned to more than one category.

2.5 Analysis of the results and state of the art

This subsection presents an analysis of the works that have been selected, evaluated and classified as a state of the art of the dissertation.

The typical pipeline of an automatic classifier system for acoustic events contains the modules and applies the classification techniques detailed in the work by Mesaros et al. published in (Mesaros et al. 2021). The most common way of approaching a classification problem of sound event detection is to apply *supervised learning* (Mesaros et al. 2021), which aims at classifying acoustic events by creating a model from annotated acoustic samples. The classification stage can be carried out by traditional ML techniques (such as Gaussian Model Mixture (GMM), Hidden Markov Models o Models Ocults de Markov (HMM) or Support Vector Machine o Màquina de Vectors de Suport (SVM)) or, lately, by DL models.

2.5.1 History of Deep Learning

The origins of DL date back to 1940's and 1950's, when the perceptron was first introduced by Frank Rosenblatt in 1958 (Rosenblatt 1957) over the work of Warren McCulloch and Walter Pitts (Fitch 1944). However, this algorithm was later criticised. For instance, Marvin Minsky and Seymour Papert published in 1969 a book explaining the limitations of the perceptron in a book entitled "Perceptrons: an introduction to computational geometry" (Minsky and Papert 2017). One of the most discussed issues that the authors presented in the book is the difficulty that a neural network would have to compute a simple XOR (exclusive OR) operation. They claimed that using Rosenblatt's algorithm, the operation could not be solved, as it would require multiple layers of perceptrons.

In 1974, Paul Werbos studied on his PhD the application of the backpropagation algorithm in neural networks (Werbos 1974), making the training of multi-layer neural networks possible. This technique did not gain popularity until 1986, when David Rumelhart, Geoffrey Hinton, and Ronald Williams published their outstanding work describing the methodology and addressing the problems discussed by Minsky's work(Rumelhart et al. 1986; Rumelhart et al. 1985).

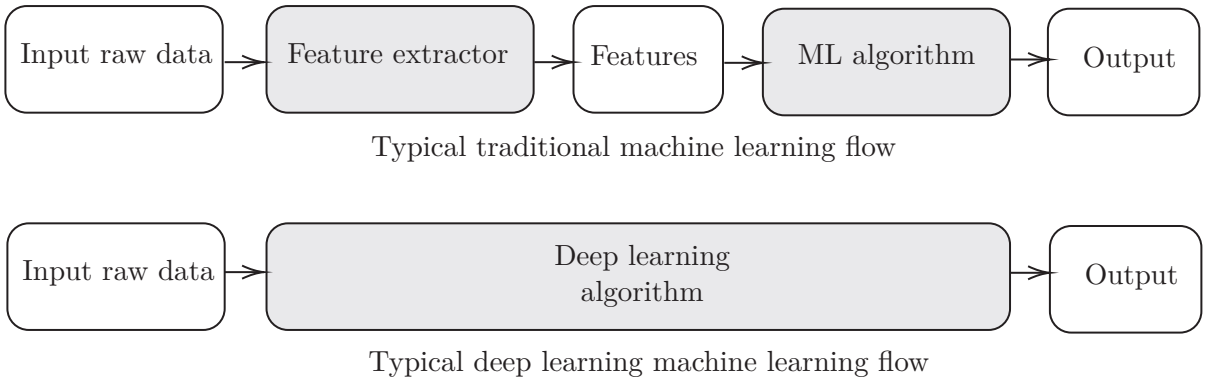


Figure 2.2: Typical workflows for machine learning and deep learning problems. Inspired by the presentation of Vivek Kumar entitled Artificial Intelligence in Audio Event (https://www.youtube.com/watch?v=388AZ2ujM9w&t=208s, accessed on 29 December 2021).

Once the key of how to train multi-layer neural networks was made publicly available, the first deep learning applications started to appear. One of the most relevant contributions of that time occurred in 1989, when Yann LeCun et. al. applied a Convolutional Neural Network o Xarxa Neuronal Convolucional (CNN) for the classification of handwritten digits (LeCun et al. 1989). At that time, even though they were not so popular, other applications started using DL techniques. For example, in 1988, researchers such as Lewis (Lewis 1988) and Todd (Todd 1988) proposed the use of neural networks for automatic composition of music. The main drawback of DL at that time was the amount of time that it took to train the models as the technology that was available had major limitations (in comparison of the Graphics Processing Unit o Unitat de Processament de Gràfics (GPU)'s that are available nowadays. As technology has greatly evolved in recent years, nowadays, and despite of the drawbacks that it may have (requires big amounts of data and computation capabilities), DL has become a popular research trend due to the exceptional accuracy results it achieves.

2.5.2 Audio classification baseline

Usually, the most common workflows followed when assessing a ML or DL problem in any field (not necessarily in the audio domain) are the ones shown on Figura 2.2. As it can be observed, whereas ML requires a big effort in the process of feature engineering (i.e., selecting the most convenient features for each particular problem), DL aims at skipping this part and using the raw input data as the input of the model, expecting it to automatically extract the most convenient features for the learning of the neural network. However, this is not necessarily the case (or, at least, yet) in the audio domain. In the current state of the art, a big amount of audio detection and classification problems are solved by means of using a neural network with spectrograms as inputs. This way, in the audio domain, in some problems, DL may still require a feature selection and tuning process for training neural networks. The detail on how this process is carried out and some relevant contributions in the field are explained in Secció 2.5.3 and Secció 2.5.4.

2.5.3 Machine learning in acoustic classification

An appropriate feature selection process becomes crucial when using traditional ML algorithms. This features can either contain information about the frequency content of the audio files, their temporal evolution or a mix between them.

The baseline or most commonly used features in this type of problems are the Mel Frequency Cepstral Coefficients o Coeficients Cepstrals de Freqüència Mel (MFCC), which aim at characterizing the audio files emulating human auditory perception. Usually, this features are combined with others such as Linear Prediction Cepstrum Coefficients o Coeficients Cepstrals de Predicció Lineal (LPCC) or Zero Crossing Rate o Taxa de Creuament per Zero (ZCR) among others (Dave 2013; Ito and Donaldson 1971).

For example, in (Giannakopoulos et al. 2015), authors present a system that aims at estimating the soundscape quality (in both natural and urban environments) by means of audio analysis. Concretely, to carry out the estimation of the soundscape quality, the system uses 68 feature statistics per each 15-second fragment of audio. These features include the aforementioned MFCC and ZCR features combined with the short term energy of the fragment, the entropy of the energy, the spectral centroid and spread, the spectral entropy, flux and roll-off and the chroma features. Once the features are calculated, four different SVM regressors are trained to estimate three context levels and the soundscape quality level.

Similarly, in (Noviyanti et al. 2019), authors aim at predicting the urban soundscape from a set of acoustic features. Specifically, the prediction is carried out using MFCC coefficients together with "acoustic ecology" parameters. With all these parameters, the work predicts the perception of relaxation, dynamic and communication in a given soundscape. Authors conclude from the results that MFCC are better features than the acoustic ecology metrics for their model's performance, that is based on a binary logistic regressor.

Other examples can be found in (Tsalera et al. 2020) or in (Lojka et al. 2014). In the first case, authors also combine features for classification. They use 8 temporal features (including ZCR), 11 spectral features and 4 perceptual features (including MFCC). In this case, the selected classification algorithm is a simple K-Nearest Neighbors o K-Veïns Propers (KNN), and the results are discussed when using different numbers of neighbors (1 to 3) and different distance metrics (Euclidean, Chebyshev and Cosine). The target of this work is to classify events occurring in urban environments from 9 different categories. In the second case, the used features are MFCC together with Frequency Bank Coefficients o Coeficients de Bancs de Freqüències (FBC) and Mel-Spectral Coefficients o Coefficients Espectrals Mel (MSC). For classification, they use a modified Viterbi decoding process together with Weighted Finite-State Transducers (WFSTs) and a HMM.

Clustering techniques are also widely used in the audio domain. For instance, in (Pita et al. 2021), authors use an unsupervised (i.e., it does not require labelled data to train the model) ML algorithm to cluster data from the city of Barcelona to detect clusters that are close to urban roads, residential areas and leisure areas. Despite this work does not perform acoustic event classification, it shows an interesting methodology to segment the city according to the noise levels presents in each area. In this case, only the equivalent levels of noise are used as

features. The ML clustering algorithm used in the work is K-means.

Another work that uses the noise level (processed in several features) as an input to a model to evaluate a certain soundscape is the one presented in (Torija et al. 2014). In that work, the algorithms used to assess the soundscape evaluation are a SVM and a Sequential Minimal Optimization o Optimització Mínima Seqüencial (SMO). Results show that SMO outperforms SVM when performing the task of soundscape classification.

As it can be seen, studying a given soundscape by means of ML techniques can be seen from different points of view: since clustering different areas of a city depending on their noise levels or correlate different noise events with the perception of the population, to automatically detecting acoustic events happening on a urban environment to know which are the most polluted noise areas. One big project that considered several points of view of two different areas (urban area and suburban area) was the LIFE+ DYNAMAP project. Concretely, in that project, researchers developed a low-cost WASN to monitor two large-scale areas in the cities of Milan and Rome using a dynamic acoustic map updated at real-time. During the project, once the WASN was deployed, long recording campaign took place and allowed a massive data gathering (Alsina-Pagès et al. 2019). Some of those data were later manually labelled by experts, which allowed (1) to study the impact of the different acoustic events to the population (Alías et al. 2020) and (2) the development of an algorithm that detects and differentiates anomalous acoustic events from road traffic noise (Socoró et al. 2017; Alías et al. 2018). For classification, authors use MFCC as features and GMM as classification model.

Another technique that has been used in the field is using a bag-of-features approach (Grzeszick et al. 2017). For example, in (Grzeszick et al. 2017), authors use MFCC and GammaTone Cepstral Coefficients o Coeficients Cepstrals de Tons Gamma (GTCC) as an underlying set of features that are later quantized with respect to a certain codebook to generate a bag-of-features. The works also shows how can robustness of models be improved by fusing acoustic data from multiple channels. One advantage of this approach is that it does not require high computational capabilities, which means that the algorithm can run over low-cost (in terms of computation) hardware or on on-line environments.

In some other works, ML techniques are mixed with signal processing techniques and the classification of acoustic events is split in different classification layers. For example, in (Luitel et al. 2016), authors conceive a two-layers classifier to categorize urban events coming mainly from vehicles (i.e., bus engine, bus horn, car horn and whistle). A first layer divides the acoustic events in two classes with signal processing techniques (looking at the frequency spectrum and filtering to the desired frequencies) and a second layer finally classifies the acoustic event. Again, the classifier uses the baseline MFCC features as inputs for the models. Several classifiers are tested, for instance an Artificial Neural Network o Xarxa Neuronal Artificial (ANN), a Naive Bayes (NB) classifier, a Decision Tree o Arbre de Decisió (DT) and a Random Forest o Bosc Aleatòri (RF). Their results show that using two layers improves the classification results with respect to a one-layer baseline system.

However, not all the classification problems rely on the MFCC features for classification. In (Salamon and Juan Pablo Bello 2015a) and (Salamon and Juan Pablo Bello 2015b), authors

prove that it is possible to achieve better classification results (in terms of accuracy only) when using unsupervised feature learning from 2D mel-spectrograms and a scattering technique. The application of a scattering transform allows to characterize the short-term temporal dynamics captured by 2D mel spectrogram patches with the added advantage of being phase invariant. This is an advantage given that they are able to characterize signals that vary in time over (relative) long windows in comparison to other methods such as MFCC computation.

Similarly, in (Waldekar and Saha 2020), authors propose the usage of wavelet transform given that it performs good when extracting characteristics information from environmental audio as it can vary in length. As environmental noise may have overlapping frequency content and also a wider frequency range (compared to other fields such as speech recognition), the mel-scaled wavelet transform presented in the work outperforms a baseline MFCC-based classification system.

To sum up, different ML techniques have been widely used in the state of the art of acoustic event detection and acoustic soundscape evaluation. Also, when using those algorithms, MFCC are the baseline features, but they are usually combined with other acoustic parameters to achieve better classification results.

2.5.4 Deep learning in acoustic event classification

In the later years, and due to the rapid development of technology, DL has gained popularity in the field of acoustic event classification. Mostly, big DL networks have been trained to classify acoustic events in machines with high-computation capabilities. However, this is not the case of all the DL classification problems. Some small networks have been conceived to classify acoustic events even in low-cost (in terms of computational capabilities) devices such as mobile phones (Stowell 2021) efficiently.

Rather than classifying acoustic events, some works focus on classifying acoustic scenes (Acoustic Scene Classification o Classificació d'Escena Acústica (ASC)). The difference between acoustic event detection and acoustic scene classification is that, whereas the first one focuses on assigning a semantic label to a concrete acoustic event coming from a concrete noise source, the second one assigns a label referring to the environment in which an acoustic stream was recorded. In this field, it is very common to use also Deep Neural Network o Xarxa Neuronal Profunda (DNN). However, not only CNN are popular in the field. For example, in (Singh et al. 2021), authors use a prototypical neural network to obtain an embedding space for the task of ASC. The hypothesis behind prototypical networks is that there exists an embedding space in which points cluster around a single prototype representation for each class (Snell et al. 2017). However, even though the hypothesis is promising, this type of networks typically obtain moderate accuracy values and must be further studied.

Some remarkable state of the art projects involving acoustic event classification for different applications with DL are explained below.

In (Genaro et al. 2010), authors use 25 features to train an ANN to predict the noise level in a urban environment. In their study, authors compare the results predicted by the ANN and the results obtained when applying Principal Component Analysis o Anàlisi de

Components Principals (PCA) from the model aiming for model simplification. Even though the results are worse after applying PCA, authors claim that they are still acceptable.

One noticeable work is the one published in (Lopez-Ballester et al. 2019; Lopez-Ballester et al. 2020). Their application focuses on evaluating the annoyance of audio sounds using DL. Concretely, they use a CNN that is capable of predicting psycho-acoustic annoyance using as inputs raw audio signals. The conclusions of their result is that their network is able to predict faster than conventional calculations for psycho-acoustic annoyance maintaining high-precision, making the deployment of their network suitable for IoT devices.

CNNs have been also widely used to classify acoustic events to monitor the biodiversity status. For instance, in (Morgan and Braasch 2021), data from the state of New York (United States) have been collected, analysed and classified to calculate the species richness and distribution from some pseudo-species. The inputs of the neural network are the spectrograms of the audio signals. Also, the work studies the correlation between the acoustic events and other abiotic parameters such as the temperature or weather conditions. Another example can be found in (Nanni et al. 2021). In this case, authors use an ensemble of CNN to classify acoustic events from different datasets (bird vocalizations, cat sounds or environmental sounds). Authors claim that their off-the-self ensemble can be trained on different datasets and reach state of the art performances (in terms of accuracy).

In (Mushtaq and Su 2020), authors also use CNN for acoustic classification, in this case for environmental sound classification. As features, authors use the mel-spectrogram together with the MFCC parameters. Actually, in this work, authors highlight the importance of using data augmentation techniques to boost the classification results. Data augmentation, in this context, refers to using techniques to increase the size of the training data set by adding more samples generated by modifying the data from the dataset or creating new data from existing samples.

The importance of the input selection to the CNN has been evaluated by some works of the field. For instance, in (Zhou et al. 2017). Their experiments conclude that the best classification performance on urban sounds is normally achieved when the input spectrograms have moderate time resolution. Also, the normalization of the input data when using spectrograms is a topic that is still open for discussion. For example, in (Ick and McFee 2021), authors explore the different parameters of Per-Channel Energy Normalization o Normalització d'Energia Per Canal (PCEN) (which is an adaptative procedure that has been proved to be useful in some audio classification problems (Lostanlen et al. 2018)) and propose a multi-rate PCEN approach to improve classification results.

Due to the large amount of data required to train DL models, data augmentation has been used in most of the DL works in the field. Apart from the work by (Mushtaq and Su 2020), this technique has been widely used by the community in the later years (Davis and Suresh 2018; Shah et al. 2019; Shen et al. 2020; Nanni et al. 2021; Dinkel et al. 2021).

More works applying audio classification using CNNs can be found in (Sang et al. 2018; AbeBer et al. 2018; Bai et al. 2019; Phan et al. 2019; Fairbrass et al. 2019; Cao et al. 2019; Shen et al. 2020; Ciaburro 2020; Ciaburro and Iannace 2020). Whereas some of the works use

the spectrograms (or a fusion between the spectrograms an other acoustic features) as inputs for their networks, the other ones use raw audio data.

Also, whereas most DL models are too large to be deployed in low-cost devices, some effort has been made in the field to study deployment strategies. This is the case, for example, of the project presented in (Arce et al. 2021). The work presents a WASN that monitors urban areas and recognizes a group of acoustic events. The nodes of the network are composed of a Rasperry Pi as a computing unit, and the classification algorithm is a CNN together with a pre-detection stage that is able to differ three relevant events from traffic and activates the CNN only when one of the three relevant events occurs. This way, using the pre-detection stage, authors are able to reduce the Central Processing Unit o Unitat de Processament Central (CPU) usage of their computing device by a factor of 6.

Finally, it must be taken into account that, in urban environments, it is common to find acoustic events occurring simultaneously (also known as polyphonic events). This is a challenging task as some events have more acoustic level than others and they have different duration and structure. Several works from the field aim at recognizing multiple events at the same time. Usually, this task is done by applying a manual threshold on the last layer of the DNN to decide if the event is present in the acoustic fragment of the input or not. However, there are works that elaborate different strategies to achieve multi-label classification. For example, in (Xia et al. 2018), authors use a multi-variable regression approach and give a confidence to each audio segments. Their results show that, this way, the classification scores are higher.

On a different work (Pankajakshan et al. 2019), authors propose a model that aims at improving the temporal localization of sound events using a combination of two models. The first model predicts which sound events are present at each time frame, and the second one predicts if a sound event is present or not in an acoustic frame. The joint models result in higher classification scores than a separate implementation of each of the models.

Another work that handles multi-label data is the one presented in (He et al. 2020). To assess the multi-label classification process, authors use a a sigmoid-sparsemax multi-activation structure.

Other studies that deal with polyphonic or multi-label data can be found at (Xia et al. 2020; Gontier et al. 2021; Luo et al. 2021). In the first work (Xia et al. 2020), authors use both the event position within a whole audio segment (task 1) and the frame position inside an audio event (task 2) for the development of a multi-task learning approach. Results on a monophonic dataset and a polyphonic dataset confirm that their approach achieves improved classification results compared to the baseline of that respective datasets. In the second work (Gontier et al. 2021), authors use polyphonic training set synthesis to improve classification results compared to a self-supervised learning method. Polyphonic training set synthesis consists in annotating a small corpus of acoustic events of interest, which are then automatically mixed at random to form a larger corpus of polyphonic scenes. In the work, authors claim that the geographical origin of the acoustic events in the training set synthesis has a great impact on the classification results. Finally, in (Luo et al. 2021), authors use a

combination model of a capsule neural network (CapsNet) and a recurrent neural network. As inputs to their model, authors use a feature aggregation method including MFCC and log-mel features. Also, they have implemented a real-world system capable of detecting the acoustic events in urban environments.

2.5.5 Wireless Acoustic Sensor Networks deployed in the modern society

This subsection gives an overview of some WASN that are currently deployed in different environments: either urban environments or suburban areas. Apart from the already mentioned DYNAMAP project, (Alsina-Pagès et al. 2019), that deployed acoustic sensors in different areas of Rome and Milan, other works around the world have carried out similar approaches.

For example, in the United Kingdom, the DREAMsys (Distributed Remote Environmental Array & Monitoring System) (Barham et al. 2010) project has developed a measurement-based approach to survey environmental noise and perform acoustic mapping using novel distributed sensors. The hardware of their sensors includes a MEMS microphone protected with a waterproof and windproof shield, a computing unit that calculates the equivalent level of noise, a GSM modem and batteries that can last up to 15 days and a tripod that allows the mobility of sensors.

In Italy, in a short-term project named SENSEable (Nencini et al. 2012) deployed a WASN in the city of Pisa to measure the noise levels in real-time in different locations of the city.

In the use-case scenario, Barcelona, several high-quality (class I) sensing nodes have been deployed as well (Farrés 2015). The aim of the WASN deployed in Barcelona is to (1) evaluate the noise levels in noisy areas, (2) quantify the noise reduction when action plans are implemented, (3) update a real-time noise map and (4) identify noise sources and evaluate them. However, this evaluation is still made manually, no automatic classification systems have been deployed. The real-time sensing values can be seen in the SENTILO platform, which supplies information regarding the acoustic situation of the city but also shows values gathered by different types of sensors (for example, it shows meteorological information). The real-time map can be seen at https://connecta.bcn.cat/connecta-catalog-web/component/map (accessed on 30 December 2021).

Another WASN can be found in Canada, in the framework of the project UrbanSense (Rainham 2016). In this case, the network is not only responsible of monitoring acoustic data (LAeq), it also takes into account other parameters of interest such as the amount of carbon dioxide, carbon monoxide, wind speed and direction, temperature, relative humidity and precipitation.

In Paris (France), the BUITPARIF organization has carried out a project which has resulted in the design and patent of a noise-monitoring device called MEDUSA (C. Mietlicki and F. Mietlicki 2018), that combines four microphones and one optical system so it is possible to represent noise levels in 360º. Then, the noise levels are projected over a geographic map creating coloured hexagons that allow to see the noise levels of the area.

The LIFE Monza project (Bartalucci et al. 2018), a LIFE project that lasted until 2020, developed a method for the identification and the management of the Noise Low Emission

Zone, which is an urban zone subject to traffic restrictions to mitigate the impact of noise to the population. Concretely, the project deployed a pilot test in the city of Monza (northern Italy). The physical implementation of the project involved the usage of low-cost sensors and a web-page interface.

Outside Europe, noise monitoring is also considered an important issue to be taken into account. For instance, the SONYC (Sound Of New York City) project (Juan P Bello et al. 2019), has deployed 55 low-cost acoustic sensors in the city of New York. Each sensor is composed of a sensor core (Raspberri Pi + WiFi antenna) and an acoustic sensing module (based on a MEMS microphone and a microcontroller) (Mydlarz et al. 2019). Besides monitoring the noise level, those sensors are capable of classifying some event occurring in real-time.

Finally, in Spain, besides the already mentioned network deployed in Barcelona, several projects in different cities have addressed the challenge of noise monitoring using acoustic sensors. For example, in Málaga (López et al. 2020), to assess the quality of life of the citizens, and as in some areas the amount of leisure noise may exceed the limits permitted by the current regulations, a sub-set of 8 acoustic sensors were deployed as part of a sub-network. Those sensors aim to obtain several (86) acoustic parameters in real-time to monitor the noise level in problematic areas. In the capital of the country, Madrid, there have been installed 31 premium sound level meters (Brüel & Kjaer, class 1) with a microphone that allows outdoor deployments (Asensio et al. 2020). To guarantee reliable results, the sensors are calibrated each year according to the regulations. Those sensors are installed on the roof of environmental conditions measurement booths, and are in charge of monitoring the noise levels of the selected locations.

2.5.6 Research gaps and community

To conclude this chapter, this section gives an overview of the current challenges in the field of acoustic event classification and what efforts are being made by the community to fill them.

As seen in this state of the art, worldwide researchers have been putting efforts to achieve year by year better results by means of using new ML algorithms, feature extraction algorithms, or a combination of both. However, most of those works are focused on achieving the best classification accuracy, without worrying about footprint of their models or the required hardware to perform classification. Actually, in different works, it is very common to see researchers worried only about the training time of their models, and they tend to sub-estimate the inference time. This is due to the big computational capabilities required for ML or, specially, DL training, which may require even months to obtain a stable state (this depends on the hardware being used for training, the size of the models and the amount of data available). However, for real world applications, inference time is more important than training time if the final objective is to supply a classification result in real-time. For this reason, this dissertation aims to contribute to this research gap rather than building a model that would obtain the best classification accuracy possible. In this sense, this dissertation will balance the inference computing time and memory on the proposed low-cost hardware and the accuracy

scores.

To push the state of the art forward, a research competition is organized every year proposing several challenges (usually 5 or 6 tasks) to be addressed in different domains of acoustic event classification. The competition is called Detection and Classification of Acoustic Scenes and Events o Detecció i Classificació d'Escenes i Esdeveniments Acústics (DCASE) (http://dcase.community, accessed on 29 December 2021) and aims at promoting several datasets and classification systems while encouraging researchers to apply new classification techniques to win the competition. The different tasks may include acoustic data from different domains such as urban events, animal vocalizations or acoustic scene classification.

References

AbeBer, Jakob, Gotze, Marco, Kuhnlenz, Stephanie, Grafe, Robert, Kuhn, Christian, ClauB, Tobias and Lukashevich, Hanna (Aug. 2018). 'A Distributed Sensor Network for Monitoring Noise Level and Noise Sources in Urban Environments'. In: *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*. Barcelona, Spain: IEEE, pp. 318–324.

Alías, Francesc, Alsina-Pagès, Rosa Ma, Orga, Ferran and Socoró, Joan Claudi (July 2018). 'Detection of Anomalous Noise Events for Real-Time Road-Traffic Noise Mapping: The Dynamap's project case study'. In: *Noise Mapping* vol. 5, no. 1, pp. 71–85.

Alías, Francesc, Orga, Ferran, Alsina-Pagès, Rosa Ma and Socoró, Joan Claudi (Jan. 2020). 'Aggregate Impact of Anomalous Noise Events on the WASN-Based Computation of Road Traffic Noise Levels in Urban and Suburban Environments'. en. In: *Sensors* vol. 20, no. 3, p. 609.

Alsina-Pagès, Rosa Ma, Orga, Ferran, Alías, Francesc and Socoró, Joan Claudi (May 2019). 'A WASN-Based Suburban Dataset for Anomalous Noise Event Detection on Dynamic Road-Traffic Noise Mapping'. en. In: *Sensors* vol. 19, no. 11, p. 2480.

Amo Filvà, Daniel, Alier Forment, Marc, García Peñalvo, Francisco Javier, Fonseca Escudero, David and Casany Guerrero, María José (2020). 'Privacidad, seguridad y legalidad en soluciones educativas basadas en Blockchain: Una Revisión Sistemática de la Literatura'. In: *RIED. Revista iberoamericana de educación a distancia* vol. 23, no. 2, pp. 213–236.

Arce, Pau, Salvo, David, Piñero, Gema and Gonzalez, Alberto (Sept. 2021). 'FIWARE based low-cost wireless acoustic sensor network for monitoring and classification of urban soundscape'. en. In: *Computer Networks* vol. 196, p. 108199.

Asensio, César, Pavón, Ignacio and De Arcas, Guillermo (2020). 'Changes in noise levels in the city of Madrid during COVID-19 lockdown in 2020'. In: *The Journal of the Acoustical Society of America* vol. 148, no. 3, pp. 1748–1755.

Bai, Jisheng, Chen, Chen and Chen, Jianfeng (Nov. 2019). 'A Multi-feature Fusion Based Method For Urban Sound Tagging'. In: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Lanzhou, China: IEEE, pp. 1313–1317.

Barham, Richard, Chan, Martin and Cand, Matthew (2010). 'Practical experience in noise mapping with a MEMS microphone based distributed noise measurement system'. In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Vol. 2010. 6. Institute of Noise Control Engineering, pp. 4725–4733.

Bartalucci, Chiara, Borchi, Francesco, Carfagni, Monica, Furferi, Rocco, Governi, Lapo, Lapini, Alessandro, Bellomini, Raffaella, Luzzi, Sergio and Nencini, Luca (2018). 'The smart noise monitoring system implemented in the frame of the Life MONZA project'. In: *Proceedings of the EuroNoise*, pp. 783–788.

Bello, Juan P, Silva, Claudio, Nov, Oded, Dubois, R Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles and Doraiswamy, Harish (2019). 'Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution'. In: *Communications of the ACM* vol. 62, no. 2, pp. 68–77.

Cao, Jiuwen, Cao, Min, Wang, Jianzhong, Yin, Chun, Wang, Danping and Vidal, Pierre-Paul (Oct. 2019). 'Urban noise recognition with convolutional neural network'. en. In: *Multimedia Tools and Applications* vol. 78, no. 20, pp. 29021–29041.

Ciaburro, Giuseppe (Aug. 2020). 'Sound Event Detection in Underground Parking Garage Using Convolutional Neural Network'. en. In: *Big Data and Cognitive Computing* vol. 4, no. 3, p. 20.

Ciaburro, Giuseppe and Iannace, Gino (July 2020). 'Improving Smart Cities Safety Using Sound Events Detection Based on Deep Neural Network Algorithms'. en. In: *Informatics* vol. 7, no. 3, p. 23.

Dave, Namrata (2013). 'Feature extraction methods LPC, PLP and MFCC in speech recognition'. In: *International journal for advance research in engineering and technology* vol. 1, no. 6, pp. 1–4.

Davis, Nithya and Suresh, K (Dec. 2018). 'Environmental Sound Classification Using Deep Convolutional Neural Networks and Data Augmentation'. In: *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. Thiruvananthapuram, India: IEEE, pp. 41–45.

Dinkel, Heinrich, Wu, Mengyue and Yu, Kai (2021). 'Towards Duration Robust Weakly Supervised Sound Event Detection'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 29, pp. 887–900.

Fairbrass, Alison J., Firman, Michael, Williams, Carol, Brostow, Gabriel J., Titheridge, Helena and Jones, Kate E. (Feb. 2019). 'CityNet—Deep learning tools for urban ecoacoustic assessment'. en. In: *Methods in Ecology and Evolution* vol. 10, no. 2. Ed. by Isaac, Nick, pp. 186–197.

Farrés, Júlia Camps (2015). 'Barcelona noise monitoring network'. In: *Proceedings of the Euronoise*, pp. 218–220.

Fitch, Frederic B (1944). 'McCulloch Warren S. and Pitts Walter. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics, vol. 5, pp. 115–133'. In: *Journal of Symbolic Logic* vol. 9, no. 2.

Genaro, N., Torija, A., Ramos-Ridao, A., Requena, I., Ruiz, D. P. and Zamorano, M. (Oct. 2010). 'A neural network based model for urban noise prediction'. en. In: *The Journal of the Acoustical Society of America* vol. 128, no. 4, pp. 1738–1746.

Giannakopoulos, Theodoros, Siantikos, Georgios, Perantonis, Stavros, Votsi, Nefta-Eleftheria and Pantis, John (July 2015). 'Automatic soundscape quality estimation using audio analysis'. en. In: *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. Corfu Greece: ACM, pp. 1–9.

Gontier, Félix, Lostanlen, Vincent, Lagrange, Mathieu, Fortin, Nicolas, Lavandier, Catherine and Petiot, Jean-Francois (2021). 'Polyphonic training set synthesis improves self-supervised urban sound classification'. In: *The Journal of the Acoustical Society of America* vol. 149, no. 6, pp. 4309–4326.

Grzeszick, Rene, Plinge, Axel and Fink, Gernot A. (June 2017). 'Bag-of-Features Methods for Acoustic Event Detection and Classification'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 25, no. 6, pp. 1242–1252.

He, Kexin, Shen, Yuhan, Zhang, Wei-Qiang and Liu, Jia (May 2020). 'Staged Training Strategy and Multi-Activation for Audio Tagging with Noisy and Sparse Multi-Label Data'. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, pp. 631–635.

Ick, Christopher and McFee, Brian (June 2021). 'Sound Event Detection in Urban Audio with Single and Multi-Rate Pcen'. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, pp. 880–884.

Ito, M and Donaldson, R (1971). 'Zero-crossing measurements for analysis and recognition of speech sounds'. In: *IEEE Transactions on Audio and Electroacoustics* vol. 19, no. 3, pp. 235–242.

Khan, Khalid S, Kunz, Regina, Kleijnen, Jos and Antes, Gerd (2003). 'Five steps to conducting a systematic review'. In: *Journal of the royal society of medicine* vol. 96, no. 3, pp. 118–121.

LeCun, Yann, Boser, Bernhard, Denker, John S, Henderson, Donnie, Howard, Richard E, Hubbard, Wayne and Jackel, Lawrence D (1989). 'Backpropagation applied to handwritten zip code recognition'. In: *Neural computation* vol. 1, no. 4, pp. 541–551.

Lewis, John Peter (1988). 'Creation by refinement: a creativity paradigm for gradient descent learning networks.' In: *ICNN*, pp. 229–233.

Lojka, Martin, Pleva, Matúš, Kiktová, Eva, Juhár, Jozef and Čižmár, Anton (2014). 'EARTUKE: The Acoustic Event Detection System'. In: *Multimedia Communications, Services and Security*. Ed. by Junqueira Barbosa, Simone Diniz et al. Vol. 429. Cham: Springer International Publishing, pp. 137–148.

López, Juan Manuel, Alonso, Jesús, Asensio, César, Pavón, Ignacio, Gascó, Luis and Arcas, Guillermo de (2020). 'A Digital Signal Processor Based Acoustic Sensor for Outdoor Noise Monitoring in Smart Cities'. In: *Sensors* vol. 20, no. 3, p. 605.

Lopez-Ballester, Jesus, Pastor-Aparicio, Adolfo, Felici-Castell, Santiago, Segura-Garcia, Jaume and Cobos, Maximo (Oct. 2020). 'Enabling Real-Time Computation of Psycho-Acoustic

Parameters in Acoustic Sensors Using Convolutional Neural Networks'. In: *IEEE Sensors Journal* vol. 20, no. 19, pp. 11429–11438.

Lopez-Ballester, Jesus, Pastor-Aparicio, Adolfo, Segura-Garcia, Jaume, Felici-Castell, Santiago and Cobos, Maximo (Aug. 2019). 'Computation of Psycho-Acoustic Annoyance Using Deep Neural Networks'. en. In: *Applied Sciences* vol. 9, no. 15, p. 3136.

Lostanlen, Vincent, Salamon, Justin, Cartwright, Mark, McFee, Brian, Farnsworth, Andrew, Kelling, Steve and Bello, Juan Pablo (2018). 'Per-channel energy normalization: Why and how'. In: *IEEE Signal Processing Letters* vol. 26, no. 1, pp. 39–43.

Luitel, Bibek, Murthy, Y. V. Srinivasa and Koolagudi, Shashidhar G. (Aug. 2016). 'Sound event detection in urban soundscape using two-level classification'. In: *2016 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. Mangalore, India: IEEE, pp. 259–263.

Luo, Liyan, Zhang, Liujun, Wang, Mei, Liu, Zhenghong, Liu, Xin, He, Ruibin and Jin, Ye (2021). 'A System for the Detection of Polyphonic Sound on a University Campus Based on CapsNet-RNN'. In: *IEEE Access* vol. 9, pp. 147900–147913.

Mesaros, Annamaria, Heittola, Toni, Virtanen, Tuomas and Plumbley, Mark D (2021). 'Sound event detection: A tutorial'. In: *IEEE Signal Processing Magazine* vol. 38, no. 5, pp. 67–83.

Mietlicki, Christophe and Mietlicki, Fanny (2018). 'Medusa: a new approach for noise management and control in urban environment'. In: *Proceedings of the 11th European Congress and Exposition on Noise Control Engineering (Euronoise2018), Crete, Greece*, pp. 27–31.

Minsky, Marvin and Papert, Seymour A (2017). *Perceptrons: An introduction to computational geometry*. MIT press.

Morgan, M.M. and Braasch, J. (Mar. 2021). 'Long-term deep learning-facilitated environmental acoustic monitoring in the Capital Region of New York State'. en. In: *Ecological Informatics* vol. 61, p. 101242.

Mushtaq, Zohaib and Su, Shun-Feng (Oct. 2020). 'Environmental sound classification using a regularized deep convolutional neural network with data augmentation'. en. In: *Applied Acoustics* vol. 167, p. 107389.

Mydlarz, Charlie, Sharma, Mohit, Lockerman, Yitzchak, Steers, Ben, Silva, Claudio and Bello, Juan Pablo (2019). 'The life of a New York City noise sensor network'. In: *Sensors* vol. 19, no. 6, p. 1415.

Nanni, Loris, Maguolo, Gianluca, Brahnam, Sheryl and Paci, Michelangelo (June 2021). 'An Ensemble of Convolutional Neural Networks for Audio Classification'. en. In: *Applied Sciences* vol. 11, no. 13, p. 5796.

Nencini, Luca, De Rosa, Paolo, Ascari, Elena, Vinci, Bruna and Alexeeva, Natalia (2012). 'SENSEable Pisa: A wireless sensor network for real-time noise mapping'. In: *Proceedings of the EURONOISE, Prague, Czech Republic*, pp. 10–13.

Noviyanti, Anastasia, Sudarsono, Anugrah Sabdono and Kusumaningrum, Dian (2019). 'Urban soundscape prediction based on acoustic ecology and MFCC parameters'. In: Padang, Indonesia, p. 050005.

Pankajakshan, Arjun, Bear, Helen L. and Benetos, Emmanouil (Aug. 2019). 'Polyphonic Sound Event and Sound Activity Detection: A Multi-task approach'. In: *arXiv:1907.05122 [cs, eess]*. arXiv: 1907.05122.

Phan, Huy, Chén, Oliver Y., Koch, Philipp, Pham, Lam, McLoughlin, Ian, Mertins, Alfred and De Vos, Maarten (Feb. 2019). 'Unifying Isolated and Overlapping Audio Event Detection with Multi-Label Multi-Task Convolutional Recurrent Neural Networks'. In: *arXiv:1811.01092 [cs, eess, stat]*. arXiv: 1811.01092.

Pita, Antonio, Rodriguez, Francisco J. and Navarro, Juan M. (Aug. 2021). 'Cluster Analysis of Urban Acoustic Environments on Barcelona Sensor Network Data'. en. In: *International Journal of Environmental Research and Public Health* vol. 18, no. 16, p. 8271.

Rainham, D (2016). 'A wireless sensor network for urban environmental health monitoring: UrbanSense'. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 34. 1. IOP Publishing, p. 012028.

Rosenblatt, Frank (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rumelhart, David E, Hinton, Geoffrey E and Williams, Ronald J (1985). *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, David E, Hinton, Geoffrey E and Williams, Ronald J (1986). 'Learning representations by back-propagating errors'. In: *nature* vol. 323, no. 6088, pp. 533–536.

Salamon, Justin and Bello, Juan Pablo (Aug. 2015a). 'Feature learning with deep scattering for urban sound analysis'. In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. Nice: IEEE, pp. 724–728.

Salamon, Justin and Bello, Juan Pablo (Apr. 2015b). 'Unsupervised feature learning for urban sound classification'. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, pp. 171–175.

Sang, Jonghee, Park, Soomyung and Lee, Junwoo (Sept. 2018). 'Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms'. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. Rome: IEEE, pp. 2444–2448.

Shah, Sayed Khushal, Tariq, Zeenat and Lee, Yugyung (Dec. 2019). 'IoT based Urban Noise Monitoring in Deep Learning using Historical Reports'. In: *2019 IEEE International Conference on Big Data (Big Data)*. Los Angeles, CA, USA: IEEE, pp. 4179–4184.

Shen, Yexin, Cao, Jiuwen, Wang, Jianzhong and Yang, Zhixin (Jan. 2020). 'Urban acoustic classification based on deep feature transfer learning'. en. In: *Journal of the Franklin Institute* vol. 357, no. 1, pp. 667–686.

Singh, Shubhr, Bear, Helen L. and Benetos, Emmanouil (June 2021). 'Prototypical Networks for Domain Adaptation in Acoustic Scene Classification'. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, pp. 346–350.

Snell, Jake, Swersky, Kevin and Zemel, Richard S (2017). 'Prototypical networks for few-shot learning'. In: *arXiv preprint arXiv:1703.05175*.

Socoró, Joan, Alías, Francesc and Alsina-Pagès, Rosa (Oct. 2017). 'An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments'. en. In: *Sensors* vol. 17, no. 10, p. 2323.

Stowell, Dan (2021). *Computational bioacoustics with deep learning: a review and roadmap.* arXiv: 2112.06725 [cs.SD].

Todd, Peter (1988). 'A sequential network design for musical applications'. In: *Proceedings of the 1988 connectionist models summer school*, pp. 76–84.

Torija, Antonio J., Ruiz, Diego P. and Ramos-Ridao, Ángel F. (June 2014). 'A tool for urban soundscape evaluation applying Support Vector Machines for developing a soundscape classification model'. en. In: *Science of The Total Environment* vol. 482-483, pp. 440–451.

Tsalera, Eleni, Papadakis, Andreas and Samarakou, Maria (Nov. 2020). 'Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm'. en. In: *Energy Reports* vol. 6, pp. 223–230.

Waldekar, Shefali and Saha, Goutam (Mar. 2020). 'Analysis and classification of acoustic scenes with wavelet transform-based mel-scaled features'. en. In: *Multimedia Tools and Applications* vol. 79, no. 11-12, pp. 7911–7926.

Werbos, Paul (1974). 'Beyond regression:" new tools for prediction and analysis in the behavioral sciences'. In: *Ph. D. dissertation, Harvard University.*

Xia, Xianjun, Togneri, Roberto, Sohel, Ferdous and Huang, David (Apr. 2018). 'Confidence Based Acoustic Event Detection'. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Calgary, AB: IEEE, pp. 306–310.

Xia, Xianjun, Togneri, Roberto, Sohel, Ferdous, Zhao, Yuanjun and Huang, Defeng (Mar. 2020). 'Multi-Task Learning for Acoustic Event Detection Using Event and Frame Position Information'. In: *IEEE Transactions on Multimedia* vol. 22, no. 3, pp. 569–578.

Zhou, Huan, Song, Ying and Shu, Haiyan (Nov. 2017). 'Using deep convolutional neural network to classify urban sounds'. In: *TENCON 2017 - 2017 IEEE Region 10 Conference.* Penang: IEEE, pp. 3089–3092.

Capítol 3

Articles del compendi

Article I

Ester Vidaña-Vila, Leticia Duboc, Rosa Ma Alsina-Pagès, Francesc Polls, Harold Vargas. 'BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset'. Publicat a: *Sustainability*. Vol. 12, no. 19 (2020), pp. 8410. DOI: 10.3390/su12198140.

Article II

Ester Vidaña-Vila, Joan Navarro, Cristina Borda-Fortuny, Dan Stowell, Rosa Ma Alsina-Pagès. 'Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring'. Publicat a: *Electronics*. Vol. 9, no. 12 (2020), pp. 2119. DOI: 10.3390/electronics9122119.

Article III

Ester Vidaña-Vila, Joan Navarro, Dan Stowell, Rosa Ma Alsina-Pagès 'Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors'. Publicat a: *Sensors*. Vol. 21, no. 22 (2021), pp. 7470. DOI: https://doi.org/10.3390/s21227470.

Chapter 3

Papers of the compendium

Paper I

Ester Vidaña-Vila, Leticia Duboc, Rosa Ma Alsina-Pagès, Francesc Polls, Harold Vargas. 'BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset'. In: *Sustainability*. Vol. 12, no. 19 (2020), pp. 8410. DOI: 10.3390/su12198140.

Paper II

Ester Vidaña-Vila, Joan Navarro, Cristina Borda-Fortuny, Dan Stowell, Rosa Ma Alsina-Pagès. 'Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring'. In: *Electronics*. Vol. 9, no. 12 (2020), pp. 2119. DOI: 10.3390/electronics9122119.

Paper III

Ester Vidaña-Vila, Joan Navarro, Dan Stowell, Rosa Ma Alsina-Pagès 'Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors'. In: *Sensors*. Vol. 21, no. 22 (2021), pp. 7470. DOI: https://doi.org/10.3390/s21227470.

Paper I

BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset

Ester Vidaña-Vila, Leticia Duboc, Rosa Ma Alsina-Pagès, Francesc Polls, Harold Vargas

Published in *Sustainability*, October 2020, volume 12, issue 19, pp. 8410. DOI: 10.3390/su12198140.

Abstract

Acoustic pollution has been associated with adverse effects on the health and life expectancy of people, especially when noise exposure happens during the nighttime. With over half of the world population living in urban areas, acoustic pollution is an important concern for city administrators, especially those focused on transportation and leisure noise. Advances in sensor and network technologies made the deployment of Wireless Acoustic Sensor Networks (WASN) possible in cities, which, combined with artificial intelligence (AI), can enable smart services for their citizens. However, the creation of such services often requires structured environmental audio databases to train AI algorithms. This paper reports on an environmental audio dataset of 363 min and 53 s created in a lively area of the Barcelona city center, which targeted traffic and leisure events. This dataset, which is free and publicly available, can provide researchers with real-world acoustic data to help the development and testing of sound monitoring solutions for urban environments.

I.1 Introduction

More than four billion people (55% of the world population) live in urban areas, and the projection is that by 2050, this number will increase to seven billion, or two-thirds of the world population (Ritchie and Roser 2020). Barcelona, for example, has a population of over 1.6 million inhabitants (https://www.idescat.cat/ (Population: 2019)) and receives nine million tourists every year (Turisme a Barcelona - ciutat i regió 2019). Big cities like Barcelona combine a large range of industrial, business, and leisure activities, which can cause several environmental problems. Among these, acoustic pollution has gained increased attention over the last few years, as research has related the urban noise with adverse effects on the life expectancy and health of people (European Environment Agency, 2020 n.d.; Cik et al. 2016;

W 1991). In particular, the exposure to nocturnal noise was found to have a greater negative impact than daytime noise on long-term cardiovascular health, probably due to the repeated autonomic arousal during sleep (Jarup et al. 2008; Basner et al. 2011). Many of such studies focus on the negative health effects of traffic noise (Dratva et al. 2012; Hofman et al. 1995). However, leisure noise is being increasingly recognized as an important challenge in cities with a high number of tourist and cultural offers, where the needs of residents and visitors need to be balanced (Ottoz et al. 2018; Easteal et al. 2014), to the extent that the World Health Organization (WHO) has included recommendations on leisure noise in its recent Environmental Noise Guidelines for the European Region (World Health Organization, 2018 n.d.).

In order to address these problems, the European Commission has created the Environmental Noise Directive 2002/49/EC (END) (Cox and Palou 2002) and the Common Noise Assessment Methods in Europe (CNOSSOS-EU) (European Commission, Joint Research Centre—Institute for Health and Consumer Protection 2012). The former requires from Member States the development of separate strategic noise maps and noise management plans every five years for major roads, airports, railways, and agglomerations of more than 100,000 inhabitants. The latter provides common methods that Member States are expected to use for such purposes. Historically, such maps have been manually built to ensure this separation of noise sources, which is a laborious and slow process that requires human intervention.

With the advances in sensor and network technologies, many cities have now deployed Wireless Acoustic Sensor Networks (WASNs). These networks have the potential to represent a paradigm change for city managers and the population alike. WASNs can enable the automatic and real-time generation of strategic noise maps and, consequently, the creation of more efficient policies and technical solutions for managing urban noise pollution and for designing sustainable urban and suburban soundscapes.

Such technical solutions often use machine learning (ML) algorithms for the automatic identification of noise sources (Socoró et al. 2017; Alías and Alsina-Pagès 2019), many of which require supervised training. That is, they use structured environmental audio datasets to train these ML algorithms. However, the creation of reliable environmental audio datasets normally involves the manual tagging of many hours of audio data, which is very labor- and time-consuming. In order to avoid this work, several ML solutions are being developed using datasets created by artificially mixing sounds from online digital repositories, such as FreeSound (http://www.freesound.org), Soundcloud (http://www.soundcloud.com), and AudioHero (http://www.audiohero.com). While these datasets allow algorithms to be trained with very large amounts of data, most of them gather sounds collected from several places and devices, which could make training more difficult. A real-life dataset recorded under controlled conditions and devices is closer to real operation conditions of the nodes of the WASN and allows for data augmentation when the classification algorithm requires a larger dataset (Nakajima et al. 2016; Salamon and J. P. Bello 2017). Therefore, developers of sound solutions based on machine learning can benefit from free and publicly available real-world environmental audio datasets.

In this paper, we report on an environmental audio dataset created from 6 h of recording in a lively area of Barcelona city center, recorded as a joint collaboration with the Barcelona City Council. The recordings took place in a selected neighborhood that had produced a high number of complaints from residents. In particular, given the impact of nocturnal traffic and leisure noise on people's health, the spot chosen for the recording is well known for having both types of noise during the night.

Therefore, the contribution of this work is a precisely labeled night urban traffic and leisure sound dataset and its analysis, which is open and freely available to researchers and technicians. The analysis includes the duration of the events, the signal-to-noise ratio, the number of occurrences, the impact of each occurrence on the background noise L_{Aeq}, and the intermittency ratio (IR) of the entire data sample (Brambilla et al. 2019; Wunderli et al. 2016), which are metrics that have been related to healthy effects in different studies (Wunderli et al. 2016). We envision this dataset being used, extended, and combined with others for different purposes, such as the development of noise identification and monitoring solutions, the creation of guidelines for designing sustainable urban and suburban soundscapes, and the comparison with other datasets for health impact studies.

I.2 Related Work

The environmental acoustic databases described in the literature, which are used by the machine listening research community to train and test different types of algorithms, are normally generated by artificially mixing sounds or from real-life recordings. The former allows the control of the signal-to-noise ratio (SNR) of the synthetic mixtures (Alías and Socoró 2017), also dealing with data scarcity by means of data augmentation (Nakajima et al. 2016; Salamon and J. P. Bello 2017). Nevertheless, for this contribution, data augmentation techniques are not a key issue because the goal is to obtain and analyze a dataset exclusively from real operation data. We next review the literature on datasets recorded in real operation environments.

Valero (Valero and Alías 2012) presents an automatic approach for the classification of road vehicles by means of their pass-by signatures. The team recorded a dataset with six categories (light vehicles, heavy vehicles, motorcycles, aircrafts, trains, and industrial noise), resulting in 90 sound samples for each category, with a duration of 4 s each. Heittola (Heittola et al. 2013) published a 1,133 min audio dataset that includes 10 different acoustic environments from indoor and outdoor recordings. Their goal is to detail how contextual information can be used in automatic sound event detection. The work attempts to simulate human behavior when detecting and identifying sound events by means of a two-stage process that includes the automatic recognition of the context and the subsequent detection of the sound event. Despite the presence of the temporal component in this analysis, their work does not refer to the dynamics and the location of the sound. Instead, it uses the surrounding events to improve its classification results. Foggia (Pasquale Foggia et al. 2015) presents a large dataset of audio events for a surveillance application using acoustic event detection. The dataset

includes both long and short sounds and presents pieces of background noise with a significant noise level. The training dataset contains about 20 h, while the test set has around 9 h. In this sense, the goal of the dataset generation is completeness in terms of events coexisting with diverse background noise. Moreover, the same research laboratory developed a smaller dataset of about 1 h—also for surveillance purposes—focused on road acoustic events, which contains sound events from tire skidding and car crashes (P. Foggia et al. 2016).

Alías (Alías and Socoró 2017) presents a 9 h and 8 min real-life acoustic database collected from the urban and suburban environments of the pilot areas of the LIFE+ DYNAMAP project (Sevillano et al. 2016). This expert-based recording was carried out for discriminating Road-Traffic Noise (RTN) from Anomalous Noise Events (ANEs) through the Anomalous Noise Event Detection (ANED) algorithm running on low-cost acoustic sensors (Socoró et al. 2017; Alsina-Pagès et al. 2018). The ANEs, which correspond to 7.5% of the labeled data, were classified into 19 different subcategories after expert annotation, and the SNR levels were evaluated, taking as a reference the background noise. The SNR results ranged from −10 to +15 dB, also showing a wide heterogeneity of intermediate SNR levels. It is worth mentioning that the recordings in the urban area were conducted at the street level at pre-selected locations within District 9 of Milan (Zambon et al. 2017), while the recordings in the suburban area were conducted on the A90 ring-road portals surrounding Rome (see (Bellucci et al. 2017) for further details). In the final stage of the LIFE+ DYNAMAP project, in (Alsina-Pagès et al. 2019), the same authors presented the production and analysis of a real-operation environmental audio database collected through the 19-node WASN of a suburban area of Rome. As a result, 156 h and 20 min of labeled audio data were obtained, differentiating among RTN and ANEs (classified in 16 subcategories). The preliminary suburban expert-based dataset contained 3.2% of the ANEs of the total recorded time, whereas this new dataset contains only 1.8%. A possible explanation to these differences is that the expert-based dataset recording was centered in daytime, and this WASN-based dataset was recorded day and night, where night shows low presence of ANEs with respect to the day. A complementary analysis to these works can be found in (Alías et al. 2020), which is focused on evaluating the aggregate impact of the ANEs occurring in the acoustic environments where the sensors of WASNs are installed.

Another WASN-based project that has collected real operation acoustic samples is the SONYC (Sounds of New York) project (Juan P. Bello et al. 2019). Bello provides a simplified taxonomy of the sounds of the city by means of a two-level hierarchy, dividing them into eight coarse categories and 23 fine labels (Cartwright et al. 2019). The generated dataset is composed of 2351 recordings in the train split and 443 in the validation split, making a total of 2794 audio samples of 10 s each. The full taxonomy and details of the SONYC project dataset can be found in a previous work from the same authors (Salamon et al. 2014). The most innovative proposal of the SONYC project is that by means of the deployed network, the researchers can locate the distribution of the outdoor noise complaints and identify whether there have been, e.g., after-hours construction noise (Juan P. Bello et al. 2019). This identification can be done by means of the occurrence time of the group of annoying events,

also allowing the retrieval and visualization of the data streams obtained for each complaint location. Nevertheless, their use of deep learning models requires a large amount of labeled data, which are unavailable for environmental sound; for this reason, the data necessary for the training of the model are obtained by means of an audio data augmentation, which deforms the data using audio transformations (Juan P. Bello et al. 2019). The final dataset used to train and test the network contains both real-life audio pieces and other synthetically mixed samples.

Mesaros (Mesaros et al. 2019) details an acoustic dataset recorded in multiple cities in Europe, which is an extension of the TUT 2018 Urban Acoustic Scenes dataset (Mesaros et al. 2016). The original dataset contains recordings from Barcelona, Helsinki, London, Paris, Stockholm, and Vienna, and TAU 2019 adds Lisbon, Amsterdam, Lyon, Madrid, Milan, and Prague. The recordings were conducted with four devices simultaneously: (i) Soundman OKM II Klassik/studio A3 electret binaural microphone, (ii) Samsung Galaxy S7, (iii) IPhone SE, and (iv) GoPro Hero5 Session. Taking into account this variety of recording devices, the scenes were manually labeled to enable training and testing of machine learning algorithms. The dataset was used in one of the DCASE 2019 Challenges (http://dcase.community/challenge2019/), which included data from different recorded acoustic scenes and where the acoustic raw pieces were used together despite their different locations and origins.

I.3 Location Selection

Since our focus is the noise pollution caused by traffic and leisure activities, we chose to study large streets with large influxes of vehicles and with nighttime leisure activities. Our contacts in the Environmental Quality Department from the Barcelona City Council provided us with maps of the most problematic places in the district of Eixample based on the noise-related complaints from neighbors. The maps, which cannot be published for confidentiality reasons, contrast the areas with the greatest numbers of noise-related complaints about bars, restaurants, and music venues, many of which have terraces. After an analysis of the maps with the Department of Environmental Quality, we chose to focus our study in the following four parallel streets: Muntaner, Aribau, Enrique Granados, and Balmes, between the streets of Consell de Cent and Mallorca. All these streets have acoustic sensors or sound level meters located in light posts at around 4 m from the ground in places that were of interest for the Barcelona city council, as shown by the numbered circles in Figure I.1. We next summarize the analysis we conducted around these sensors:

1. Sensor 1 is a TA120 from CESVA with protection against external agents (such as birds, wind, rain, insects, etc.), which is a Class 1 precision sensor with programmable noise measurement integration time ranging from 1 s to 60 min, and is connected via optic fiber with the city council. It is located at Balmes street close to the corner with Consell de Cent street. Balmes is a street with a heavy traffic flow, but few leisure activities. The street has a sidewalk of two meters and four lanes dedicated to vehicles that circulate

downwards, from the mountain to the sea. This is an important street to access the city center. Consell de Cent street, on the other hand, connects the city from west to east. Even with three lanes intended for traffic, it not a busy street. Since this area does not have many leisure venues, such as bars or nightclubs, in the street, most noise will be generated by vehicles, which also tend to be fewer than in upward streets. See picture 1 in Figure I.2 for a photo of this street near the sensor.

2. Sound Level Meter 2 is TA025 from CESVA with an outdoor cabinet AR054 and an SC420 sound level meter, and is connected to the city council via 3G. It is located in Enric Granados, between the streets of Mallorca and Valencia. Enric Granados Street is one of the few pedestrian streets in the area, where the movement of vehicles is limited to only one lane and is at a reduced speed. It is also a street with many entertainment venues. These leisure activities are basically concentrated in bars and restaurants that have a closing time between 0:00 and 2:00 am. See picture 2 in Figure I.2 for a photo of this street near the sensor.

3. Sound Level Meter 3 is TA025 from CESVA with an outdoor cabinet AR054 and a SC420 sound level meter, and is connected to the city council via 3G. It is located in Aribau, between València and Mallorca. This street connects the city center with the northern part of Barcelona and has three lanes dedicated to vehicles. Furthermore, traffic circulates uphill (from the sea to the mountain), which increases the noise from vehicles, which have to use more engine power to get around. This street also has a very active night life, with many bars and restaurants that close between 2:00 and 3:00 am. As such, the noise in this street is caused both by heavy traffic and by leisure activities.

4. Sound Level Meter 4 is a TA024 from CESVA with an outdoor cabinet AR054 and a SC310 sound level meter, and is connected to the city council via 3G. It is located in Muntaner close to the corner of Consell de Cent. While the latter is not a busy street, Muntaner has a very high density of vehicles. However, as in Balmes, its traffic runs downhill, meaning that it generates less road traffic noise than in Aribau. On the other hand, around this corner, there is an important concentration of nightlife venues and, hence, noise generated from leisure activity. See picture 4 in Figure I.2 for a photo of this street near the sensor location.

Table I.1 summarizes the main characteristics of the streets considered in this study. Since our aim was to create a dataset for distinguishing between traffic and leisure noise, we chose to carry out our recording campaign in Aribau street.

I.4 Recording campaign

The recording campaign took place in two stages between March and June 2018. Since our goal was to create a dataset to collect raw acoustic data to distinguish between traffic and leisure noise and traffic is constant in the chosen location, we chose to carry out the study during the peak of the leisure activity hours; that is, on Saturdays between 22:00 and 03:00.

Table I.1: Summary of the main characteristics around the sensors in the area of interest.

Sensor	Direction	Traffic	Leisure Venues
1	Downhill	Heavy	Few
2	Downhill	Light	Many
3	Uphill	Heavy	Many
4	Downhill	Heavy	Many

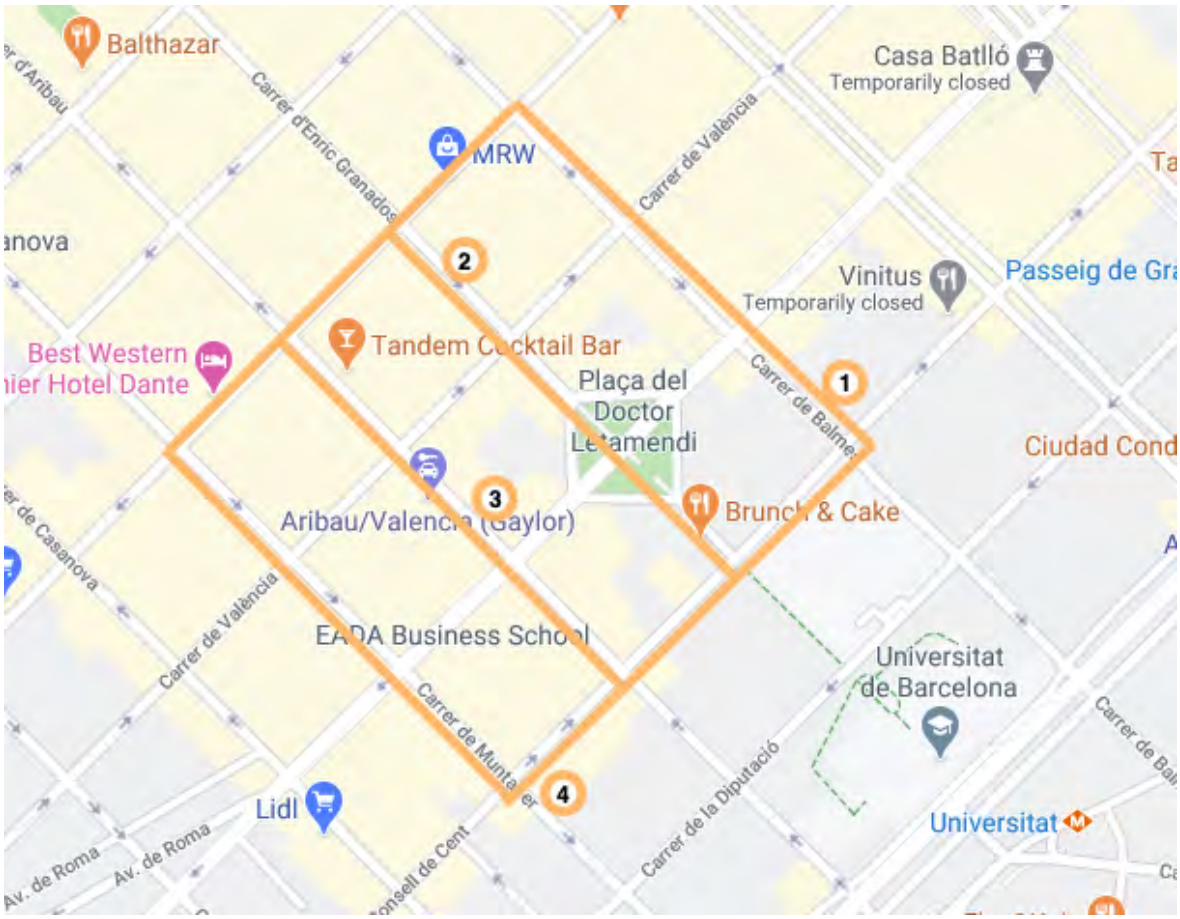


Figure I.1: Studied area in Eixample, Barcelona, with numbers 1 to 4 representing the positions of acoustic sensors in these streets. Source: Google Maps (last access 26/07/2020).

Therefore, the first campaign was carried out on the day 17 of March of 2018, and resulted in two audio files. The first audio file has a duration of 124 minutes and 13 seconds and the second audio file has a duration of 115 minutes and 27 seconds. The second campaign took place on the 9 of June of 2018 and resulted in a single audio file of 124 minutes and 13 seconds just like in the first campaign. Hence, the presented dataset has a total duration of 363 minutes and 53 seconds.

Despite having acoustic sensors in all the locations mentioned above, these had technical limitations. They were unable to make recordings for long periods of time (just for a few seconds) and could only store the sound level and frequency, as well as the time when a noise event went over a particular dB level. For this reason, we chose to use a ZOOM H5 (*H5*

Figure I.2: Photos from the studied streets in Eixample, Barcelona, with numbers 1–4 representing the views of the streets close to the acoustic sensors. Picture **1** is Balmes street, picture **2** is Enric Granados street, picture **3** is Aribau street, and picture **4** is Muntaner street. Source: Google StreetView.

Handy Recorder - Operation Manual 2014) recorder with an attached microphone working at 44100 Hz with a microphone sensitivity of -45dB/Pa that saved the recordings in *.WAV* format instead, as shown in Figure I.3.a. The recorder was placed close to the location of the acoustic sensor 3, in a first floor balcony at 4.5 meters above the street level (see Figure I.3.b) This also allowed us to record the sound without intervening in the street and altering people or car's behavior for placing the equipment at the sidewalk. Finally, two technicians, standing in the street under the recorder, observed the area and took independent notes about the noises and activities throughout the recording campaign in order to facilitate data labelling and analysis.

I.5 Data Labeling

In order to create a dataset that could be used to train artificial intelligence (AI) algorithms, we labeled each audio event using the Audacity program. This is an audio recording and editing software that allows one to name sections of sound (i.e., noise events) and associate a text label to them, as shown in Figure I.4. The result for each of the audio files recorded was a text file (.txt) containing the beginning and end of each section as well as their corresponding labels, with the following structure: "`starting_time_event(seconds) ending_time_event(seconds) label`". The seconds are always referenced to the beginning (second 0) of each individual audio file.

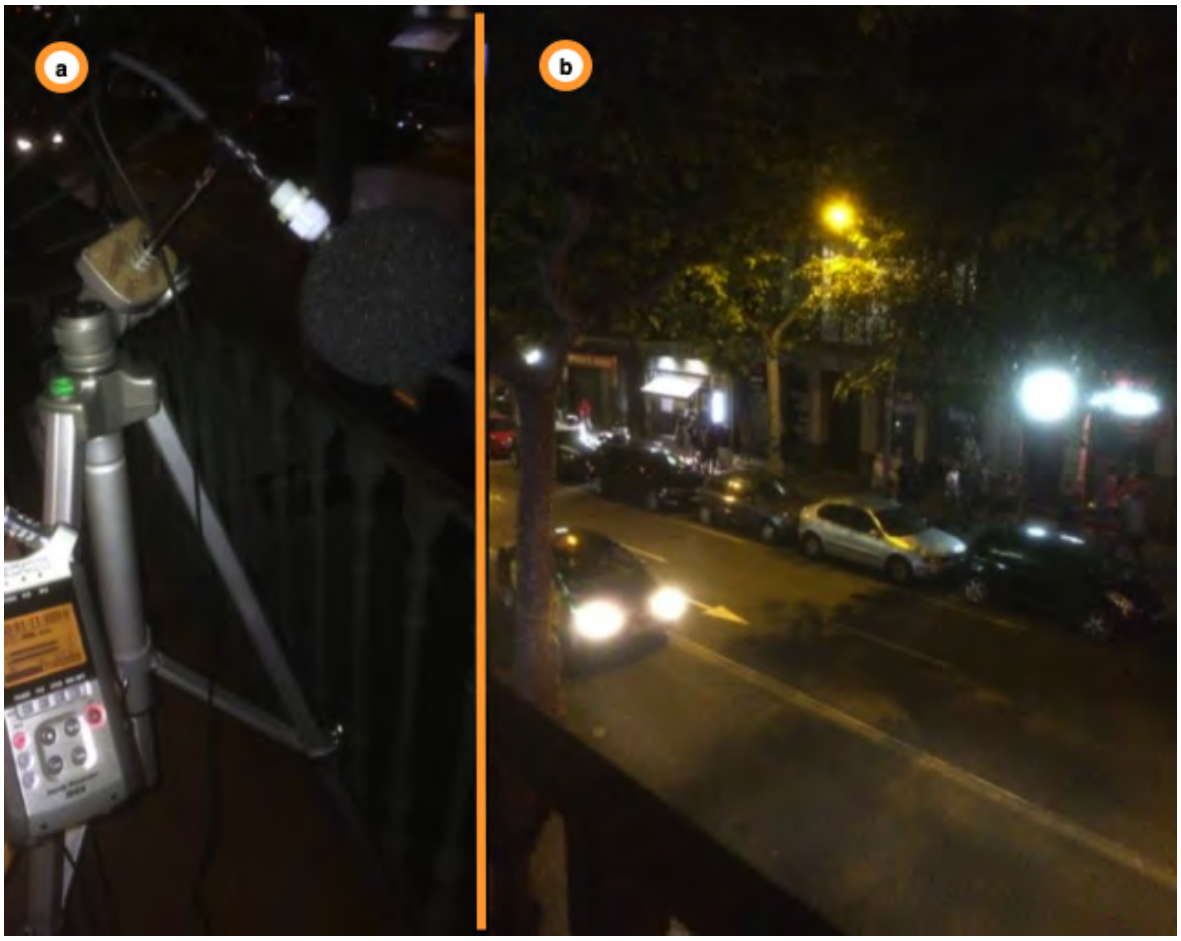


Figure I.3: Photos of the recording device and its relation to the street level. (**a**) Shows the Zoom recorder on a first-floor balcony, while (**b**) shows the view of the street from this balcony.

The labeling process was independently carried out by two technicians. Careful and consistent labeling is very important to ensure an effective training of AI algorithms. For this reason, fragments of the labeled audio were cross-checked by experts of the LIFE+ DYNAMAP project (Sevillano et al. 2016), who have extensive experience in labeling similar recordings. The labeling process was repeated up to three times, until experts confirmed that the labels from the different fragments were consistent. The resulting dataset is composed of the events and their respective labels, which are shown in Table I.2.

Since the main goal for which this dataset was created was to study the distinction between leisure and traffic in the city of Barcelona, the third column in Table I.2 also shows how each event was classified between the leisure and traffic categories. We have classified as leisure all sounds related to people, blinds (bars and restaurants), and music. The traffic category, on the other hand, contains sound events related to vehicles. The authors would like to highlight that the *rtn* event represents the road traffic noise produced by different vehicles. The sound of vehicles could be considered background noise in cities instead of an acoustic event, as it does not have a clear start and stop time and it is more or less stationary. However, as the purpose of this dataset is to compare the impact of traffic noise and leisure noise in the city center of Barcelona, only the road traffic noise that presented a noise level high enough

to mask any other events occurring simultaneously has been tagged as *rtn*. This will be compared to the rest of the acoustic events in future analyses in Section I.6.

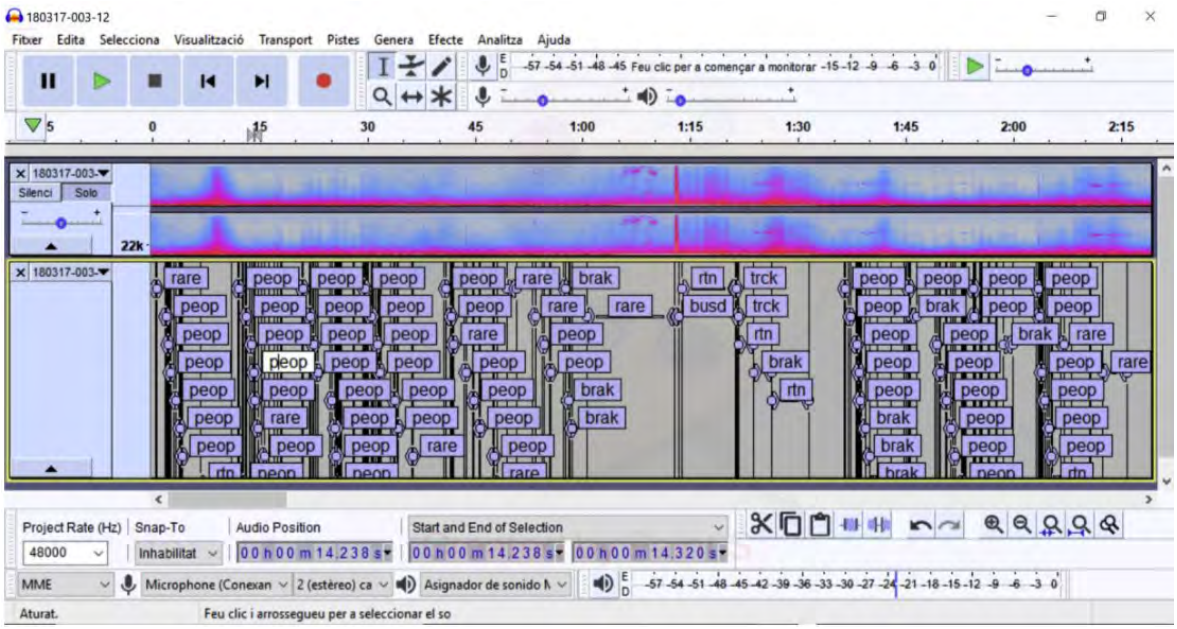


Figure I.4: Screenshot of the Audacity program showing a labeled audio fragment.

It is also worth mentioning the *rare* noise event, which does not belong to any of the aforementioned categories. Such events are sounds not easily recognized by a human—their source was not possible to determine, even after consulting the notes taken by the observers during the recording campaigns—or a mix of sound events that cannot be classified in a single category, such as two events occurring at the same time (e.g., high-level *rtn* and high-level *peop* events occurring at the same time).

Table I.2: Event types considered for the dataset and their respective descriptions and categories (leisure/traffic).

Label	Event	Category
bkmu	A mix of background city noise and music	Leisure
blin	The opening and closing of blinds	
coug	Person coughing	
door	Door or knock noise (house, car, or object)	
musi	Music in a car or in the street	
peop	People talking	
troll	Sound of wheels of suitcases (trolley)	
whtl	Whistle	
brak	Noise of brake or car's timing belt	Traffic
busd	Opening of a bus or tramway door	
horn	Horns of vehicles (cars, motorbikes, or trucks)	
rtn	High-intensity road traffic noise	
sire	Sirens of ambulances or police cars	
rare	Unrecognizable noise	None

I.6 Dataset Analysis

After recording and labeling the audio files, a detailed analysis was performed in the dataset to determine the main features of the sounds. Table I.3 shows the numbers of events detected on each of the audio files of the recording campaigns. As can be observed, the class that presents the most events is—by far—*peop*, followed by *rtn*, *brak*, *door*, and *rare*. As *peop* and *door* are categorized as leisure sounds, we can deduct that the zone where the audio files were recorded contains mostly leisure-time noises. In order to be able to confirm this deduction, deeper analyses were carried out.

Table I.3: Number of events labeled on each audio file and their durations in seconds.

Event	File #1	File #2	File #3	Total
bkmu	11 events, 49.65 s	0 events, 0 s	7 events, 25.61 s	18 events, 75.27 s
blin	1 event, 2.66 s	10 events, 6.34 s	9 events, 29.41 s	20 events, 38.42 s
brak	520 events, 428.51 s	134 events, 144.68 s	156 events, 239.57 s	810 events, 812.77 s
busd	15 events, 8.62 s	36 events, 14.22 s	12 events, 9.42 s	63 events, 32.27 s
coug	15 events, 5.50 s	0 events, 0 s	5 events, 4.14 s	20 events, 9.65 s
door	455 events, 108.85 s	138 events, 52.69 s	138 events, 95.46 s	731 events, 257.01 s
horn	36 events, 29.17 s	21 events, 22.46 s	36 events, 37.54 s	93 events, 89.19 s
musi	12 events, 51.88 s	2 events, 7.95 s	2 events, 9.15 s	16 events, 68.95 s
peop	810 events, 578.93 s	1068 events, 803.82 s	578 events, 2383.50 s	2456 events, 3765.86 s
rare	404 events, 787.65 s	135 events, 319.68 s	132 events, 474.50 s	671 events, 1581.84 s
rtn	574 events, 2132.89 s	343 events, 3789.41 s	233 events, 1968.52 s	1150 events, 7889.90 s
sire	4 events, 36.25 s	2 events, 16.41 s	3 events, 23.74 s	9 events, 76.42 s
troll	8 events, 25.38 s	1 event, 0.56 s	2 events, 8.35 s	11 events, 34.30 s
whtl	2 events, 0.91 s	0 events, 0 s	6 events, 7.13 s	8 events, 8.05 s
Total	2867 events, 4245.81 s	1890 events, 5178.34 s	1319 events, 5316.11 s	6076 events, 14,739.95 s

Apart from the number of occurrences of each type of sound, the duration of each of the events is important when considering the noise impact of each class. Hence, a boxplot displaying the duration of each labeled situation is presented in Figure I.5. As can be observed in the figure, on average, the class that presents the greatest duration is *sire*, but considering that there are only nine samples of this type of sound, this event type may not be as relevant as other classes with greater numbers of occurrences. However, it is worth noting that the class that contains more events (i.e., *peop*) is usually short in time (less than 1 s on average) in comparison to other events that also appear several times, such as *rtn*. The higher number of occurrences of the *peop* class and the short duration of each of the occurrences are balanced with the fewer number of occurrences of the *rtn* class and the longer duration of each of the occurrences of this class. This fact is explained by the characteristics of each noise source: *peop* is labeled each time anybody speaks, as a conversation is mainly not considered a continuous event; *rtn* is usually considered a continuous event, despite that it contains several passes or other vehicles.

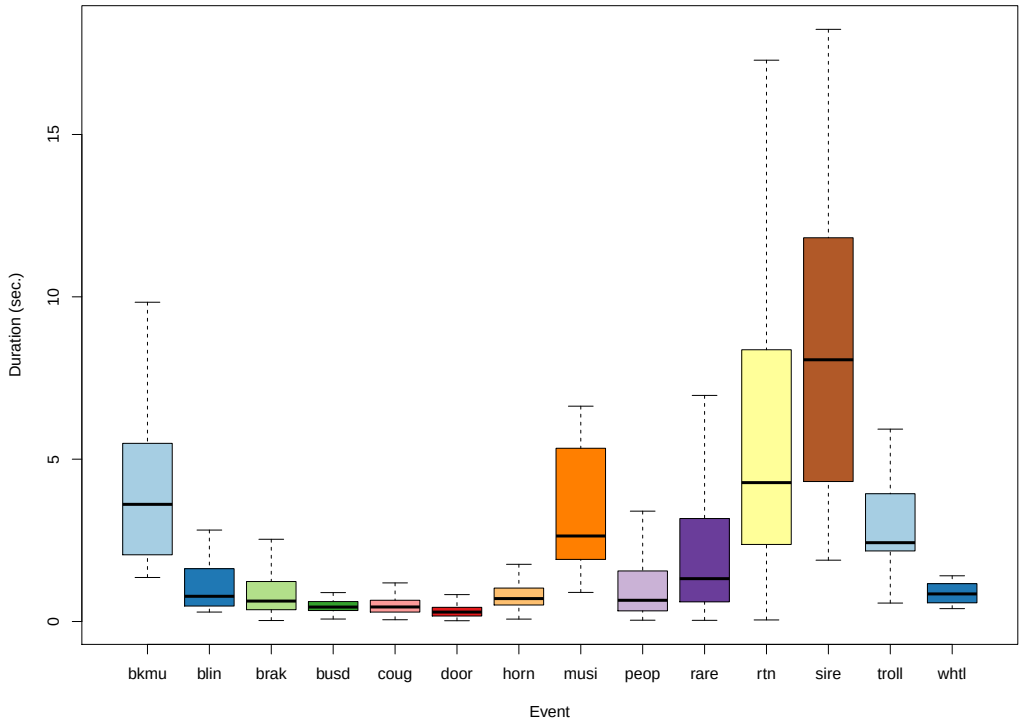


Figure I.5: Boxplot of the durations (in seconds) of the labeled events for each of the classes of the dataset.

I.6.1 Signal-to-Noise Ratio Calculation

In fact, calculating the duration of an audio event is meaningless without considering the level of noise that it produces. Indeed, short events with a high level of noise (impulse noises) are usually perceived as more annoying by neighbors in comparison to long events with low noise level (Jarup et al. 2005). Hence, two important parameters to be taken into account are the SNR (signal-to-noise ratio) and the impacts of the different events. For this reason, and to be able to compare the traffic noise against the leisure noise, a boxplot of the SNR is shown in Figure I.6.

To calculate the SNR, and as the background noise of the different labeled events is not stationary, we applied the methodology detailed in (Orga et al. 2017). That is, we first calculated the power of the spectrum of the labeled event (considering that the event is the "Signal"), and then obtained the power of the background noise by getting samples from before and after the event. After that, we divided the power of the signal and the power of the noise to obtain the final value of the SNR in dB. This means that the obtained value is always relative to the sounds that happen right before and right after the event. In the case that an event is followed by signal with more power, the SNR would have a negative value, indicating that the event is less noisy than its environmental noise before and after the event.

As an example, Figure I.7 depicts the spectrogram of an event labeled as *door*. In the figure, samples used as "Signal" or "Noise" have been marked with arrows. The N central samples (labeled as *door*) were the ones used to calculate the power of the signal, and the $\frac{N}{2}$

samples before and after the event were used to calculate the power of the noise. This means that the SNR was computed as:

1. Power of the event:
$$PS = \frac{\sum_{i=1}^{N} Signal\ sample_i^2}{N}. \tag{I.1}$$

2. Power of the background noise around the event, considering the $\frac{N}{2}$ samples before and after the event depicted in Figure I.7:
$$PN = \frac{\sum_{i=1}^{N/2} Noise\ 1\ sample_i^2 + \sum_{i=1}^{N/2} Noise\ 2\ sample_i^2}{N}. \tag{I.2}$$

3. Finally, the ratio is calculated and converted to dB:
$$SNR = 10 \log_{10}(\frac{PS}{PN}). \tag{I.3}$$

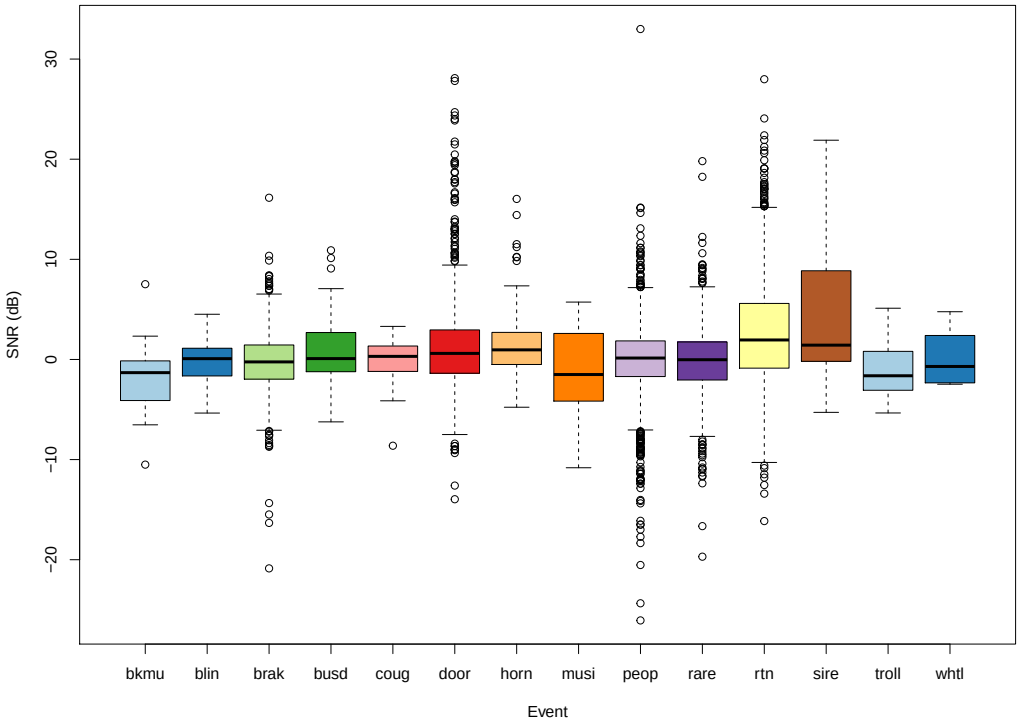


Figure I.6: Boxplot of the signal-to-noise ratio (SNR; in dB) of the labeled events for each of the classes of the dataset.

From this analysis, we can conclude that the "traffic" events have, on average, a higher value of SNR. Concretely, the events that have, on average, the highest SNR values are *sire*, *horn*, and *rtn* (classes included in the traffic category). Considering together the durations and SNRs of these events, we can see that, whereas road traffic noise and siren sounds typically present a duration of few seconds and a high value of SNR, the *horn* event is almost an impulse noise (very short in time and with a high level of SNR). However, we can see in the boxplot that a few occurrences of *rtn* and *sire* events have SNR values of around 20 dB,

meaning that independently of their duration, they present an extremely high noise level in comparison to their surrounding environment.

Regarding the *door* events, we can see that, whereas some occurrences have a high SNR (reaching a maximum of about 30 dB), some other occurrences present negative values. The main reason behind this phenomenon is that the dataset contains two main door types tagged with the same label. On the one hand, the closing of car doors has been labeled as *doors*. Because of the materials of the car, and as they are very heavy, when people close these types of doors, they make an impulse sound (very short in time and with a lot of energy). One example of this type of door event is shown in Figure I.7. On the other hand, doors related to leisure places, such as such as bars, have also been tagged as *doors*. These doors are lighter and typically present lower levels of energy.

Another interesting observation is that all the musical events (*musi* and *bkmu*) present SNR values smaller than 0 dB on average, meaning that they are less noisy than their surrounding environmental noise. However, something that must be taken into account is that, when labeling the recorded audio files, the authors noticed that all the musical sounds originated from cars passing by with their windows opened, so they were all surrounded by *rtn*. Hence, as musical sounds are always surrounded by an event that typically presents a positive SNR by itself, they are partially masked in the recordings.

Finally, analyzing the most common event in the dataset, *peop*, we can see that some occurrences present a positive SNR and some other occurrences present a negative value. The main reason for this is that, during the recording campaign, there were two types of people in the street. On the one side, there were people walking by the street and talking normally to each other. On some occasions, these occurrences were masked by other events or could not be distinguished from background noise, so they present negative values of SNR. On the other side, there were people standing in the street and having loud conversations—that even included a few shouts—close to the recording sensor. These are the occurrences that present high SNR values.

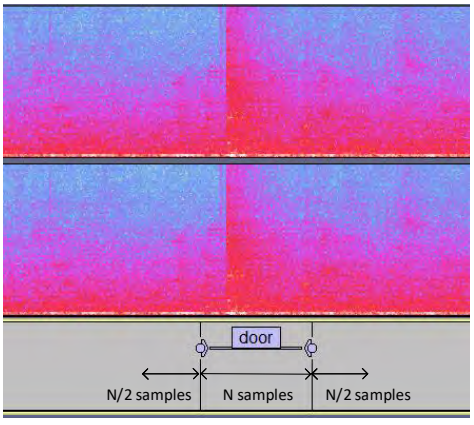


Figure I.7: Spectrogram of a *door* event indicating which samples were used as signal or noise for the SNR calculation.

I.6.2 Event Impact Analysis

Apart from the SNR calculation, we also calculated the impact of each event. The impact measures the contribution of the labeled event over the equivalent level of a certain period of time after applying the A-weighting filter (St Pierre Jr and Maguire 2004). This indicator was calculated following the methodology explained in (Orga et al. 2017). As in the cited work, the impact is relative to the 5 min of L_{Aeq} measured surrounding the event. To obtain the final impact value, the L_{Aeq} of the signal is obtained by first applying the A-weight filter and then obtaining the equivalent level. Then, the labeled event is removed from the audio file and replaced by an interpolated value of the background noise to maintain a continuous energy of the signal. Finally, the impact is measured as the subtraction between the initial L_{Aeq} and the L_{Aeq} without the labeled event. For more details about this procedure, the reader is referred to (Orga et al. 2017).

The value of impact of an event is highly related to the type of event that is being measured. If it is an event that presents a high value of SNR or its duration is long, the impact will have a high value. If both conditions are met (the SNR is high and the duration is long), the impact will be extremely high. As there are some events that usually have similar durations (e.g., a *door* event is not likely to last more than 1 s, while a *sire* event will often last for several seconds), the impacts of the events from the same class may have similar values. Impact values can be deduced by looking at the boxplots presented in Figures I.5 and I.6. Events that present smaller boxes in the boxplots (such as the *busd*, *cough*, or *door* will typically have similar values of impact. However, events that present bigger boxes, such as the *rtn* or *sire*, will have a wider range of impact values, as the duration and SNR can be very different when comparing events belonging to the same class.

Figure I.8 shows the impact of all the labeled events divided among the three recording campaigns. As expected, the Figure suggests that the events that have longer durations also present the highest impact values. It can also be observed that both the traffic sounds and the leisure sounds have similar values of impact (the circles presented in the Figure have similar sizes). Actually, on the one hand, only a few events present an increase of L_{Aeq} greater than 0.01, which means that all the events have similar contributions to the noisiness of the environment. On the other hand, there are several events—which usually have a duration close to milliseconds—that present a negative impact, which means that their noisiness is lower than the average background noise. The results observed are highly correlated with the two previous boxplots (Figures I.5 and I.6). More precisely, the most remarkable event is the *rtn*, which, apart from being the class that presents occurrences with the highest durations, also presents the biggest circle, meaning that the impact is more notable.

Figure I.8 is also useful to see the duration, SNR, and impact of the events for each individual audio recording, as there are several notorious differences between the features of the events of the different classes. Whereas in the first and second files, the *peop* class typically has a duration smaller than 10 s (with only two exceptions in audio file #1), the third audio file presents several samples of *peop* talking in the street with longer duration. Given that the recording campaigns took place on different days (audio files #1 and #2 were

recorded one day and audio file #3 was recorded on a another day), it is normal that the number of people in the street standing close to the sensor is slightly different.

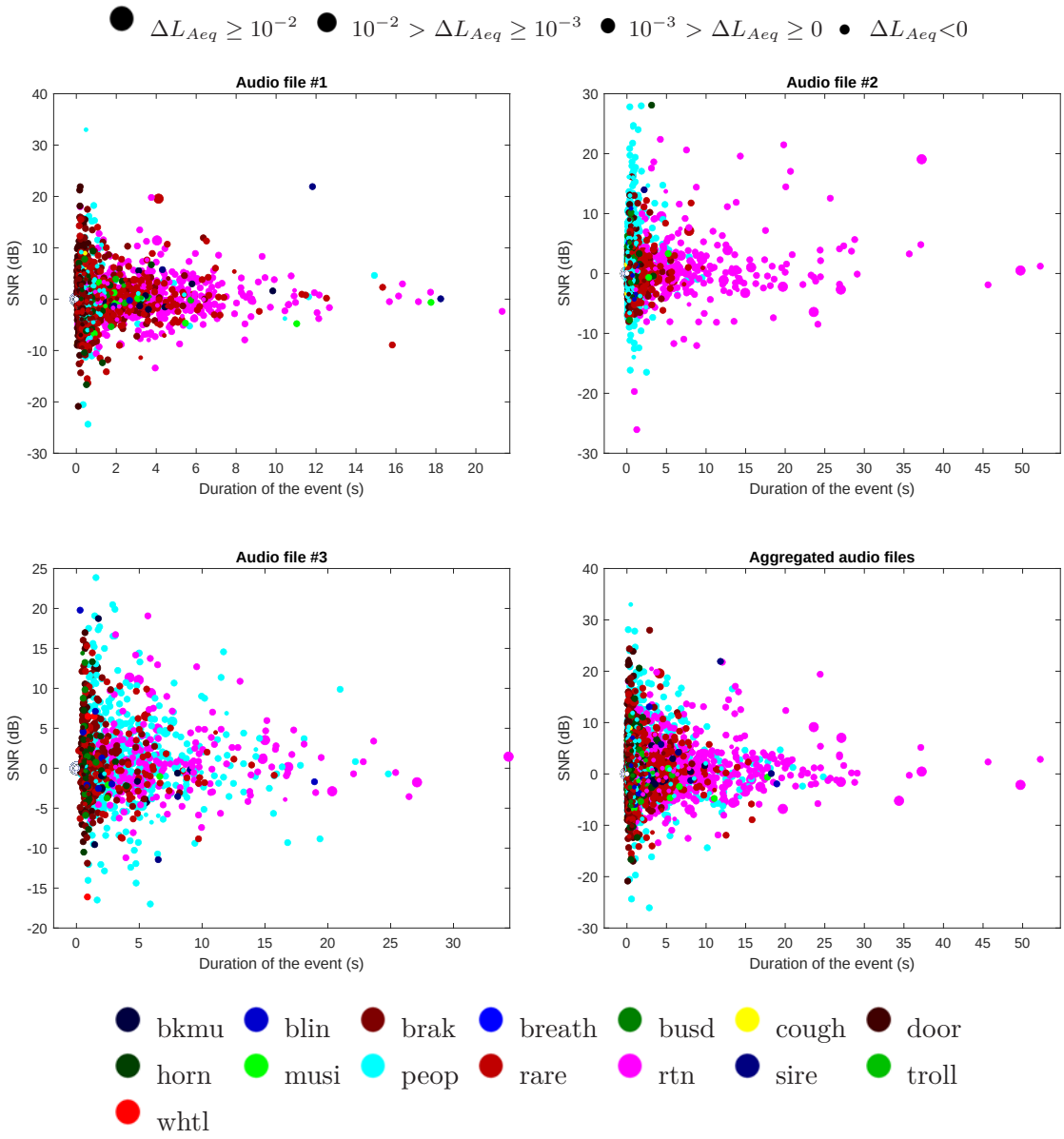


Figure I.8: Analysis of the impacts of the audio events.

I.6.3 Analysis of the Time–Event Distribution

Figure I.9 depicts the occurrence of the events in time. The figure shows three sub-plots, each one representing one of the audio files used to generate the dataset. The x-axis of each sub-plot represents the time of the audio file in minutes, and the y-axis contains the 14 different possible categories of the labels shown in Table I.2. To display each event's occurrence, a colored dot has been drawn at its starting second in the x-axis and at the height of its label. For example, if a *whtl* event happened at minute 0 in the third audio file, a dot would be drawn around the top-left corner of the last sub-plot.

The color of the dot represents the SNR value of that event, calculated as explained in

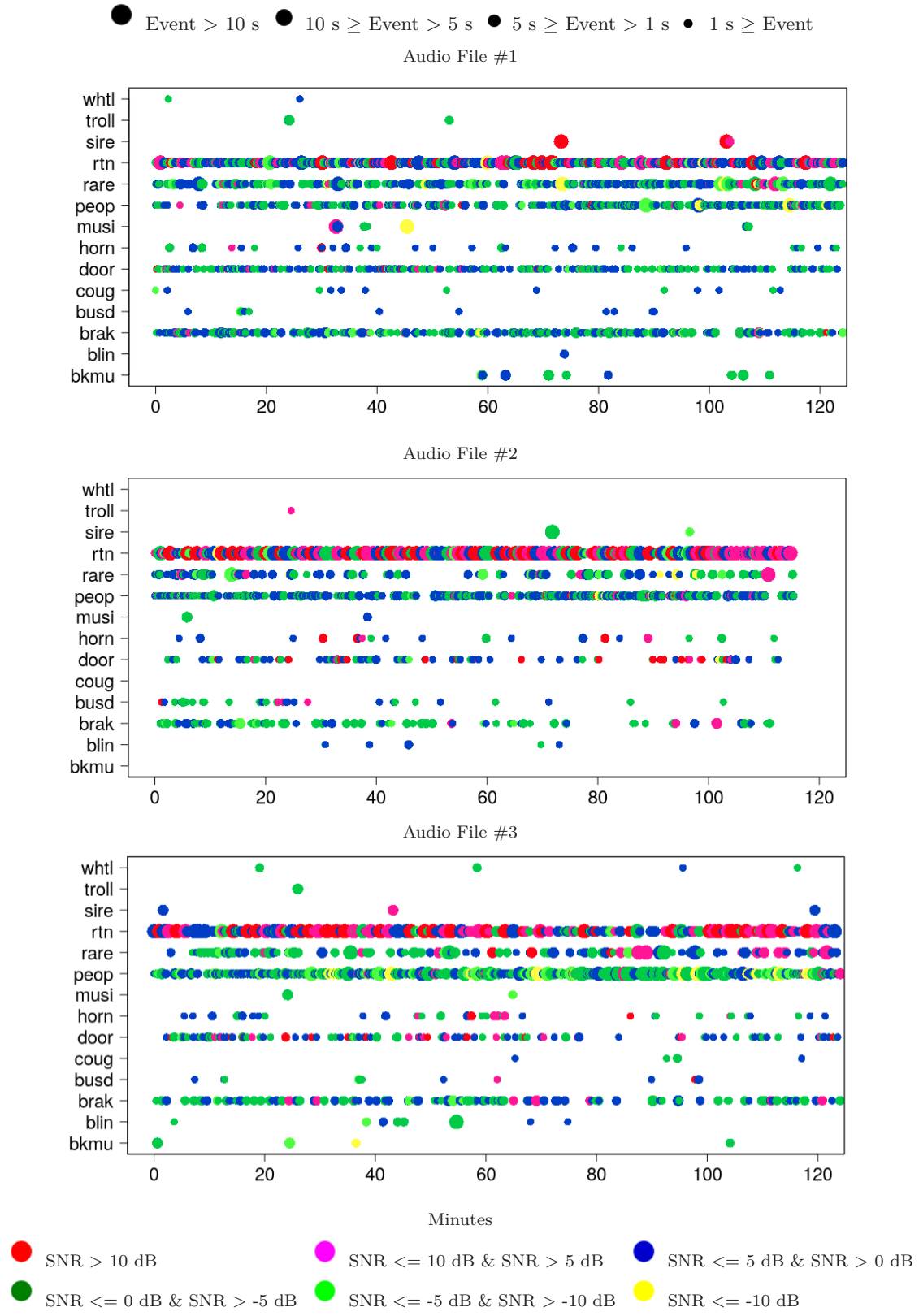


Figure I.9: Distribution of the labeled events of each audio file in time. Each subplot stands for the results of an audio file. The x-axis is the time in minutes and the y-axis represents the different labeled categories that can be found in the dataset. Each dot corresponds to an event of the y-axis type starting at the moment indicated in the x-axis. The color of a dot represents the SNR of that concrete event and the size of the dot represents the duration of that event.

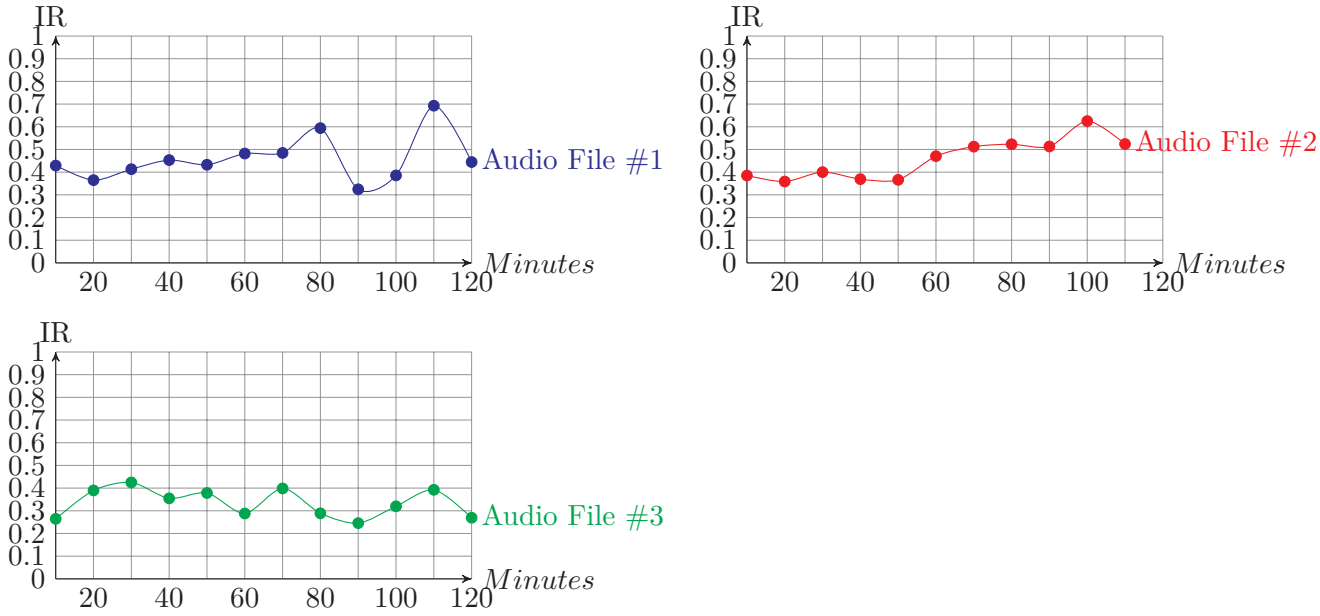


Figure I.10: Intermittency Ratio of the three audio files presented in the dataset calculated with windows of 10 minutes. The y-axis represents the IR of each audio file and the x-axis represents the time evolution (in minutes) of the audio file that is being evaluated.

Section I.6.1. The red, purple, and blue dots represent the events that have a positive SNR, and the rest of dots represent the events that have a negative SNR. As shown in the plots, typically, the events that present higher SNR in the three audio files are *rtn* and *sire*, as well as some *brak* and *door* events. Whereas *peop* is the category that has more occurrences, there are just a few events with SNR values greater than 10 dB. The size of each dot represents the duration of the event. Longer events are associated to bigger dots, whereas the events shorter than 1 s are represented with the smallest dots.

Regarding the time distribution of the events, we can see that there is not an accurate pattern for the occurrences of events of different classes or among the different audio files. Both traffic and leisure noises happen all along the audio files in a uniform distribution. The events that are more stable in the dataset are *peop* and *rtn*; they are present during all the audio files (we can see there are dots over all the horizontal axis, creating almost a constant line for these two categories). Then, audio files number 1 and 3 present higher occurrences of *door* and *brak* events, and they are also distributed across the audio files. In the second audio file, however, the events of those types are present mainly at the beginning, and there are just a few occurrences at the end. A remarkable fact is that, concretely, the few *door* events happening at the end of the second audio file (from minute 90 to minute 110), are the ones that present higher SNR values in that category.

Looking at the sizes of the dots, we can observe that even though the number of occurrences of *peop* and *rtn* seems to be constant, there is a considerable difference regarding the duration of these two events. In the three audio files, the *rtn* category has more occurrences of long events, and the SNR value is greater, too. This is consistent with the results previously observed in Figure I.8.

I.6.4 Analysis of the Intermittency Ratio

As a final analysis, the intermittency ratio (IR) has been calculated and is presented in Figure I.10. The intermittency ratio is a metric that was first introduced by Wunderli et al. in (Wunderli et al. 2016) and measures the "eventfulness" of a traffic environment. Concretely, it reflects the contributions of events that surpass a certain threshold to the total amount of energy in a certain period of time, measuring the impact on the total L_{Aeq} of all the individual loud events. Concerning the focus of the contribution of this work, with a standard metric, the IR supports the idea that the events detected and labeled (e.g., *rtn*) present a clear impact on the global value of the equivalent level measured in the street.

The procedure to calculate the ratio is as follows (Wunderli et al. 2016; Brambilla et al. 2019): First, the equivalent level of energy of a window of size T is calculated as $L_{eq,T,tot}$. This is the amount of energy contained inside the window. For this study, we chose a window of 10 min, following the fact that the L_{Aeq} is mainly stationary, and we intend to define the impacts of the events with the shortest window frames possible to evaluate the differences in the axes of the time series. This trade-off time window allows us to have 12 IR values for each audio file (except in the case of Audio File #2, where we only have 11 values because the audio file is shorter). Then, the equivalent level of energy of each 1-s fragment inside the window was also calculated. In order to follow the methodology of the previous works (Wunderli et al. 2016; Brambilla et al. 2019) and to be able to obtain comparable results, those 1-s fragments presenting a L_{eq} greater than $L_{eq,T,tot} + 3$ dB were considered as "events", independently of their labels in the dataset. Then, considering all the 1-s windows that surpassed the +3 dB threshold, the Heaviside step function was applied to remove the non-event sounds inside the 10-min window, and a new $L_{eq,T,events}$ was calculated to obtain the energy of only those 1-s fragments presenting a L_{eq} level greater than $L_{eq,T,tot}+$ 3 dB. Finally, the ratio was calculated by dividing the $L_{eq,T,events}$ by $L_{eq,T,tot}$.

Summarizing, we used the next three equations to obtain the IR of each 10-min window of the three audio files of the dataset:

$$L_{eq,T,tot} = 10 \log_{10} \left(\frac{\sum_{n=0}^{N} X[n]^2}{N} \right) \text{[dB]}, \tag{I.4}$$

where $X[n]$ are the samples of the audio file and N is the number of samples of a window (in our case, 10 min times the sampling frequency of the audio file).

$$L_{eq,T,events} = 10 \log_{10} \left(\frac{\sum_{n=0}^{N} H[X[n] - K] X[n]^2}{N} \right) \text{[dB]}, \tag{I.5}$$

where $H[X[n] - K]$ is the Heaviside step function and K is the $L_{eq,T,tot}$ plus the threshold—in our case, set to 3 dB, as in (Brambilla et al. 2019).

$$IR = \left(\frac{10^{0.1 L_{eq,T,events}}}{10^{0.1 L_{eq,T,tot}}} \right) \tag{I.6}$$

To interpret the results of this ratio, we have to consider that, on the one hand, a ratio higher than 0.5 indicates that more than half of the energy of the signal is due to events

(understanding events as parts of the signal with L_{eq} greater than K). This situation occurs when the events clearly stand out from the parts of the signal that are not considered as events (i.e., background noise), meaning that the ratio is high when the amount of energy of the events is significantly higher than the background noise. On the other hand, a small ratio value means that the amount of energy of the events is considerably low compared to the background noise. This situation occurs when the noise level of the events is close to the threshold. Hence, a small ratio in a street does not mean that the general noise level is lower than the noise level in a street with a higher ratio, but that the noise level of the events is similar to the background noise.

Evaluating the results in Figure I.10, we can conclude that the three audio files presented in the dataset have an IR ranging from 0.24 (minute 90 from the Audio File #3) to 0.69 (minute 110 from Audio File #1). The relatively low values of IR (comparing them, for example, with the ones obtained from measurements in a local street in the work of Brambilla et al. in (Brambilla et al. 2019)), together with the impact values and number of events analyzed in previous sections (Figures I.3 and I.8), suggest that the street where the recordings took place is pretty noisy in terms of background noise, and the energy contribution is balanced between background noise and events. Nevertheless, there are several IR evaluations over 50%, so, at certain moments, mainly passes of *rtn* and other events related to leisure (e.g., *peop* shouting and others) can have a relevant contribution to the value of L_{Aeq} total of the acoustic file.

I.7 Materials

The labeled dataset can be downloaded from https://doi.org/10.5281/zenodo.3956503. The dataset is structured in six files: three audio files (.wav) and three label files (.txt). The two audio files recorded in the first campaign have been named *File-1.wav* and *File-2.wav*, and the audio file recorded in the second campaign has been named *File-3.wav*. The names of the label files belonging to each of the audio files follow the same naming scheme, adding a *_labels* at the end of the name (e.g., File-1_labels.txt).

I.8 Conclusions

This work has presented the creation and the analysis of a real-life environmental audio dataset in the district of Eixample, Barcelona. The dataset is composed of six hours of audio, was recorded on two Saturdays between 22:00 and 03:00, and contains 14 types of events, with a total of 6076 event occurrences. These events were classified into two categories: leisure and traffic, with the exception of the *rare* events, whose sources were not possible to determine or were a mix of sounds. The most common type of noise event is *people*, followed by road traffic noise, *brakes*, *doors*, and *rare* events. The fact that people and door events are among the most common indicates that the area and time chosen for the recording campaigns are suitable to measure the impact of leisure activities.

An SNR analysis comparing traffic noise with leisure noises revealed that traffic events have, on average, a higher value of SNR; *siren*, *horn*, and *road traffic noise* have the highest

ones. An impact analysis suggested that the events with longer durations also had the highest impact. Both traffic and leisure had similar values of impact, but only a few had an increase in L_{Aeq} of greater than 0.01, meaning that they, in fact, contribute in a similar way to the noisiness of the environment.

The time event distribution indicates that the noises of people and traffic are constant during the recording, with the traffic noises being longer and confirming the greater SNR values observed previously. Finally, an analysis of the intermittency ratio shows that the recordings present low values of IR compared to other studies, which, together with the impact values and the number of events, indicates a noisy street with a balanced contribution of energy background noise and events over 3 dB, with some punctual exceptions, including loud *peop* or *rtn* street noise.

The dataset described in this paper is open and freely available to the community and may be used for different purposes. It can also be extended by means of further recordings or data augmentation, and also combined and compared with other datasets. A greater understanding of leisure and traffic events at night could help policymakers to regulate the noise produced in leisure locales, such as restaurants, bars, or discotheques. In particular, if an automatic sensor-based system is implemented to reliably distinguish and measure leisure activities, city councils would be able to continuously measure such noises, both in specific places in the city and during local festivities. Such information could inform urban planners and provide evidence to change the design of certain places in the city to improve the soundscape perceived by the neighbors.

Our future work is centered on validating the completeness of the dataset published. For this purpose, we plan to record another pair of days close to the locations of the other sensors in Figure I.2, as pointed out by our colleagues from the Barcelona City Council. If more leisure events are detected in the new recordings, we would complete this corpus before proceeding to the event detection. Having more data points would also allow us to correlate the types of noises identified with the numbers of complaints in each area, which might give us some pointers for the types of noises that are most annoying to neighbors. In addition, given that the analysis focused on commonly used metrics in well-being- and health-related studies, we might be able to compare the characteristics of the noises observed with similar studies in other cities (Dratva et al. 2012; Hofman et al. 1995; Wunderli et al. 2016; Ottoz et al. 2018; Easteal et al. 2014).

Furthermore, a deeper analysis of the impact of the labeled sounds will be conducted with a wider comparison between events belonging to the same category in order to determine differences between the impacts of each event, with special focus on the surrounding environmental noise, which is a key issue for the evaluation of SNR and impact. Once these analyses are conducted, the *rare* category has to be vertically analyzed in order to determine which types of events usually correspond to that fuzzy label; they cannot be classified inside any of the other categories, but maybe we can make more acoustic information about all the *rare* events available. Finally, we plan to train an ML system (similar to (Socoró et al. 2017)) in order to automatically classify noise events at night, taking into account at

least leisure and traffic.

Author's contributions

All authors have significantly contributed to this work. E.V.-V. contributed to the tagging of the recordings and writing and carried out the data analyses. L.D. was involved in the project conceptualization and coordination, as well as writing. R.M.A.-P. participated in the tagging and writing, and also offered conceptual and technical support. F.P. and H.V. carried out the recording campaigns, participated in the tagging, and carried out some preliminary data analyses. All authors have read and agreed to the published version of the manuscript.

Funding

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 712949 (TECNIOspring PLUS) and from the Agency for Business Competitiveness of the Government of Catalonia. The authors thank the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement (Generalitat de Catalunya) under grant 2017 SGR 966 and the Ramon Llull University (ref. 2020.URL-Proj-053).

Acknowledgements

The authors would like to thank our colleagues from the Environmental Department of Barcelona City Council for the guidance in this work, especially to Julia Camps.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANEs	Anomalous Noise Events
ANED	Anomalous Noise Event Detector
AI	Artificial Intelligence
CNOSSOS-EU	Common Noise Assessment Methods in Europe
END	Environmental Noise Directive 2002/49/EC
IR	Intermittency Ratio
SNR	Signal-to-Noise Ratio
WASN	Wireless Acoustic Sensor Networks

References

Alías, Francesc and Alsina-Pagès, Rosa Ma (2019). 'Review of Wireless Acoustic Sensor Networks for Environmental Noise Monitoring in Smart Cities'. In: *Sensors* vol. 2019, p. 13.

Alías, Francesc, Orga, Ferran, Alsina-Pagès, Rosa Ma and Socoró, Joan Claudi (2020). 'Aggregate Impact of Anomalous Noise Events on the WASN-Based Computation of Road Traffic Noise Levels in Urban and Suburban Environments'. In: *Sensors* vol. 20, no. 3, p. 609.

Alías, Francesc and Socoró, Joan Claudi (2017). 'Description of anomalous noise events for reliable dynamic traffic noise mapping in real-life urban and suburban soundscapes'. In: *Applied Sciences* vol. 7, no. 2, p. 146.

Alsina-Pagès, Rosa Ma, Alías, Francesc, Socoró, Joan Claudi and Orga, Ferran (2018). 'Detection of Anomalous Noise Events on Low-Capacity Acoustic Nodes for Dynamic Road Traffic Noise Mapping within an Hybrid WASN'. In: *Sensors* vol. 18, no. 4, p. 1272.

Alsina-Pagès, Rosa Ma, Orga, Ferran, Alías, Francesc and Socoró, Joan Claudi (2019). 'A WASN-Based Suburban Dataset for Anomalous Noise Event Detection on Dynamic Road-Traffic Noise Mapping'. In: *Sensors* vol. 19, no. 11, p. 2480.

Basner, Mathias, Müller, Uwe and Elmenhorst, Eva-Maria (Jan. 2011). 'Single and Combined Effects of Air, Road, and Rail Traffic Noise on Sleep and Recuperation'. In: *Sleep* vol. 34, pp. 11–23.

Bello, Juan P., Silva, Claudio, Nov, Oded, Dubois, R. Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles and Doraiswamy, Harish (2019). 'SONYC: A System for Monitoring, Analyzing, and Mitigating Urban Noise Pollution'. In: *Communications of the ACM* vol. 62, no. 2, pp. 68–77.

Bellucci, Patrizia, Peruzzi, Laura and Zambon, Giovanni (2017). 'LIFE DYNAMAP project: The case study of Rome'. In: *Applied Acoustics* vol. 117, pp. 193–206.

Brambilla, Giovanni, Confalonieri, Chiara and Benocci, Roberto (2019). 'Application of the intermittency ratio metric for the classification of urban sites based on road traffic noise events'. In: *Sensors* vol. 19, no. 23, p. 5136.

Cartwright, Mark, Mendez, Ana Elisa Mendez, Cramer, Jason, Lostanlen, Vincent, Dove, Graham, Wu, Ho-Hsiang, Salamon, Justin, Nov, Oded and Bello, Juan (2019). 'Sonyc urban sound tagging (sonyc-ust): a multilabel dataset from an urban acoustic sensor network'. In.

Cik, Michael, Lienhart, Manuel and Lercher, Peter (2016). 'Analysis of Psychoacoustic and Vibration-Related Parameters to Track the Reasons for Health Complaints after the Introduction of New Tramways'. In: *Applied Sciences* vol. 6, no. 12, p. 398.

Cox, P and Palou, J (2002). 'Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 Relating to the Assessment and Management of Environmental Noise-Declaration by the Commission in the Conciliation Committee on the Directive Relating to the Assessment and Management of Environmental Noise (END)'. In: *Annex I, OJ* vol. 189, no. 18.7, p. 2002.

Dratva, Julia et al. (2012). 'Transportation Noise and Blood Pressure in a Population-Based Sample of Adults'. eng. In: *Environmental health perspectives* vol. 120, no. 1, pp. 50–55.

Easteal, Matthew, Bannister, Simon, Kang, Jian, Aletta, Francesco, Lavia, Lisa and Witchel, Harry (Sept. 2014). 'Urban Sound Planning in Brighton and Hove'. In.

European Commission, Joint Research Centre—Institute for Health and Consumer Protection (2012). *Common Noise Assessment Methods in Europe (CNOSSOS-EU) for strategic noise mapping following Environmental Noise Directive 2002/49/EC*.

European Environment Agency, 2020 (n.d.). *The European environment — state and outlook 2020*. https://www.eea.europa.eu/soer/2020 (accessed on 01 April 2020).

Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N. and Vento, M. (Jan. 2016). 'Audio Surveillance of Roads: A System for Detecting Anomalous Sounds'. In: *IEEE Transactions on Intelligent Transportation Systems* vol. 17, no. 1, pp. 279–288.

Foggia, Pasquale, Petkov, Nicolai, Saggese, Alessia, Strisciuglio, Nicola and Vento, Mario (2015). 'Reliable detection of audio events in highly noisy environments'. In: *Pattern Recognition Letters* vol. 65, pp. 22–28.

H5 Handy Recorder - Operation Manual (2014). Zoom Corporation.

Heittola, Toni, Mesaros, Annamaria, Eronen, Antti and Virtanen, Tuomas (2013). 'Context-dependent sound event detection'. In: *EURASIP Journal on Audio, Speech, and Music Processing* vol. 2013, no. 1, pp. 1–13.

Hofman, W.F., Kumar, A. and Tulen, J.H.M. (1995). 'Cardiac reactivity to traffic noise during sleep in man'. In: *Journal of Sound and Vibration* vol. 179, no. 4, pp. 577–589.

Jarup, Lars et al. (2005). 'Hypertension and exposure to noise near airports (HYENA): study design and noise exposure assessment'. In: *Environmental health perspectives* vol. 113, no. 11, pp. 1473–1478.

Jarup, Lars et al. (2008). 'Hypertension and Exposure to Noise Near Airports: the HYENA Study'. eng. In: *Environmental health perspectives* vol. 116, no. 3, pp. 329–333.

Mesaros, Annamaria, Heittola, Toni and Virtanen, Tuomas (Aug. 2016). 'TUT database for acoustic scene classification and sound event detection'. In: *24th European Signal Processing Conference (EUSIPCO 2016)*. Vol. 2016. Budapest, Hungary: IEEE, pp. 1128–1132.

Mesaros, Annamaria, Heittola, Toni and Virtanen, Tuomas (2019). 'Acoustic scene classification in DCASE 2019 challenge: closed and open set classification and data mismatch setups'. In.

Nakajima, Yasutaka, Sunohara, Masahiro, Naito, Taisuke, Sunago, Norihito, Ohshima, Toshiya and Ono, Nobutaka (Aug. 2016). 'DNN-based Environmental Sound Recognition with Real-recorded and Artificially-mixed Training Data'. In: *Proc. 45th International Congress and Exposition on Noise Control Engineering (InterNoise 2016)*. Hamburg, Germany: German Acoustical Society (DEGA), pp. 3164–3173.

Orga, Ferran, Alías, Francesc and Alsina-Pagès, Rosa (Dec. 2017). 'On the Impact of Anomalous Noise Events on Road Traffic Noise Mapping in Urban and Suburban Environments'. In: *International Journal of Environmental Research and Public Health* vol. 15, p. 13.

Ottoz, Elisabetta, Rizzi, Lorenzo and Nastasi, Francesco (Apr. 2018). 'Recreational noise: Impact and costs for annoyed residents in Milan and Turin'. In: *Applied Acoustics* vol. 133, pp. 173–181.

Ritchie, Hannah and Roser, Max (2020). 'Urbanization'. In: *Our World in Data*. https://ourworldindata.org/urbanization.

Salamon, J. and Bello, J. P. (Mar. 2017). 'Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification'. In: *IEEE Signal Processing Letters* vol. 24, no. 3, pp. 279–283.

Salamon, J., Jacoby, C. and Bello., J. P. (Nov. 2014). 'A dataset and taxonomy for urban sound research'. In: *Proc. of 22nd ACM International Conference on Multimedia*. Orlando, Florida, USA: ACM, pp. 1041–1044.

Sevillano, Xavier et al. (May 2016). 'DYNAMAP – Development of low cost sensors networks for real time noise mapping'. In: *Noise Mapping* vol. 3 (1), pp. 172–189.

Socoró, Joan Claudi, Alías, Francesc and Alsina-Pagès, Rosa Ma (2017). 'An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments'. In: *Sensors* vol. 17, no. 10, p. 2323.

St Pierre Jr, Richard L and Maguire, Daniel J (2004). 'The impact of A-weighting sound pressure level measurements during the evaluation of noise exposure'. In: *Noise-Con 04. The 2004 National Conference on Noise Control EngineeringInstitute of Noise Control EngineeringTransportation Research Board*.

Turisme a Barcelona - ciutat i regió, Observatori del (2019). *Informe de l'Actividad Turistica 2019 - Capsula 1*.

Valero, Xavier and Alías, Francesc (2012). 'Hierarchical classification of environmental noise sources considering the acoustic signature of vehicle pass-bys'. In: *Archives of Acoustics* vol. 37, pp. 423–434.

W, Clark W (1991). 'Noise exposure from leisure activities: a review'. In: *Journal Acoustic Soc Am*, pp. 175–181.

World Health Organization, 2018 (n.d.). *Environmental Noise Guidelines for the European Region*. URL: https://www.euro.who.int/__data/assets/pdf_file/0008/383921/noise-guidelines-eng.pdf. (accessed on 31 July 2020).

Wunderli, Jean Marc, Pieren, Reto, Habermacher, Manuel, Vienneau, Danielle, Cajochen, Christian, Probst-Hensch, Nicole, Röösli, Martin and Brink, Mark (2016). 'Intermittency ratio: A metric reflecting short-term temporal variations of transportation noise exposure'. In: *and environmental epidemiology* vol. 26, no. 6, pp. 575–585.

Zambon, Giovanni, Benocci, Roberto, Bisceglie, Alessandro, Roman, H. Eduardo and Bellucci, Patrizia (2017). 'The LIFE DYNAMAP project: Towards a procedure for dynamic noise mapping in urban areas'. In: *Applied Acoustics* vol. 124, pp. 52–60.

Authors' addresses

Ester Vidaña-Vila GTM – Grup de Recerca en Tecnologies Mèdia, La Salle Campus Barcelona - Universitat Ramon Llull Quatre Camins, 30, 08022 Barcelona, Spain

ester.vidana@salle.url.edu

Paper II

Low-Cost Distributed Acoustic Sensor Network for Real-Time Urban Sound Monitoring

Ester Vidaña-Vila, Joan Navarro, Cristina Borda-Fortuny, Dan Stowell, Rosa Ma Alsina-Pagès

Published in *Electronics*, December 2020, volume 9, issue 12, pp. 2119. DOI: 10.3390/electronics9122119.

Abstract

Continuous exposure to urban noise has been found to be one of the major threats to citizens' health. In this regard, several organizations are devoting huge efforts to designing new in-field systems to identify the acoustic sources of these threats to protect those citizens at risk. Typically, these prototype systems are composed of expensive components that limit their large-scale deployment and thus reduce the scope of their measurements. This paper aims to present a highly scalable low-cost distributed infrastructure that features a ubiquitous acoustic sensor network to monitor urban sounds. It takes advantage of (1) low-cost microphones deployed in a redundant topology to improve their individual performance when identifying the sound source, (2) a deep-learning algorithm for sound recognition, (3) a distributed data-processing middleware to reach consensus on the sound identification, and (4) a custom planar antenna with an almost isotropic radiation pattern for the proper node communication. This enables practitioners to acoustically populate urban spaces and provide a reliable view of noises occurring in real time. The city of Barcelona (Spain) and the UrbanSound8K dataset have been selected to analytically validate the proposed approach. Results obtained in laboratory tests endorse the feasibility of this proposal.

II.1 Introduction

Research dating back to the last century (Alexander 1968) has acknowledged that continuous exposure to high levels of noise is harmful for human beings, as recently highlighted by the World Health Organization (WHO) (*WHO/Europe | Noise - Data and statistics* n.d.). For instance, noise can negatively affect sleep quality (Test et al. 2011), induce chronic effects on the nervous sympathetic system (Su-bei 2007), or even cause psycho-physiological effects

such as annoyance, reduced performance or aggressive behavior (Moudon 2009). In this context, noise is often defined as a type of unwanted and/or harmful sound that disturbs communication between individuals (Moudon 2009; Juan P Bello et al. 2019), i.e., the overall acoustic energy measured in Sound Pressure Levels (SPLs) exceeds a predefined limit (Flindell and Walker 2004).

Accordingly, several agencies and public departments (e.g., NSW Environment Protection Agency, NYC Department of Environmental Protection, European Commission) have defined regulations (Flindell and Walker 2004) to limit the amount of noise (i.e., equivalent averaged level *LAeq*) that the population can be exposed to. For instance, the WHO recommends that noise must be below 35 dBA in classrooms to enable good teaching and learning conditions, or below 30 dBA in bedrooms to enable good quality sleep (Hurtley 2009). Most of these regulations define the maximum level of noise allowed in a specific scenario (e.g., home buildings, factories, schools) and a specific acoustic source (e.g., motor vehicles, air conditioners, machinery, water heaters, etc.) (Office 2017). However, such a standard way of defining and regulating noise faces two important challenges when applied and enforced in the real world (Flindell and Walker 2004): acoustic source isolation and identification and practical on-field noise measurement for automatic acoustic surveillance:

1. It is very difficult to isolate and identify a specific noise source from the overall acoustic landscape since the aforementioned SPL measurements aggregate the energy level from all the acoustic sources at the same time (Juan P Bello et al. 2019). Indeed, in a real-world environment, several different acoustic sources may emerge over time and, thus, the definition of a fixed SPL threshold for a given area is not always appropriate (Mun and Geem 2009), i.e., the acoustic threshold should be dynamic according to the sound (noise) that is currently occurring. Unfortunately, the SPL value *per se* does not provide enough practical information to facilitate the identification of the sound (noise) source(s) (Mun and Geem 2009), which complicates the task of verifying whether an acoustic landscape meets the local regulations or not.

2. Also, effectively measuring the amount of noise in large-scale environments (e.g., urban areas) requires a considerable number of resources in terms of highly qualified professionals—it has been reported that the NYC Department of Environmental Protection has up to 50 professionals designated to dealing with noise complaints in the city of New York (despite this, their average response time is still about 5 days) (Juan P Bello et al. 2019)—and expensive equipment (Juan P Bello et al. 2019). Indeed, this equipment can range from $1500 up to $20,000 depending on the type, measurement range, and capability of the microphone to produce noise spectral data (Mydlarz et al. 2017). Therefore, conducting scalable, long-term (i.e., conducting measurements 24 h a day 365 days a year) noise surveillance tasks in wide span areas has emerged as a hot research topic in recent years.

Over the last decade, Ubiquitous Sensor Networks (USNs) (13 2008) have emerged as a powerful alternative to address the challenges of scalable and cost-effective (Ferrández-

Pastor et al. 2016) sensing in large-scale areas (Murty et al. 2008). The benefits of USNs have been exploited in several domains, ranging from water pollution monitoring (Shin et al. 2007) to smart agriculture (Ferrández-Pastor et al. 2016), including Wireless Acoustic Sensor Networks (WASNs) for Ambient Assisted Living (Navarro et al. 2018). Indeed, USNs provide a design reference to conceive versatile architectures able to interconnect a high number of devices—typically with limited capabilities in terms of storage, computation and communications—while providing fault tolerance and robustness with the aim of increasing the performance of individual sensors (Shin et al. 2007; Bagula et al. 2012; Koucheryavy et al. 2015). Indeed, the idea of using an interconnected set of inexpensive commodity hardware to beat the performance of individual high-end devices is well known in the literature of distributed systems and has been massively exploited— the Google File System (Ghemawat et al. 2003) being one of its most representative examples.

This work aims to extrapolate this idea to the field of urban sound monitoring, i.e., the use of a set of low-cost microphones deployed in a redundant topology—being the sensing layer (13 2008) of an ubiquitous sensor network that will later provide them with additional storage and computing features—to *listen* to events from large-scale areas in a cost-effective way while obtaining a reasonable accuracy. Hence, the modest performance of the low-cost microphones can be compensated by the robustness of the computing algorithms running on top of the ubiquitous sensor network (Piper et al. 2017). Therefore, the purpose of this paper is to propose a low-cost cost distributed acoustic sensor network for real-time urban sound monitoring in large-scale scenarios. More specifically, the proposed approach aims to present a network composed of inexpensive hardware (i.e., Raspberry Pi Model 2B (RPi) (*Raspberry Pi Official web site* n.d.)) in which each node is conceived to (1) process a real-time audio stream from a directly connected low-cost microphone, (2) locally identify the occurring events in this audio stream by means of a deep neural network, (3) communicate the identified events to the neighboring nodes of the network by means of a custom planar antenna with almost isotropic radiation, and (4) globally validate these locally discovered events by means of a distributed consensus protocol.

To sum up, the main contributions of this work are:

- A deep-learning algorithm for urban sound identification in real time to be deployed in low-cost devices with modest computing and storage capabilities.
- A custom planar antenna with almost isotropic radiation pattern for robust and low-energy consumption communications between nodes.
- A distributed consensus protocol to compare the detection results of each individual node with its neighboring nodes.

To further validate the proposed approach, we evaluated automatic recognition against the UrbanSound8K dataset (J. Salamon et al. 2014) as a source of typical urban audio events and selected the city of Barcelona (Spain) as a reference model to deploy the proposed system. Indeed, Barcelona was designed following a particular square block grid (see Figure II.1) that

makes it an ideal scenario to deploy urban ubiquitous sensor networks. However, current noise surveillance initiatives in Barcelona only focus on sound pressure levels and span an average area of 1 square kilometer per sensor. This work aims to enrich the measurements by identifying the sound source and providing a fine-grained analysis of their location. The evaluation of the proposed system has been done as follows: (1) the communication antenna has been validated by means of simulation, and (2) the acoustic recognition together with the distributed consensus protocol have been validated by means of laboratory testing rather than full real-world deployment, planned for future work. The regularly defined urban grid of Barcelona greatly facilitates this aspect of spatial modelling.

Figure II.1: Aerial view of the urban grid structure of the city of Barcelona (Wikipedia contributors 2020).

The remainder of this paper is organized as follows. Section II.2 reviews the related work on acoustic sensor networks for environmental noise monitoring. Section II.3 details the proposed system architecture and details its three layers: data processing, distributed consensus, and communications. Section II.4 presents the conducted experimental evaluation. Section II.5 discusses the obtained results. Finally, Section II.6 concludes the paper and proposes some future work directions.

II.2 Related Work

In this section, we describe related works to the main WASN-based approaches developed in recent years to monitor environmental noise. The main goal of these networks is to collect the L_{Aeq} levels alone or together with extra information obtained in each node. In some situations, this extra information gathered in each node corresponds to data about the sound source measured in each sensor.

II.2.1 WASNs to Monitor the Noise Levels

Most of the WASNs in this first category use commercial sound level meters as sensor nodes. These devices are usually connected to a central server of the WASN, which collects all the L_{Aeq} information collected by the nodes. Projects such as Telos (Polastre et al. 2005), which correspond to one of the first experiences in this WASN design by means of an ultra-low power wireless sensor module designed by the University of California (Berkeley). Some years after that experience, Santini et al. in (Santini and Vitaletti 2007; Santini et al. 2008) showed how a WASN can be used in a wide variety of environmental monitoring applications, with a special focus on urban noise.

More recent projects include the deployment of a network to monitor the traffic noise in Xiamen City (China) for environmental purposes (Wang et al. 2013). The project covers 35 roads in 9 green spaces in the city, and the scientists use the data from the monitoring stations to model the traffic of 100 other roads in the city. The deployment included noise level meters, with ZigBee and GPRS communications.

The FI-Sonic Project is focused on continuous noise monitoring surveillance (Paulo et al. 2015); the main goal is to develop the technology required to process urban sounds by means of artificial intelligence, enabling the generation of noise maps but also the identification and location of groups of sound events (Paulo et al. 2016). It is based on a FIWARE platform (https://www.fiware.org/). Finally, the RUMEUR project (Urban Network of Measurement of the sound Environment of Regional Use) is based on a hybrid wireless sensor network deployed by BruitParif (F. Mietlicki et al. 2015) in Paris and its surroundings. The network has high accuracy on monitoring critical places (for example, airports) but also uses other less precise measuring equipment, whose final goal is to evaluate the equivalent noise level of the environment. The RUMEUR project has evolved to Medusa (C. Mietlicki and F. Mietlicki 2018), a system that combines four microphones and two optical systems so that noise levels can be represented on a 360° image of the environment, by means of the identification of the source location. Its computational load is high, and it cannot be resolved by most of the low-cost acoustic sensor systems.

The Barcelona Noise Monitoring Network (NMN) was described in (Camps-Farrés 2015) and reviewed in (Camps-Farrés and Casado-Novas 2018). The network is designed to reduce the impact of urban infrastructures on the environment in the city of Barcelona. The results of the analysis carried out in (Camps-Farrés and Casado-Novas 2018) suggest that both the costs and the number of manual tasks carried out by technicians should be reduced. In Barcelona, several other initiatives to empower the citizens of critical urban areas, such as Plaza del Sol (Coulson et al. 2018), have also been developed, but so far they have been only able to complement the measurements conducted by the calibrated sensors deployed by the City Council.

II.2.2 WASNs Based on Ad-Hoc Designed Nodes

To satisfy the increasing demand of an automatic monitoring of noise levels in urban areas, as described in (Basten and Wessels 2014), several WASN-based projects are being developed in

different countries and designed, then deployed ad-hoc for their application; some of these projects include other environmental measurements used to determine other aspects of citizens' quality of life besides noise pollution.

The CENSE project (Characterization of urban sound environments) focuses on the design of noise maps in France (*Cense - Characterization of urban sound environments* n.d.). It integrates both simulated and measured data by means of a wide network of low-cost sensors. The project includes environmental acoustics, statistics, Graphical Information System (GIS) to plot the results, as well as network sensor design, signal processing and the proposal of the production of perceptive noise maps. The IDEA project (Intelligent Distributed Environmental Assessment) (Botteldooren et al. 2011) focuses on noise and air quality pollution in several urban areas of Belgium. It integrates a sensor network based on a cloud platform, and it measures noise and air quality (Domínguez et al. 2014). The MESSAGE project (Mobile Environmental Sensing System Across Grid Environments) (Bell and Galatioto 2013) not only monitors noise, carbon monoxide, nitrogen dioxide, temperature, but also humidity and traffic occupancy, and it gives real-time noise data information in the United Kingdom. The UrbanSense project (Rainham 2016) and the MONZA project (Bartalucci et al. 2018) follow both the idea of monitoring urban noise real-time together with other air pollutants; UrbanSense in Canada and MONZA in the Italian city of Monza.

The urban acoustic environment of New York City is monitored using a low-cost static acoustic sensor network in the framework of a project named SONYC project (Sounds of New York City) (Mydlarz et al. 2017). The goal of this project is to describe the acoustic environment while monitoring noise pollution. It collects longitudinal urban acoustic data, to process them and have generous sampling to work with acoustic event detection (Juan P Bello et al. 2019).

Another interesting approach of the monitoring network projects is the hybrid approach of crossing the acoustic information with subjective perception surveys, to consider the typology of the events in relationship with sleep quality (De Coensel and Botteldooren 2014). A sound recognition system is applied to provide information about the detected sounds and establish a relationship between the perception surveys and the identified events related to road traffic noise (Brown and Coensel 2018). However, this project is only aimed at the identification of the acoustic events and their perception, it has no impact on any noise maps generation.

The DYNAMAP project (Sevillano et al. 2016) achieves a good trade-off between cost and accuracy in the design of a WASN. The project deployed two pilot areas in Italy, located in Rome (Bellucci et al. 2017) and Milan (Zambon et al. 2017), so as to evaluate the noise impact of road infrastructures in suburban and urban areas, respectively. The two WASNs monitor road traffic noise reliably collecting data at 44.1kHz to remove specific audio events, which are unrelated to road traffic (Socoró et al. 2017; Rosa Ma Alsina-Pagès et al. 2018) for the noise map computation (Bellucci and Cruciani 2016). Based on their experience in this project and particularly the time and effort spent transforming the original prototyping code into real operable language, the team developed a low-cost flexible acoustic sensor for rapid real-time algorithm development and testing (Rosa Maria Alsina-Pagès et al. 2020).

II.3 System Architecture

This section details the proposed system architecture and further elaborates on the rationales to implement the ubiquitous acoustic sensor network for urban sound identification. The main constraints (Murty et al. 2008) and design guidelines that have driven the conception of this distributed system are the following:

Cost affordability. The system must be conceived to cover large-scale areas (i.e., hundreds of km^2) in a redundant topology (Piper et al. 2017) (i.e., at least 4 nodes per city block (As a matter of reference, in Barcelona city (Spain) the sides of the blocks measure around 110 m on average). Therefore, the individual cost of each of the nodes that articulate the ubiquitous sensor network must be kept as low as possible. This prevents us from using expensive high performance computing devices (e.g., GPUs (Navarro et al. 2018)) and leads us to consider alternative solutions with more modest computing and storage features.

Physical distance between neighboring nodes. As individual nodes must be composed of inexpensive hardware—in terms of both acoustics and computing— they need to take advantage of each other to provide robust answers and good performance (Piper et al. 2017). In this regard, each node will need to constantly communicate with its neighbors to check, compare, and validate the identified acoustic events. Therefore, there is a trade-off on the physical distance between nodes: one the one hand it must be kept low so that an event can be *heard* by more than one node, and on the other hand, the larger distance, the more area will be covered.

Real-world deployment. The system must be deployed in urban spaces, which makes it vulnerable to extreme weather conditions (e.g., heat, cold, wind, rain), vandalism, or theft (Murty et al. 2008). Therefore, the nodes that compose the proposed USN must be as small as possible so they can be installed in existing street furniture (e.g., traffic lights (Ji et al. 2020b)). Also, the power consumption of each node must be low to facilitate its integration. This means that the proposed approach will need to be efficient both in terms of communications (i.e., exchanging little data among nodes) and of computing (i.e., using as low computing resources as possible to obtain the maximum event identification accuracy).

Fault tolerance and recovery. Since the nodes of the system will be exposed to harsh environmental conditions and given the difficulty of physically accessing them to conduct maintenance and reparation duties (e.g., reboot), the nodes must be self-managed, i.e., a node must keep operating even in case of failure of their neighboring nodes.

Acoustic quality. The nodes must be capable of acquiring and processing data at a minimum sample rate of 22,050 samples per second (to be able to analyze frequency information ranging from 0 to 11,025 Hz) and a depth of 16 bits per sample. Before deployment, the

microphones must be calibrated, and the gain must be adjusted so all the microphones of the USN capture similar signal levels when exposed to the same sounds.

To meet all these requirements, we propose the use of a Raspberry Pi device augmented with acoustic and communications capability. We select the Raspberry Pi Model 2B (*Raspberry Pi Official web site* n.d.) that has a 900 MHz quad-core ARM Cortex-A7 processor, 1 GiB RAM, and an average power consumption of 200 mA. An external USB microphone for acoustic data processing and a custom communications antenna for data exchanging among nodes must be plugged to the RPi (see Figure II.2). The remainder of this section (1) describes the proposed acoustic data-processing framework to locally identify acoustic events in urban areas, (2) details and justifies the design of a communications module (modem and custom antenna) to enable data communications among nodes, and (3) presents a distributed consensus protocol aimed at comparing the locally identified acoustic events to obtain robust and global-scope acoustic findings.

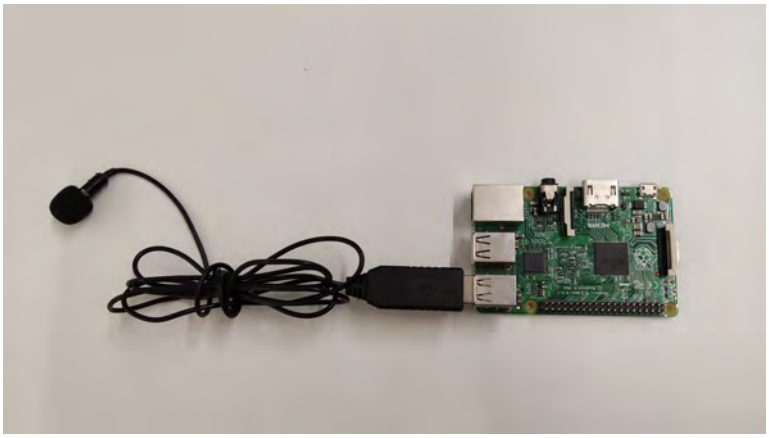


Figure II.2: Raspberry Pi Model 2B with USB microphone.

II.3.1 Data Processing

For data acquisition and processing, a low-cost omnidirectional electret USB microphone is used. The reference number for the microphone is OUT-AMLO-0872 and it is manufactured by Seacue. The frequency response is almost flat for the frequency range where the events are taking place (50 Hz–10 KHz), meaning that it does not generate *colorations* (i.e., alterations or distortions) on that frequencies. The price is as low as 12 EUR and it is plug-and-play, meaning that there is no need for an external Analog-to-Digital converter (ADC). Once one window (audio fragment of a certain duration) of audio is captured, the spectrogram is calculated and fed to the neural network. Using spectrograms as input features for the network is a technique that has been proven to be effective for sound classification tasks (Huzaifah 2017), since they provide information about acoustic energy in both frequency and time. The selected network architecture is a Convolutional Neural Network (CNN). The reason behind using a CNN lies in the fact that they typically require storing fewer parameters than traditional deep neural networks, which reduces the model size (Goodfellow et al. 2016).

Moreover, CNNs have been extensively validated for sound event detection (Mesaros et al. 2017).

The output result of this data-processing layer is an events vector with as many components as acoustic event types (i.e., classes), where each component is a value between 0 and 1 representing the probability of the event belonging to that class. For instance, in the case of UrbanSound8K dataset being used in this work, there are 10 event types.

This resulting vector will be sent to the neighboring nodes using the communications antenna and the distributed consensus protocol that is detailed in the following sections.

II.3.2 Communications

For inter-communication between neighboring nodes, a custom bespoke antenna has been designed to achieve higher specifications with the limited physical space available on the RPi. The performance of the whole communications system is calculated using the Friis Transmission Equation (Pozar 2011). This equation states that the signal received by the communications module (i.e., antenna plus transceiver) is calculated and compared to the noise level, giving a Signal-to-Noise Ratio (SNR). If the SNR is high enough, the signal will be successfully decoded. The Friis Transmission Equation shows that losses increase with frequency, and higher power is lost at higher frequencies. Therefore, the very first design constraint to be addressed is the operating frequency.

The 2.4 GHz band (UN-51) (Ministerio de Energía n.d.; CNAF n.d.[c]; CNAF n.d.[b]) (i.e., Wi-Fi) is a convenient choice for communications in USNs (Murty et al. 2008). However, this band is often absorbed by structural elements, such as walls and floors or ceilings and it also coincides with the resonant frequency of water, which makes it inappropriate for urban spaces. In addition, such a high frequency limits the communication range of each node (Pozar 2011). Alternatively, the 433 MHz band (UN-30) is slightly better than the 2.4 GHz band for the range—as it operates at a lower frequency—but it does not guarantee a secure data transmission, due to a lot of interference in this specific frequency band, such as remote controls and parking remote controls which can produce high levels of interference. The frequency band of 868 MHz is an ISM band designated by the UN-39 in Spain (Ministerio de Energía n.d.; CNAF n.d.[c]; CNAF n.d.[b]) that offers a better range than the 2.4 GHz band, increased by 2–3 times, and is less populated than the 433 MHz band. This frequency band is used by LoRa, Zigbee and Sigfox in ITU region 1 (Europe) (CNAF n.d.[a]). In case of the system being used in other regions, such as the US, the ITU-RR-5.150 specifies a band in 915 MHz in the ITU region 2. In this case, the communication system would need to be accordingly updated and fine-tuned with the new requirements to radiate at the specific frequency band for the new region.

To summarize, transmission in the 868 MHz band is (1) able to penetrate obstructions in the line-of-sight and (2) suitable for connecting medium and long-distance remote monitoring systems. However, it presents limited maximum data rates compared to other bands. As in the proposed large-scale urban sound monitoring use-case, low data rates (a few kbps) for medium range are enough to transmit the vector with the classification results (see Section

II.3.1), there is no need to use a higher frequency band.

After selecting the 868 MHz operating band, a transceiver for this frequency to be attached to the RPi is required. There are many off-the-shelf communication modules available in the market for RPi—which is in fact one of the advantages of using this device. For a very low price there are numerous modules to radiate at the frequency band of 868 MHz, some with a short range and others with medium range. For example, the ENOCEAN PI 868, RTX-868-FSK and SX1272 (González et al. 2016; Links 2016). The SX1272 module for RPi operates at 868 MHz but uses a simple monopole antenna. The monopole antenna is troublesome as it could lead to null communication in certain directions. An isotropic antenna would best fit the requirements for this project. The design of a custom bespoke antenna to be connected to the SX1272 module is presented below.

As shown in Figure II.3, the antenna is designed with two planar crossed dipoles in a low-cost FR-4 substrate to present an isotropic pattern. Consequently, the same signal will be received at the receiver due to its isotropic properties, regardless of the orientation of the antenna. Therefore, there is no risk that communication will be lost when the sensor is in certain orientations. The selected design is based on an isotropic Crossed Dipoles designed for a higher frequency band (Pan et al. 2012). The proposed design must be optimized to operate at the 868 MHz frequency band and to fit the RPi case. The planar crossed dipoles placed on top of a low-cost FR-4 substrate with 1 mm thickness and relative permittivity of 4.4, are a low-cost solution to fit in the RPi Shield and provide isotropic radiation. The crossed dipoles are fed at a 90° phase shift to achieve isotropic radiation (Pan et al. 2012).

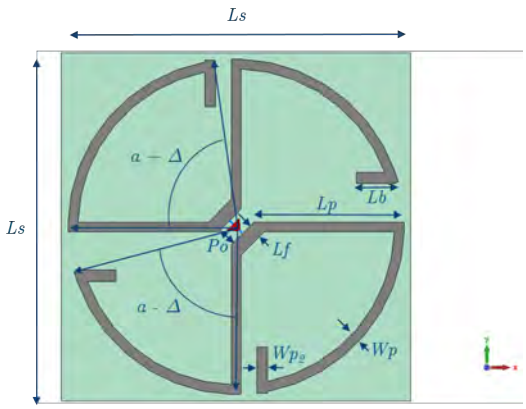


Figure II.3: Proposed Planar Crossed Dipoles for isotropic radiation.

To optimize the antenna parameters and achieve the aforementioned requirements of the communication system (operating frequency, isotropic radiation and size constraints), CST Microwave Studio has been used to conduct the parametric studies depicted in Figure II.4:

1. Delta (Δ) is the variation between the length of the horizontal dipole and the length of the vertical dipole. By varying the parameter Δ, the current of the two crossed dipoles can be excited at the same magnitude and 90 degrees phase shift, which is only in this conditions that isotropic radiation can be achieved (Pan et al. 2012). The top left plot

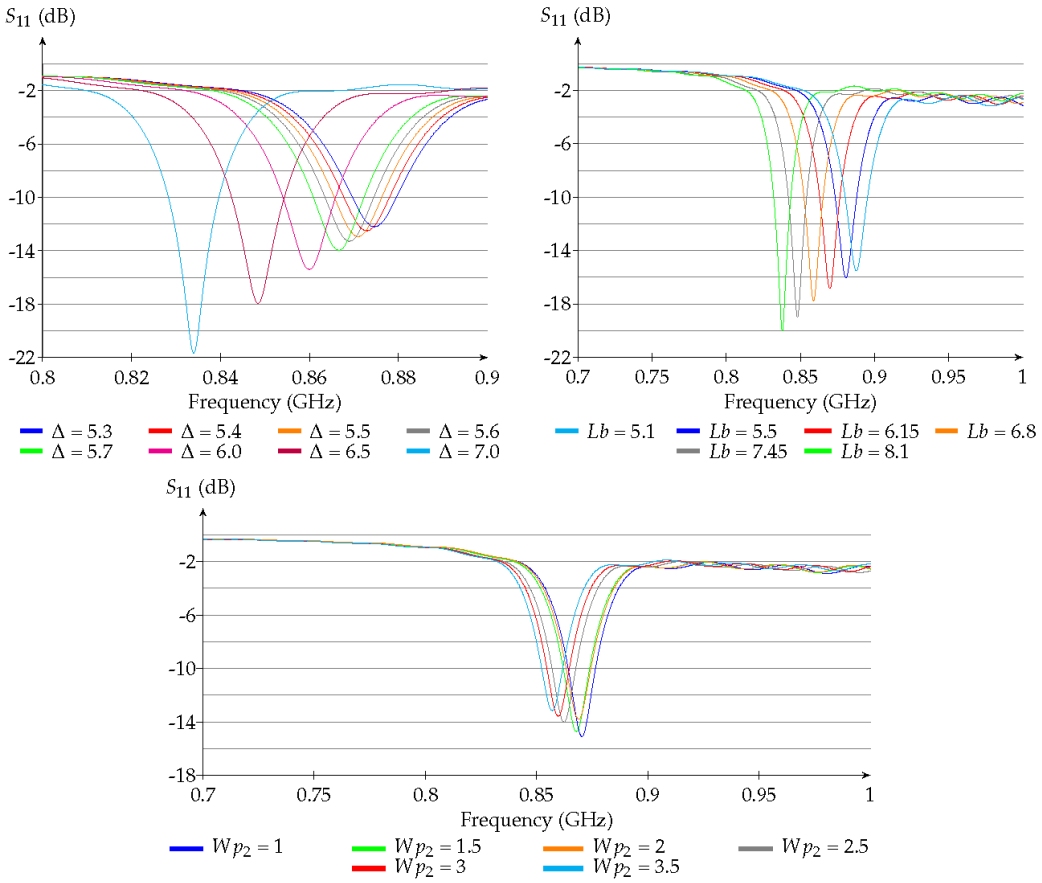


Figure II.4: Parametric study of the reflection coefficient S_{11}. Top left: changing the Δ. Top right: changing the branch length (Lb) . Bottom: changing the branch width (Wp_2).

in Figure II.4 shows that when the 90 degrees phase shift between dipoles is achieved the operating frequency is better matched and, thus, S_{11} parameter becomes lower.

A Δ of 7.0° is selected because it better matches the input impedance of the antenna and it presents isotropic radiation.

Once isotropic radiation is achieved, the input impedance can be matched at the operating frequency by varying the dipole length and width as shown in the next steps.

2. The length of each dipole branch is a crucial parameter to match the antenna at 868 MHz frequency band. The branch length (Lb) is the length at the end which will be longer for lower frequencies or shorter if the antenna needs to operate at higher frequencies. Lb is added to the design to match the input impedance of the antenna at the desired operating frequency of 868 MHz, considering the size of the RPi cannot accommodate long-enough dipoles to be resonant at this frequency. Otherwise, antenna efficiency would be reduced at the required operating frequency band. By making the dipole branches longer a lower frequency can be matched. As expected, the resonant frequency decreases as the length of the dipole increases. Therefore, the desired operating frequency can be achieved by varying this parameter.

   Lb is the same for both dipoles, as the only difference in length comes from the parameter

Δ, mentioned before. Δ is used to achieve isotropic radiation, Lb adjusts the matched input impedance so that the antenna operates at the required operating band.

The top right plot in Figure II.4 shows the effect of varying the Lb parameter in the reflection coefficient of the antenna (S_{11} parameter).

3. Finally, the length of the 2 dipole branches is determined, although the width will also impact the operating frequency of the antenna, as shown in Figure II.4 (bottom). A parametric study of the $Wp2$ is used to fine-tune the value of this parameter and match the antenna at exactly 868 MHz with a value of 1.5 mm.

Combining the results of these studies, the optimized parameters of the proposed antenna are presented in Table II.1:

Table II.1: Values of the Optimized design parameters for the antenna geometry.

Ls (mm)	Lp (mm)	Lb (mm)	Lf (mm)	Po (mm)	Wp (mm)	Wp_2 (mm)	α	Δ
60	25.8	5.1	3.18	3	1.7	1.5	78.6°	7°

With this configuration, a matching of -22 dB can be achieved at the 868 MHz band. As a result, Figure II.5 presents the simulated radiation pattern obtained from exciting each dipole at a time. It can be observed that the combination of the planar crossed dipoles is essentially isotropic.

As shown in Figure II.6, the antenna is matched at the 868 MHz frequency band and has a good rejection rate of other bands. Also, it presents the higher gain at the frequency that is matched (868 MHz). Outside this band the gain severely declines. The gain obtained at the frequency of interest is 1.41 dBi which is close to the isotropic gain radiation expected for the proposed antenna. Recall that the antenna gain needs to be low to produce isotropic radiation, so that the physical orientation of the node will not affect the communication link in deployment.

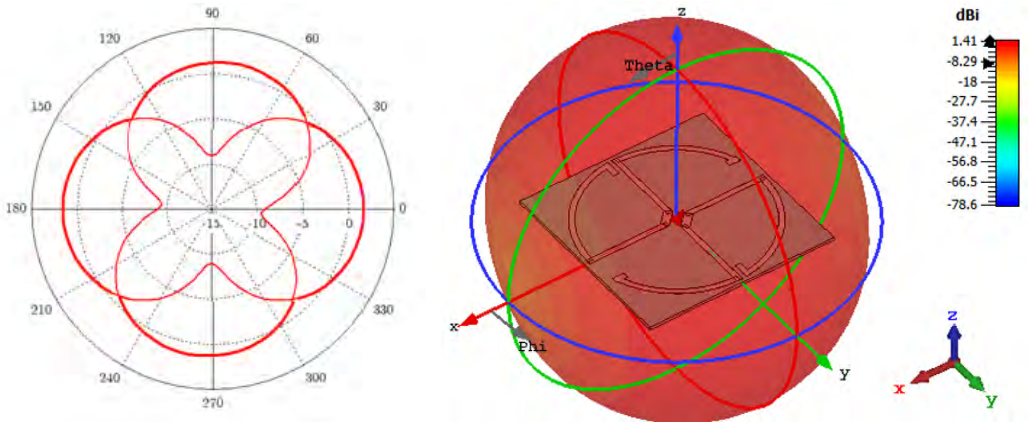


Figure II.5: Radiation patterns of the Planar Crossed Dipoles for isotropic radiation in 3D (right) and the combination for Theta=0 exciting each dipole at a time (left).

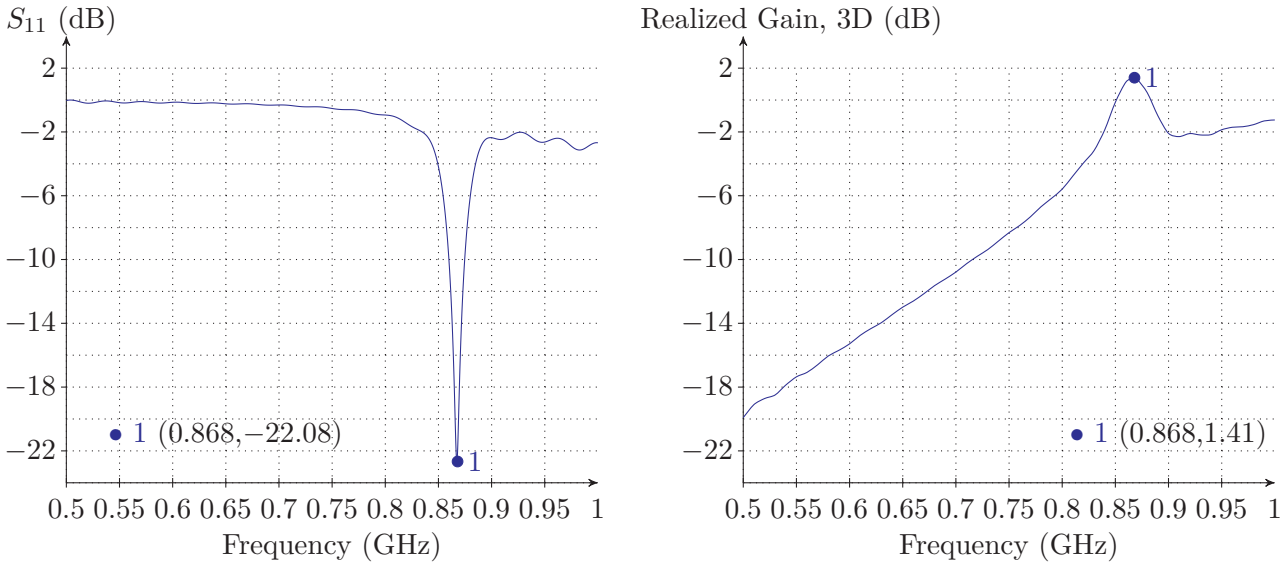


Figure II.6: Performance results of the proposed Planar Crossed Dipoles. Reflection coefficient for isotropic radiation on the left. Realized Gain over frequency on the right.

II.3.3 Distributed Consensus

Designers of USNs typically select cloud or edge computing architectures (Armbrust et al. 2010) to outsource the heavy computation tasks associated with data streams processing (Navarro et al. 2018). This alleviates the requirements in terms of storage and computing of USN nodes but requires a reliable communications infrastructure able to transfer a large amount of data traffic to (and from) the cloud. However, in the specific scenario of large-scale urban sound monitoring, streaming the sensed acoustic data to a central entity (or cloud) would increase the complexity (in terms of codecs and connectivity to the Internet), the delay, the power consumption (Ji et al. 2020a) and the overall cost of the nodes (Pham and Cousin 2013). Therefore, the proposed USN has been designed to be autonomous (i.e., it can reliably identify acoustic events without the aid of powerful cloud devices) and self-managed. In this regard, a custom distributed consensus layer that enables synchronous communications among nodes has been implemented.

This layer is committed to increasing the robustness of the local acoustic event identification by comparing the identified local events at a single node with the events detected by neighboring nodes with the aim to emulate an ensemble decision system (Nanni et al. 2020). For instance, if a node detects a car horn but none of the surrounding nodes have detected this event, the system may decide to discard such event.

As shown in Figure II.7, nodes are organized following a token ring topology. To keep the size of the ring small—recall that the purpose of the proposed USN is to take advantage of physical redundancy to enable more than one node listen the same event—and minimize the delay, all the nodes that keep a close physical distance are assigned to the same ring. Therefore, to cover a large-scale physical area the same node can belong to more than one ring, which results in a multiring topology (Aiello et al. 2001).

The behavior of each node from the ring is as follows:

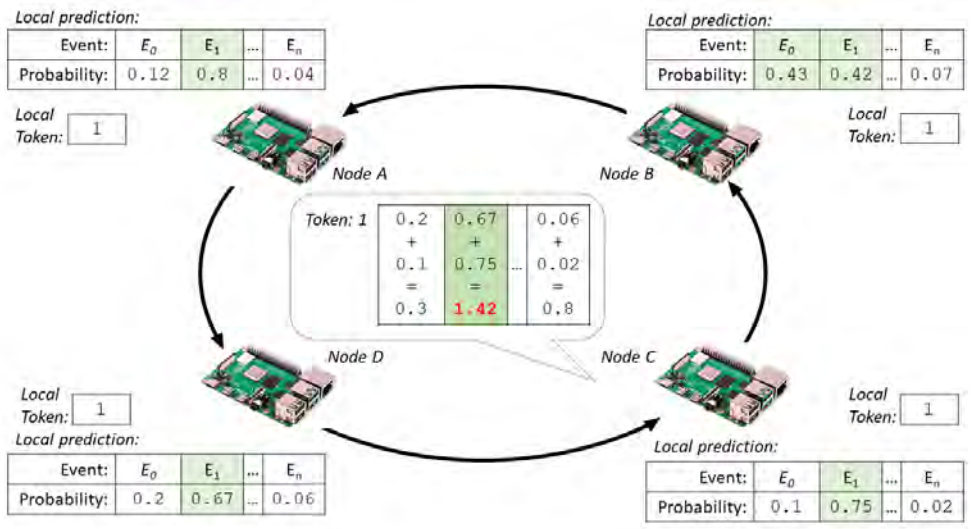


Figure II.7: Logical organization of nodes.

1. Run the local data processing (i.e., CNN classification algorithm) and wait for the classifier to populate the events vector (see Section II.3.1).

2. Next, increment a local token number, which will give a notion of virtual synchrony among the nodes, i.e., a natural number indicating the logical time in which the classification has been conducted.

3. If the node has the lowest identifier among all the nodes of the ring, it sends a message to the next neighbor in the ring containing the local token number and the obtained events vector. The remaining nodes will wait to receive their associated message.

4. When a node receives the vector with a token number matching its local token number for the first time, it will make a component-wise addition between its own events vector and the vector contained in the message. Next, it will forward the resulting vector and the token number to the subsequent node.

5. When a node receives the vector with a token number matching its local token number for the second time, it means that all the nodes of the ring have contributed to the events vector contained in the message. At that moment, the node will apply a set of heuristic rules to determine the final label of the event. Specifically:

   - If the node locally classifies an event whose *Leq* is typically low (i.e., *air conditioner, children playing, dog bark and engine idling*) with a probability of more than 90%, the results obtained from the rest of the nodes (i.e., events vector from the message) of the ring will be ignored. In this case, the local events vector will be examined and the component with the highest value will be considered the winning label. The rationale behind this decision is that it is not likely than in a noisy street these sounds can be heard by different sensors of the ring, as background noise will probably mask them.

- However, for the rest of the events (that typically have higher *Leq* such as horns or sirens), or if the network was not completely sure of whether one of the other events had actually occurred, the events vector from the message will be examined and the component with the highest value will be considered the winning label.

6. Increment the local token and go back to the first step. Please note that thanks to the local token, the system can associate the events vector with a logical time frame, which would be very useful in the case of faults (e.g., node crash, or communication fading).

Figure II.7 shows an example of the proposed approach with a ring of 4 nodes. It can be seen that the most probable event detected at *Nodes* 0, 1, and 3 is E_1; however, *Node* 2 believes that the most probable event is E_0. However, after sharing the events vector with all the nodes of the ring, it will correct the local classification and agree with its neighboring nodes that the most probable event is E_1.

II.4 Experimental Evaluation

This section aims to validate the feasibility of the proposed approach by means of two experiments. In the first experiment, authors evaluate several deep network architectures. Using the original UrbanSound8K dataset (J. Salamon et al. 2014), audio files are tested in a RPi to find out which classification algorithm offers the best trade-off between classification accuracy and memory/computing requirements. The aim of this experiment is to find an algorithm capable of classifying acoustic data in real-time with the resources provided by a single low-cost device. The results obtained in this first experiment will give a best-case scenario accuracy values that will be used as a baseline to compare the results of the second experiment.

The second experiment aims to evaluate how different neighboring nodes connected as described in Section II.3 would behave in an emulated (i.e., laboratory) real-world operation. For this purpose, the audio files from the dataset are modified emulating the air channel and physical topology of the streets of Barcelona. Moreover, road traffic noise recorded in the city of Barcelona (Vidaña-Vila et al. 2020) has been randomly added to each of the audio files so each sensor perceives the event partially masked by traffic noise. This experiment shows how a ubiquitous sensor network can improve the classification results over individual sensors when perceiving the same acoustic data from different locations and masked with traffic noise.

II.4.1 Experiment 1: Event Detection in Each Individual Sensor

UrbanSound8K is an online free dataset containing 8732 labelled sound events of 4 s or less from 10 different urban categories: *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren,* and *street music*. The dataset has a total duration of 31,500 s and is preorganized in 10 different folds that must not be mixed according to (J. Salamon et al. 2014). For this work, we have used folds 1,2,3,4,6,7 and 8 as training folds, fold 5 as validation fold and fold 10 as testing fold. Figure II.8 shows a spectrogram example of each of the classes of the dataset.

Each sensor will be constantly running a deep-learning pre-trained network to be able to classify events in real time. For the experiment, we have obtained a 4-s window spectrogram of each of the audio files of the dataset. For those audio files on the UrbanSound8K dataset with a duration shorter than 4 s, we have applied the same methodology as Singh et al. in (Singh et al. 2019), which consists of replicating the same audio file until it reaches the uniform length of 4 s.

Table II.2 details the different deep networks that were evaluated to find which classifier offers the best trade-off between accuracy, the number of floating-point operations (FLOPs) (He et al. 2016; Huang et al. 2016; Zhang et al. 2017; Sandler et al. 2018; Ma et al. 2018), and the size of the model after training and storing it into disk. For this experiment, all the networks were first trained using ImageNet (Deng et al. 2009), and we then applied transfer learning to fine-tune them so they could classify the spectrograms of the selected dataset. As the expected inputs of the network are RGB images such as the ones contained in ImageNet, each 4-s gray-scale spectrogram has been normalized in the range of [0, 1], then replicated three times (one per each RGB channel), and then normalized again with the mean and standard deviation of the ImageNet dataset.

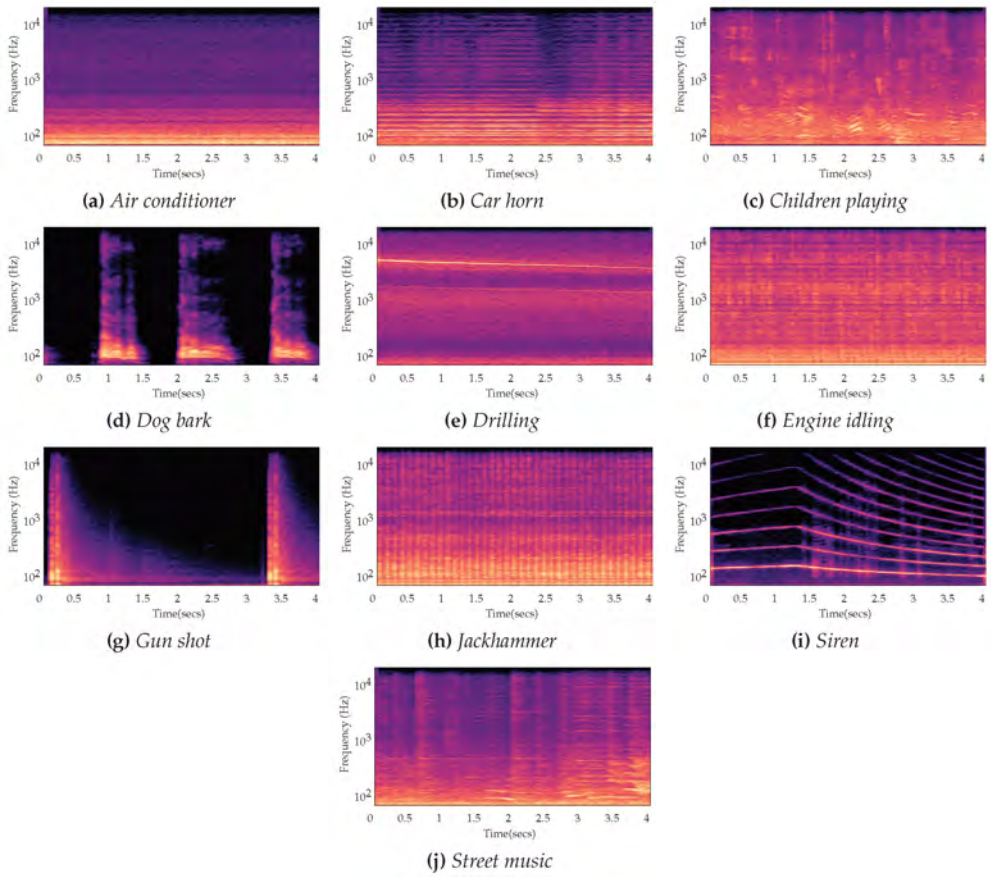


Figure II.8: Spectrograms of the ten types of sounds of the UrbanSound8K dataset.

The training of the network was carried out following standard good practice in deep learning, as follows. We used a batch size of 16 spectrograms per batch. The learning rate was initially set to 0.01, and a scheduler was programmed to decrease it by a factor of 0.1

Table II.2: Number of FLOPs, model size and accuracy on the testing fold for different network architectures.

Network Architecture	FLOPs	Accuracy	Model Size
ResNet 152	11.3×10^9	79.71%	223 MB
DenseNet 121	6×10^9	77.31%	28 MB
AlexNet	0.725×10^9	77.31%	218 MB
MobileNet v2	$\mathbf{0.3 \times 10^9}$	78.75%	8.8 MB
ShuffleNet v2	0.591×10^9	51.74%	5 MB
ResNet 18	1.8×10^9	77.19%	43 MB
VGG 16	15.3×10^9	77.91%	513 MB
SqueezeNet 18	0.833×10^9	**80.19%**	**2.9 MB**

with a patience of 3, using an SGD optimizer. Moreover, an early stopping criterion was used to obtain the optimal network configuration by using the validation fold.

As shown in Table II.2, several network architectures obtain considerable high accuracy values compared to the baseline system that Salamon and Bello presented in (Justin Salamon and Juan Pablo Bello 2015), which obtained an accuracy of 68%. Concretely, the network architecture that provides the best results is SqueezeNet 18 (Iandola et al. 2016), followed by Resnet 152 (He et al. 2016). Comparing the size of both networks in terms of numbers of operations, Squeezenet 18 is clearly a smaller network, with only 0.833 GFLOPs compared to the 11.3 GFLOPs of ResNet 152. As SqueezeNet 18 fits into the purposed architecture (i.e., RPi) and can perform the classification in less than one window time (i.e., 4 s), it has been selected—adapting the last layer with a 2D Convolutional layer with 10 outputs—to be the network installed in every node of the USN. Figure II.9 depicts a diagram of the final architecture of the classifier system, and Figure II.10 depicts the accuracy and loss when training and validating the selected model.

The acoustic processing and classification of a single 4 s window in the RPi (Quad-Core Cortex A7 at 900 MHz and 1 GiB of RAM), using the aforementioned deep network, was carried out in 2 s. This time includes (1) taking the 4 s audio data acquired by the microphone, (2) calculating the spectrogram of a 4 s audio file of the UrbanSound dataset, (3) processing the spectrograms as explained above, and (4) passing the audio file through the deep net to obtain a local classification result. Hence, we can conclude that the proposed hardware platform is able to classify in real time—considering real-time as getting a classification result in less time than one window—in a 4 s basis, which is considerably faster than existing solutions that provide minute by minute data (Bell and Galatioto 2013).

II.4.2 Experiment 2: Network of Sensors

The second experiment is aimed to assess how physical redundancy can increase the accuracy of the proposed system. To emulate the acoustic characteristics of a real-world scenario, the attributes of the Barcelona city center have been taken as a reference. According to *Pla Cerdà* (Aibar and Bijker 1997), the physical sizes of the building blocks and streets from Barcelona are depicted in Figure II.11: blocks size of 113.3 m × 113.3 m with a separation of 20 m

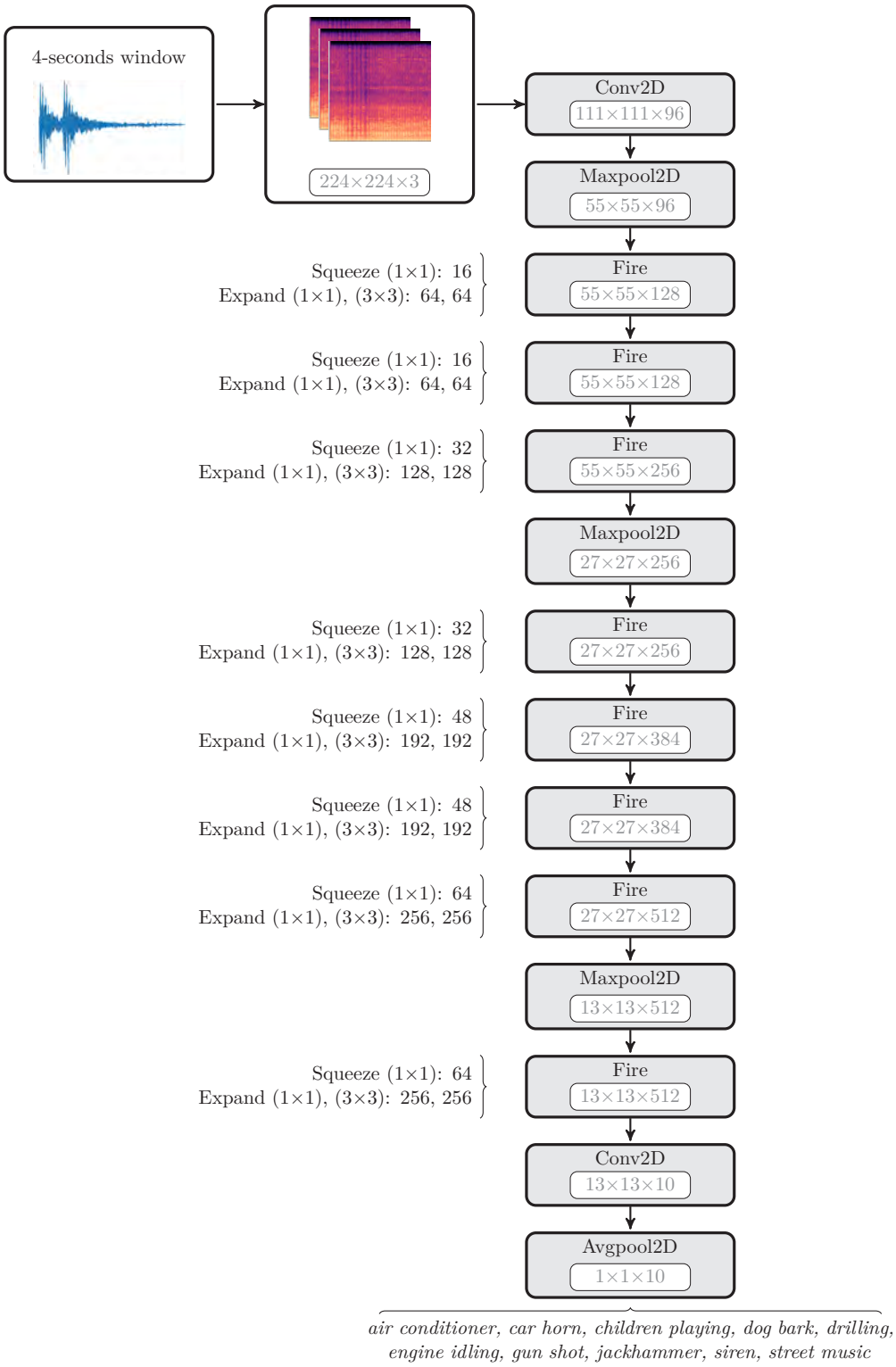


Figure II.9: Deep network architecture for the local data processing.

between each block (horizontally and vertically). Please note that in Figure II.11, white and green icons represent the acoustic sensors (microphone, antenna, and RPi), and the red dot represents an acoustic event that would be detected on nodes A, B, C and D. To deploy the proposed system, the following laboratory environment has been configured:

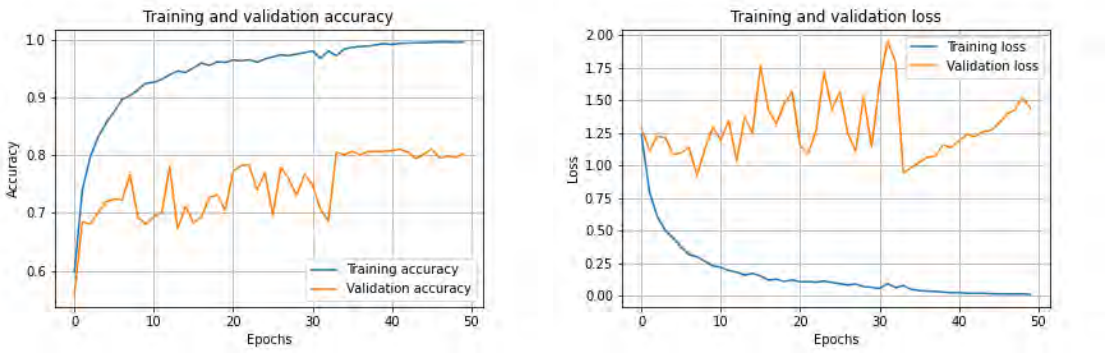


Figure II.10: Training and validation accuracy and loss of the selected model.

1. The trained model from the deep neural network used in Experiment 1 has been deployed in four different nodes (i.e., RPis). This aims to emulate the placement of the sensors in a street intersection (see Figure II.11).

2. The distributed consensus protocol has been installed at each node.

3. A sound source has been placed in the scenario (i.e., red dot in Figure II.11). To assess the advantages of the physical redundancy, a position located at different, yet reachable, distances from the four nodes was considered to be of interest (otherwise, more than one node could *hear* the same identical signal and, thus, the effect of physical redundancy could not be perceived). For this experiment, the selected distances between the sound source and nodes A, B, C, and D are 23.91 m, 57.95 m, 55.76 m, and 33.50 m respectively. This aims to emulate the location of the sound source just before the street crosswalk.

4. A synthetic acoustic test set has been generated for each node to later perform the event classification emulating how each sound would be perceived by each node. To obtain comparable results with the previous experiment, the test set has been derived from the same test fold as Experiment 1 and modified as follows:

   - The amplitude and phase of all the audio files from the testing fold have been changed according to the distance between the sound source (i.e., red dot in Figure II.11) and the sensors A, B, C and D, respectively. Hence, the same audio file has been modified as many times as neighbor nodes have been considered (four, in this case). The modifications of phase and amplitude of the samples have been carried out following the work of Bergadà et al. (Bergadà and Rosa Ma Alsina-Pagès 2019) in which, essentially, they propose the following equation:

$$y(n) = x(n - \tau)ae^{i\tau\omega}, \tag{II.1}$$

     where x is the original signal, τ is the delay on the direct path considering the speed of sound and a is the absorption coefficient times the distance between the audio source and the sensor. Please note that the absorption coefficient is dependent on the frequency, temperature and humidity. For this experiment, the average

values in the Barcelona city center have been taken: temperature of 20°C and 70% relative humidity.

- Last, but not least, urban recordings of *road traffic noise* recorded on the city center of Barcelona (Vidaña-Vila et al. 2020) have been added (i.e., weighted sum) to each audio sample to further emulate a real-world environment. This is aimed to assess what happens when the background noise partially masks the acoustic event of interest. After conducting a grid search on which realistic combinations of weights better explain the effects of physical redundancy to improve detection accuracy, the following configuration for the attenuation factors of *road traffic noise* has been selected: 0.9 for node A, 0.88 for node B, 0.7 for node C, 0.68 for node D. This configuration ensures that the events to be detected are not completely masked. Also, this makes an uneven distribution of *road traffic noise* over the nodes—note that if all the nodes were exposed to the same amount of *road traffic noise* the individual output would be the same at all of them and, thus, physical redundancy would not improve accuracy. This configuration can be best seen as emulating the presence of traffic going from South to North in the street between nodes C and D of Figure II.11 being the traffic noise closer to sensors D and C.

With this configuration, each node will later classify, at the same time, the same root acoustic sample with differences on its amplitude and phase and slightly different values of background noise.

5. Each node takes the acoustic sample from its test set, runs the local deep neural network, shares the classification vector to the neighboring nodes, and applies the consensus protocol to obtain the final result.

Table II.3 depicts the confusion matrix obtained by averaging the results of the modified audio files on the four nodes. The local accuracy of the classifier with the modified dataset is: 68.19% on node A, 62.67% on node B, 59.94% on node C and 60.42% on node D. Hence, adding the *road traffic noise* to the audio files has made accuracy decrease by a ∼20% on average (recall that in Section II.4.1, the accuracy obtained using the unaltered audio files was 80.19%). The reason behind this phenomenon is that the network has not been retrained with the modified audios. We have chosen not to retrain it for this experiment because we aim to emulate a real-world generic deployment where the background noise would be, a priori, unknown. Therefore, when deploying the system in a real-world urban environment with background traffic noise, we would expect to have the same accuracy drop.

However, after applying the distributed consensus protocol described in Section II.3.3 and considering the neighboring event vectors obtained from nodes A, B, C and D; the accuracy values obtained in each of the four nodes are: 64.47% in node A, 64.35% in node B, 62.6% in node C and 64.42% in node D, which is, on average, higher than the accuracy obtained by a single node as shown below:

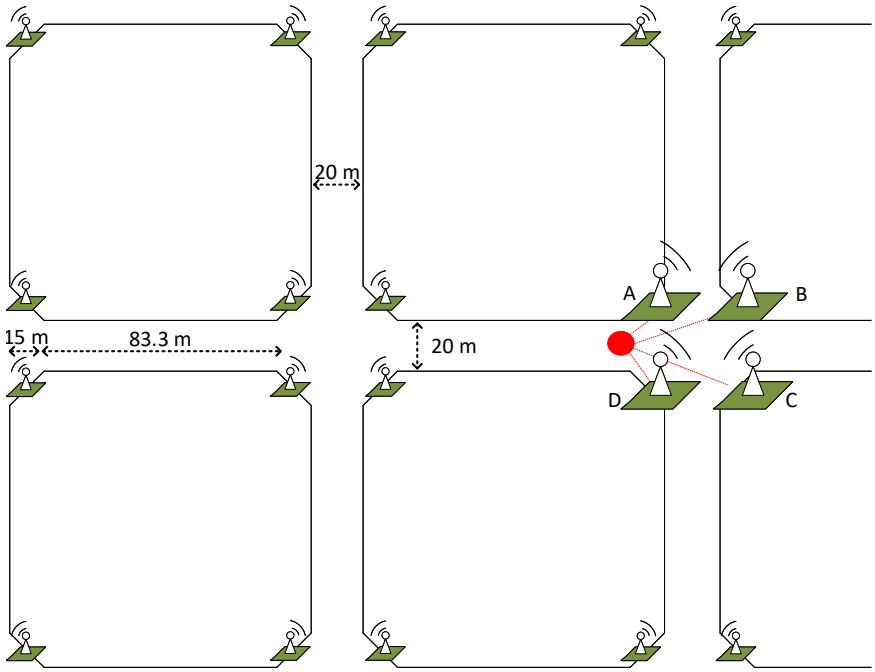


Figure II.11: Diagram of the network of sensors (nodes) in the building blocks of the city of Barcelona. The green and white icons represent the sensor devices and the red dot represents an acoustic event.

	Node A	Node B	Node C	Node D
Local accuracy	68.19%	62.67%	59.94%	60.42%
After consensus accuracy	64.47%	64.35%	62.60%	64.42%
Improvement	**−3.72%**	**+1.68%**	**+2.66%**	**+4.00%**

Table II.4 shows the average confusion matrix after applying the consensus algorithm. Observing both confusion matrices, we can see that the network tends to classify events from other categories as *Engine idling* (specially *Air conditioner*), probably because of the similarities between the spectral distribution of the *Road traffic noise* class from the BCNDataset and the *Air conditioner* and *Engine idling* classes of the UrbanSound8K dataset. The similarity between these two last events, which can be seen in Figure II.8, is further emphasized when adding *Road traffic noise* to the audio files to be classified, as it contains noises from passing cars and motorbikes that can contain fragments of engines idling.

Comparing the accuracy values obtained before and after applying the distributed consensus protocol, we can conclude that the multi-sensor approach has improved the classification results in all the sensors except for node A. As node A is the node that is closer to the acoustic event and it is the sensor where less background noise has been added, the consensus protocol has slightly decreased the accuracy of the classifier. However, this loss is compensated by the improvement in neighboring nodes.

Table II.3: Confusion matrix considering the classification of the modified audio files in a single sensor.

		Predicted Class									
		Air conditioner	*Car horn*	*Children playing*	*Dog bark*	*Drilling*	*Engine idling*	*Gunshot*	*Jackhammer*	*Siren*	*Street music*
Actual Class	*Air conditioner*	**30%**	0%	6%	0%	0%	50%	0%	0%	1%	13%
	Car horn	3%	**85%**	3%	0%	0%	3%	0%	0%	3%	3%
	Children playing	0%	0%	**84%**	6%	0%	3%	0%	0%	1%	6%
	Dog bark	0%	0%	8%	**82%**	2%	2%	0%	0%	0%	6%
	Drilling	5%	3%	3%	0%	**36%**	9%	1%	24%	2%	17%
	Engine idling	0%	0%	17%	0%	0%	**72%**	0%	0%	4%	7%
	Gunshot	6%	0%	15%	15%	0%	44%	**17%**	0%	0%	3%
	Jackhammer	0%	3%	0%	0%	2%	29%	0%	**56%**	0%	11%
	Siren	2%	0%	7%	29%	0%	1%	0%	0%	**60%**	1%
	Street music	0%	1%	23%	0%	0%	2%	0%	0%	0%	**74%**

II.5 Discussion

So far, we have shown the potential of taking advantage of the physical redundancy to increase the classification accuracy and robustness of an individual node. Alternative approaches such as the WASNs deployed in the cities of Rome and Milan on the DYNAMAP project (Sevillano et al. 2016) aim to deploy several sensors distributed in an area to enable noise map generation by interpolation, without taking into account physical redundancy. Our proposed system takes advantage of physical redundancy as the same physical space is *heard* by more than one sensor concurrently (four, in this case). According to (Mydlarz et al. 2017), the main factors that drive scalability, accuracy, adaptability, and autonomy in urban sensor networks are the following:

Monitoring sound pressure levels accurately. In our case, the proposed approach is aimed at classifying acoustic events, but as the microphone is small enough to fit a standard acoustic calibrator, the modification of the RPi software to measure sound pressure levels would be relatively straightforward.

Providing intelligent, in situ signal processing, and wireless raw audio data transmission capabilities. In our case, although the raw audio data transmission would be feasible, we have reduced the amount of data to be transmitted by taking advantage of the edge computing paradigm. In this way, the amount of data (i.e., event labels) to be transmitted among nodes is lower, which avoids bottlenecks in the communication network and, thus, shall improve the overall scalability.

Table II.4: Confusion matrix considering the classification of the modified audio files in a network of four nodes.

		Predicted Class									
		Air conditioner	*Car horn*	*Children playing*	*Dog bark*	*Drilling*	*Engine idling*	*Gunshot*	*Jackhammer*	*Siren*	*Street music*
Actual Class	*Air conditioner*	**31%**	0%	5%	0%	0%	46%	0%	0%	1%	17%
	Car horn	0%	**97%**	0%	0%	0%	3%	0%	0%	0%	0%
	Children playing	0%	0%	**82%**	7%	1%	3%	0%	0%	0%	7%
	Dog bark	0%	0%	6%	**84%**	2%	2%	0%	0%	0%	6%
	Drilling	3%	3%	2%	1%	**51%**	3%	1%	21%	3%	12%
	Engine idling	0%	0%	21%	0%	0%	**71%**	0%	0%	2%	6%
	Gunshot	3%	0%	22%	25%	0%	19%	**25%**	0%	0%	6%
	Jackhammer	0%	3%	0%	0%	1%	24%	0%	**58%**	0%	14%
	Siren	2%	0%	5%	33%	0%	0%	0%	0%	**59%**	1%
	Street music	0%	1%	17%	0%	0%	1%	0%	0%	0%	**81%**

As it is autonomous in its operation. In our case, the proposed system has been conceived considering fault tolerance by design. Therefore, if one node fails, the distributed protocol will be able to reconfigure itself to continue operating.

Having a price per node lower than or close to 100 USD. In our case, all the components of the proposed system have been conceived with low-cost devices to meet this requirement.

After demonstrating the feasibility of our proposal for urban sound monitoring in the proof-of-concept described in the previous section, we would like to share some lessons and experiences learnt during the design and development stages of the platform that might contribute to improving future versions.

II.5.1 Alternative Requirements for the Communications Antenna

The chosen operating band is 868 MHz (UN-39) because it presents advantages compared to other bands, such as robustness against absorption and higher data rates, and is not affected by a high number of other devices using this band (for example, remote controls). The antenna is designed for the UN-39 band in the ITU region 1. If the system were going to be used in another ITU region, the antenna design would need to be tuned to match the new frequency. The changes would affect the length of the crossed dipoles (*Lp* and *Lb*, in Table II.1). Accordingly, the size of the support for the antenna may also increase to accommodate the longer arms.

The isotropic radiation pattern is a good option to eliminate arrangement issues with the sensors. The current bespoke antenna design is prepared to radiate in all directions so that

the position of the sensor will not affect the communication link. This is an advantage of the current setup, with a medium range of 200 m maximum. For a longer range, the gain of the antenna may need to increase, losing the isotropic radiation property. In this case, by increasing the Δ in Table II.1, the antenna will increase its gain, which will compensate for losses for the longer path–as shown in Friis Transmission Equation (Pozar 2011). Still, this new configuration will lose the ability to be unaltered by the sensor positions.

II.5.2 Fault Tolerance

As discussed in Section II.3, the proposed USN must tolerate a certain degree of faults. This means that the system must keep operating in case of failure in a node or communication link. The system is designed to support a *fail-stop* failures in a limited number of nodes or communication links.

For *fail-stop* failures in the nodes or the communication links, the system behaves as follows. When a node detects that the last time it received the token of the distributed consensus protocol is higher than a predefined threshold, it will try to reach the following node of the ring. If the communication is successful the ring will be reconfigured. If the communication is not successful, the node will change the token direction (i.e., clockwise or counterclockwise) and the system will adopt a token bus behavior instead of a token ring.

Additionally, each node should incorporate self-reboot policies (e.g., every 48 h, and/or when connection with neighbors is lost for more than 1 min) to avoid the nodes being frozen forever.

II.5.3 Real-World Deployment of the Proposed System

So far, the proposed system (i.e., deep network, communications antenna, and consensus protocol) has been assessed under laboratory conditions as shown in Section II.4. Methodologically, this has enabled us to individually validate each component of the system and its end-to-end performance under a controlled environment. The lessons learnt during this process have let us consider the following points when deploying the system in a real-world scenario:

- The location of the nodes should be selected according to the architectonic profile of the scenario to enable physical redundancy. Please note that the proposed approach tolerates the addition or the removal of nodes at will. Also, if larger communication distances—while ensuring that several nodes can *hear* the same high *Leq* acoustic event—were required due to the scenario characteristics, alternative strategies such as simultaneous wireless information and power transfer could be explored (Ji et al. 2020b).
- In the case of microphones different from the OUT-AMLO-0872 being selected (e.g., MEMS), the most important requirements during deployment would be (1) omnidirectional pattern so they are able to pick up signal equally from all directions—to facilitate the installation of each node—(2) flat frequency response in the frequency

range of, at least, 50 Hz–10 KHz (i.e., where acoustic events are taking place), and (3) 16 bits resolution per sample.

- As the same way as the microphone, the communications antenna has been designed to radiate following an isotropic pattern to facilitate its real-world deployment (i.e., no matter how the node is oriented). Although the electromagnetic interferences that may degrade the performance of the proposed antenna have not been considered in this work, it is worth mentioning that the Transport Control Protocol (TCP/IP) can be used to detect when the frames between nodes are lost or corrupted. If that happened, the consensus protocol would reconfigure the ring accordingly as described above.

- Experiments conducted over the RPi platform using the UrbanSound8K dataset suggest that the proposed system architecture would be capable of detecting acoustic events in real time using a deep convolutional neural network. However, when deploying the system in a real-world scenario, the classifier might struggle to distinguish anomalous acoustic events in noisy environments (i.e., locations with traffic background noise partially masking the acoustic events). Hence, the Squeezenet model should be retrained using data collected on the location where the nodes would be located. If, in the future, other types of events (not included in the UrbanSound8K) were to be detected, the following modifications should be made to the system:

  1. Add as many neurons as new event types to the last CONV2D layer of the deep network. In this case, the network should be retrained to be able to classify the new categories.
  2. Increase the size of the events vector sent to the neighbor nodes.
  3. Adapt the heuristic rules of the distributed consensus protocol to decide whether the new class contains noises that typically have a low value of L_{eq} or not.

  Therefore, the proposed acoustic USN could be easily adapted to potential classification of new event types.

II.6 Conclusions and Future Work

This research presents a low-cost acoustic sensor network to monitor urban sounds in large-scale areas. The proposed approach uses a pipeline composed of the following stages: (1) acoustic data acquisition and spectrogram computation, (2) local classification using a convolutional neural network (SqueezeNet architecture), (3) a custom bespoke antenna with isotropic radiation to share the local predictions with neighboring nodes, and (4) a distributed consensus protocol and a set of heuristic rules to unify the local predictions conducted at each node. To validate this proposal, the urban environment of the city of Barcelona has been selected. The proposed system detects the most probable events occurred on an acoustic sample taking advantage of the physical redundancy of the nodes. Regarding the physical redundancy, there are several reasons to consider four nodes per street intersection. The first of them is that

the authors have chosen the Eixample district of Barcelona to conduct these experiments due to its symmetric structure. All the street intersections are of the same size and distance, which facilitates the design of a symmetric network. This leads us to the second reason. The number of nodes per street is probably too redundant, and possibly two nodes would have been sufficient to detect the noise events occurring around the intersections. However, the goal of the design is to have lots of low-cost nodes, collecting the same type of data at the same time for a large number of locations simultaneously. This data redundancy is based on the concept that a low-cost node can appear as a commodity for the project, and it is the only way of gathering a huge amount of data to, not only by reliably detecting the acoustic events occurring, but also by having enough available information for other future applications such as drawing a precise map of the noise levels and their noise source.

A further potential application of this system is to automatically test whether a specific urban area meets certain acoustic regulations: for instance, when a specific event (e.g., air conditioner) is detected, it could be straightforward to decide whether the L_{Aeq} is below its associated threshold. Indeed, the obtained results of the proposed system encourage researchers to continue working on this direction, which in later stages will go through its implementation in a real-world and real-operation environment. This will enable practitioners to (1) evaluate the validity of the training carried out using UrbanSound8K and BCNDataset, and (2) verify the completeness of the model used for acoustic propagation, assuming that in a real-world situation the additive noise from the street will be more relevant.

Actually, in a real-world environment multiple events may occur simultaneously in the same acoustic sample. To maintain the excellent results in single-event-detection obtained in this validation test, the proposed deep network should move into a multi-label acoustic samples training, hence assuming that multiple events will occur simultaneously (Cartwright et al. 2019). In this future stage, the consensus function of the distributed protocol should be adapted to tolerate the identification of multiple events.

Author's contributions

All authors have significantly contributed to this work. Conceptualization, E.V.-V., J.N., C.B.-F. and R.M.A.-P.; Data curation, E.V.-V.; Formal analysis, E.V.-V., J.N., C.B.-F., D.S. and R.M.A.-P.; Funding acquisition, R.M.A.-P.; Investigation, E.V.-V., J.N., C.B.-F. and D.S.; Methodology, J.N. and D.S.; Project administration, R.M.A.-P.; Resources, R.M.A.-P. and J.N.; Software, E.V.-V. and C.B.-F.; Supervision, J.N., D.S. and R.M.A.-P.; Validation, D.S. and R.M.A.-P.; Visualization, C.B.-F.; Writing – original draft, E.V.-V., J.N., C.B.-F. and R.M.A.-P.; Writing – review & editing, D.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research was partially funded by the Secretaria d'Universitats i Recerca of the Department of Business and Knowledge of the Generalitat de Catalunya under grants 2017-SGR-966 and

2017-SGR-977. Ester Vidaña-Vila and Rosa Ma Alsina-Pagès would like to thank La Salle Campus BCN - URL for partially funding the joint research with Queen Mary University (London) in the framework of Ms Vidaña-Vila PhD Thesis. Also, this work was partially funded by the Spanish Ministry of Science, Innovation and University, the Investigation State Agency and the European Regional Development Fund (ERDF) under grant RTI2018-097066-B-I00 for Joan Navarro and Cristina Borda-Fortuny.

Acknowledgements

The authors would like to thank Lisa Kinnear for her never-ending patience, support and thorough review of this work. Also, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADC	Analog-to-Digital Converter
CNN	Convolutional Neural Network
FLOP	Floating-Point Operation
GPU	Graphics Processing Unit
L_{Aeq}	Equivalent Level
MEMS	Micro Electrical Mechanical System
NMN	Noise Monitoring Network
USN	Ubiquitous Sensor Network
RGB	Red-Green-Blue
RPi	Raspberry Pi Model 2B
SGD	Stochastic Gradient Descent
SNR	Signal-to-Noise Ratio
SPLs	Sound Pressure Levels
WHO	World Health Organization
WASN	Wireless Acoustic Sensor Network

References

13, ITU-T Study Group (2008). *Ubiquitous Sensor Networks (USN)*. Tech. rep. East Lansing, Michigan: ITU-T Technology Watch Briefing Report Series, No. 4, p. 10.

Aibar, Eduardo and Bijker, Wiebe E (1997). 'Constructing a city: The Cerdà plan for the extension of Barcelona'. In: *Science, technology, & human values* vol. 22, no. 1, pp. 3–30.

Aiello, William, Bhatt, Sandeep N, Chung, Fan RK, Rosenberg, Arnold L and Sitaraman, Ramesh K (2001). 'Augmented ring networks'. In: *IEEE Transactions on Parallel and Distributed Systems* vol. 12, no. 6, pp. 598–609.

Alexander, W (1968). 'Some harmful effects of noise.' In: *Canadian Medical Association Journal* vol. 99, no. 1, p. 27.

Alsina-Pagès, Rosa Ma, Alías, Francesc, Socoró, Joan Claudi and Orga, Ferran (2018). 'Detection of Anomalous Noise Events on Low-Capacity Acoustic Nodes for Dynamic Road Traffic Noise Mapping within an Hybrid WASN'. In: *Sensors* vol. 18, no. 4, p. 1272.

Alsina-Pagès, Rosa Maria, Hervás, Marcos, Duboc, Leticia and Carbassa, Jordi (2020). 'Design of a Low-Cost Configurable Acoustic Sensor for the Rapid Development of Sound Recognition Applications'. In: *Electronics* vol. 9, no. 7, p. 1155.

Armbrust, Michael et al. (2010). 'A view of cloud computing'. In: *Communications of the ACM* vol. 53, no. 4, pp. 50–58.

Bagula, Antoine, Zennaro, Marco, Inggs, Gordon, Scott, Simon and Gascon, David (2012). 'Ubiquitous sensor networking for development (usn4d): An application to pollution monitoring'. In: *Sensors* vol. 12, no. 1, pp. 391–414.

Bartalucci, Chiara, Borchi, Francesco, Carfagni, Monica, Furferi, Rocco, Governi, Lapo, Lapini, Alessandro, Bellomini, Raffaella, Luzzi, Sergio and Nencini, Luca (2018). 'The smart noise monitoring system implemented in the frame of the Life MONZA project'. In: *Proceedings of the EuroNoise*, pp. 783–788.

Basten, Tom and Wessels, Peter (2014). 'An overview of sensor networks for environmental noise monitoring'. In: *Proceedings of the 21st International Congress on Sound and Vibration (ICSV21)*. Beijing, China, pp. 1–8.

Su-bei, MENG (2007). 'Harm to human health from low frequency noise in city residential area [J]'. In: *China Medical Herald* vol. 35.

Bell, Margaret Carol and Galatioto, Fabio (2013). 'Novel wireless pervasive sensor network to improve the understanding of noise in street canyons'. In: *Applied Acoustics* vol. 74, no. 1, pp. 169–180.

Bello, Juan P, Silva, Claudio, Nov, Oded, Dubois, R Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles and Doraiswamy, Harish (2019). 'Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution'. In: *Communications of the ACM* vol. 62, no. 2, pp. 68–77.

Bellucci, Patrizia and Cruciani, Francesca Romana (2016). 'Implementing the Dynamap system in the suburban area of Rome'. In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. 253 3. Institute of Noise Control Engineering, pp. 5518–5529.

Bellucci, Patrizia, Peruzzi, Laura and Zambon, Giovanni (2017). 'LIFE DYNAMAP project: The case study of Rome'. In: *Applied Acoustics* vol. 117, pp. 193–206.

Bergadà, Pau and Alsina-Pagès, Rosa Ma (2019). 'An Approach to Frequency Selectivity in an Urban Environment by Means of Multi-Path Acoustic Channel Analysis'. In: *Sensors* vol. 19, no. 12, p. 2793.

Botteldooren, Dick, De Coensel, Bert, Oldoni, Damiano, Van Renterghem, Timothy and Dauwe, Samuel (Nov. 2011). 'Sound monitoring networks new style'. eng. In: *Acoustics 2011: Breaking New Ground : Proceedings of the Annual Conference of the Australian Acoustical Society*. Ed. by J Mee, David and Hillock, Ian DM. Queensland, Australia: Australian Acoustical Society, 93:1–93:5.

Brown, AL and Coensel, B De (2018). 'A study of the performance of a generalized exceedance algorithm for detecting noise events caused by road traffic'. In: *Applied Acoustics* vol. 138, pp. 101–114.

Camps-Farrés, Júlia (2015). 'Barcelona noise monitoring network'. In: *Proceedings of the Euronoise*, pp. 218–220.

Camps-Farrés, Júlia and Casado-Novas, Javier (May 2018). 'Issues and challenges to improve the Barcelona Noise Monitoring Network'. In: *Proceedings of EuroNoise 2018*. Heraklion, Crete – Greece: EAA — HELINA, pp. 693–698.

Cartwright, Mark, Mendez, Ana Elisa Mendez, Cramer, Jason, Lostanlen, Vincent, Dove, Graham, Wu, Ho-Hsiang, Salamon, Justin, Nov, Oded and Bello, Juan (2019). 'Sonyc urban sound tagging (sonyc-ust): a multilabel dataset from an urban acoustic sensor network'. In.

Cense - Characterization of urban sound environments (n.d.). http://cense.ifsttar.fr/.

CNAF (n.d.[a]). *Artículo 5 del Reglamento de Radiocomunicaciones*. https://avancedigital.gob.es/espectro/CNAF/notasRR-2017.pdf. Accessed: 2020-07-30.

CNAF (n.d.[b]). *ATRIBUCIÓN A LOS SERVICIOS según el RR de la UIT*. https://avancedigital.gob.es/espectro/CNAF/tablas_2017.pdf. Accessed: 2020-07-30.

CNAF (n.d.[c]). *Notas UN*. https://avancedigital.gob.es/espectro/CNAF/notas-UN-2017.pdf. Accessed: 2020-07-30.

Coulson, Saskia, Woods, Mel, Scott, Michelle, Hemment, Drew and Balestrini, Mara (2018). 'Stop the noise! enhancing meaningfulness in participatory sensing with community level indicators'. In: *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 1183–1192.

De Coensel, Bert and Botteldooren, Dick (Nov. 2014). 'Smart sound monitoring for sound event detection and characterization'. In: *Proceedings of the 43rd International Congress on Noise Control Engineering (Inter-Noise 2014)*. Melbourne, Australia, pp. 1–10.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai and Fei-Fei, Li (2009). 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Domínguez, Federico, Dauwe, Samuel, Cuong, Nguyen The, Cariolaro, Dimitri, Touhafi, Abdellah, Dhoedt, Bart, Botteldooren, Dick and Steenhaut, Kris (2014). 'Towards an environmental measurement cloud: Delivering pollution awareness to the public'. In: *International Journal of Distributed Sensor Networks* vol. 10, no. 3, p. 541360.

Ferrández-Pastor, Francisco Javier, García-Chamizo, Juan Manuel, Nieto-Hidalgo, Mario, Mora-Pascual, Jerónimo and Mora-Martínez, José (2016). 'Developing ubiquitous sensor network platform using internet of things: Application in precision agriculture'. In: *Sensors* vol. 16, no. 7, p. 1141.

Flindell, IH and Walker, JG (2004). 'Environmental noise management'. In: *Advanced Applications in Acoustics, Noise and Vibration*, p. 183.

Ghemawat, Sanjay, Gobioff, Howard and Leung, Shun-Tak (2003). 'The Google file system'. In: *Proceedings of the nineteenth ACM symposium on Operating systems principles*, pp. 29–43.

González, Lisardo Prieto, Jaedicke, Corvin, Schubert, Johannes and Stantchev, Vladimir (2016). 'Fog computing architectures for healthcare'. In: *Journal of Information, Communication and Ethics in Society*.

Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron and Bengio, Yoshua (2016). *Deep learning*. Vol. 1. MIT press Cambridge.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing and Sun, Jian (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Huang, Gao, Liu, Zhuang and Weinberger, Kilian Q. (2016). 'Densely Connected Convolutional Networks'. In: *CoRR* vol. abs/1608.06993. arXiv: 1608.06993.

Hurtley, Charlotte (2009). *Night noise guidelines for Europe*. WHO Regional Office Europe.

Huzaifah, Muhammad (2017). 'Comparison of time-frequency representations for environmental sound classification using convolutional neural networks'. In: *arXiv preprint arXiv:1706.07156*.

Iandola, Forrest N., Moskewicz, Matthew W., Ashraf, Khalid, Han, Song, Dally, William J. and Keutzer, Kurt (2016). 'SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size'. In: *CoRR* vol. abs/1602.07360. arXiv: 1602.07360.

Ji, Baofeng, Chen, Zhenzhen, Chen, Sudan, Zhou, Benchuan, Li, Chunguo and Wen, Hong (2020a). 'Joint optimization for ambient backscatter communication system with energy harvesting for IoT'. In: *Mechanical Systems and Signal Processing* vol. 135, p. 106412.

Ji, Baofeng, Chen, Zhenzhen, Mumtaz, Shahid, Liu, Jianghui, Zhang, Yong, Zhu, Jia and Li, Chunguo (2020b). 'SWIPT Enabled Intelligent Transportation Systems with Advanced Sensing Fusion'. In: *IEEE Sensors Journal*.

Koucheryavy, Andrey, Vladyko, Andrey and Kirichek, Ruslan (2015). 'State of the art and research challenges for public flying ubiquitous sensor networks'. In: *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Springer, pp. 299–308.

Links, Extreme Range (2016). 'LoRa 868/900 MHz SX1272 LoRa Module for Arduino Waspmote and Raspberry Pi'. In: *Online in https://www. cooking-hacks. com/documentation/tutorials/. Consulted in February*.

Ma, Ningning, Zhang, Xiangyu, Zheng, Hai-Tao and Sun, Jian (2018). 'ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design'. In: *CoRR* vol. abs/1807.11164. arXiv: 1807.11164.

Mesaros, Annamaria, Heittola, Toni, Benetos, Emmanouil, Foster, Peter, Lagrange, Mathieu, Virtanen, Tuomas and Plumbley, Mark D (2017). 'Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 26, no. 2, pp. 379–393.

Mietlicki, Christophe and Mietlicki, Fanny (2018). 'Medusa: a new approach for noise management and control in urban environment'. In: *Proceedings of the 11th European Congress and Exposition on Noise Control Engineering (Euronoise2018), Crete, Greece*, pp. 27–31.

Mietlicki, Fanny, Mietlicki, Christophe and Sineau, Matthieu (May 2015). 'An innovative approach for long-term environmental noise measurement: RUMEUR network'. In: *Proceedings of EuroNoise 2015*. Maastrich, Netherlands: EAA-NAG-ABAV, pp. 2309–2314.

Ministerio de Energía, Turismo y Agenda Digital (n.d.). *Real Decreto 123/2017, de 24 de febrero, por el que se aprueba el Reglamento sobre el uso del dominio público radioeléctrico*. https://www.boe.es/buscar/act.php?id=BOE-A-2017-2460&tn=1&p=20170308#ar-6. Accessed: 2020-07-30.

Moudon, Anne Vernez (2009). 'Real noise from the urban environment: how ambient community noise affects health and what can be done about it'. In: *American journal of preventive medicine* vol. 37, no. 2, pp. 167–171.

Mun, Sungho and Geem, Zong Woo (2009). 'Determination of individual sound power levels of noise sources using a harmony search algorithm'. In: *International Journal of Industrial Ergonomics* vol. 39, no. 2, pp. 366–370.

Murty, Rohan Narayana, Mainland, Geoffrey, Rose, Ian, Chowdhury, Atanu Roy, Gosain, Abhimanyu, Bers, Josh and Welsh, Matt (2008). 'Citysense: An urban-scale wireless sensor network and testbed'. In: *2008 IEEE conference on technologies for homeland security*. IEEE, pp. 583–588.

Mydlarz, Charlie, Salamon, Justin and Bello, Juan Pablo (2017). 'The implementation of low-cost urban acoustic monitoring devices'. In: *Applied Acoustics* vol. 117, pp. 207–218.

Nanni, Loris, Costa, Yandre MG, Aguiar, Rafael L, Mangolin, Rafael B, Brahnam, Sheryl and Silla, Carlos N (2020). 'Ensemble of convolutional neural networks to improve animal audio classification'. In: *EURASIP Journal on Audio, Speech, and Music Processing* vol. 2020, no. 1, pp. 1–14.

Navarro, Joan, Vidaña-Vila, Ester, Alsina-Pagès, Rosa Ma and Hervás, Marcos (2018). 'Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios'. In: *Sensors* vol. 18, no. 8, p. 2492.

Office, Parliamentary Counsel's (2017). 'Protection of the Environment Operations (Noise Control) Regulation 2017'. In: *Legal Service Bull.* vol. 1, p. 44.

Pan, G., Li, Y., Zhang, Z. and Feng, Z. (2012). 'Isotropic Radiation From a Compact Planar Antenna Using Two Crossed Dipoles'. In: *IEEE Antennas and Wireless Propagation Letters* vol. 11, pp. 1338–1341.

Paulo, J, Fazenda, P, Oliveira, T, Carvalho, C and Félix, M (2015). 'Framework to monitor sound events in the city supported by the FIWARE platform'. In: *Proceedings of the 46o Congreso Español de Acústica*. Valencia, Spain, pp. 21–23.

Paulo, J, Fazenda, Pedro, Oliveira, Tiago and Casaleiro, Joao (June 2016). 'Continuos sound analysis in urban environments supported by FIWARE platform'. In: *Proceedings of the EuroRegio2016/TecniAcústica*. Porto, Portugal, pp. 1–10.

Pham, Congduc and Cousin, Philippe (2013). 'Streaming the sound of smart cities: Experimentations on the smartsantander test-bed'. In: *2013 IEEE international conference on green computing and communications and IEEE internet of things and IEEE cyber, physical and social computing*. IEEE, pp. 611–618.

Piper, Ben, Barham, Richard, Sheridan, Steven and Sotirakopoulos, Kostas (2017). 'Exploring the "big acoustic data" generated by an acoustic sensor network deployed at a crossrail construction site'. In: *Proceedings of the 24th International Congress on Sound and Vibration (ICSV), London, UK*, pp. 23–27.

Polastre, Joseph, Szewczyk, Robert and Culler, David (2005). 'Telos: enabling ultra-low power wireless research'. In: *Proceedings of the 4th international symposium on Information processing in sensor networks*. IEEE Press, p. 48.

Pozar, D.M. (2011). *Microwave Engineering, 4th Edition*. Wiley.

Rainham, D (2016). 'A wireless sensor network for urban environmental health monitoring: UrbanSense'. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 34. 1. IOP Publishing, p. 012028.

Raspberry Pi Official web site (n.d.). https://www.raspberrypi.org. Accessed: 2020-10-01.

Salamon, J., Jacoby, C. and Bello, J. P. (Nov. 2014). 'A Dataset and Taxonomy for Urban Sound Research'. In: *22nd ACM International Conference on Multimedia (ACM-MM'14)*. Orlando, FL, USA, pp. 1041–1044.

Salamon, Justin and Bello, Juan Pablo (2015). 'Unsupervised feature learning for urban sound classification'. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 171–175.

Sandler, Mark, Howard, Andrew G., Zhu, Menglong, Zhmoginov, Andrey and Chen, Liang-Chieh (2018). 'Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation'. In: *CoRR* vol. abs/1801.04381. arXiv: 1801.04381.

Santini, Silvia, Ostermaier, Benedikt and Vitaletti, Andrea (2008). 'First experiences using wireless sensor networks for noise pollution monitoring'. In: *Proceedings of the 2008 Workshop on Real-World Wireless Sensor Networks (REALWSN)*. ACM, pp. 61–65.

Santini, Silvia and Vitaletti, Andrea (2007). 'Wireless sensor networks for environmental noise monitoring'. In: *6. Fachgesprach Sensornetzwerke*, p. 98.

Sevillano, Xavier et al. (2016). 'DYNAMAP–Development of low cost sensors networks for real time noise mapping'. In: *Noise mapping* vol. 1, no. open-issue.

Shin, Daejung, Na, Seung You, Kim, Jin Young and Baek, Seong-Joon (2007). 'Fish robots for water pollution monitoring using ubiquitous sensor networks with sonar localization'.

In: *2007 International Conference on Convergence Information Technology (ICCIT 2007)*. IEEE, pp. 1298–1303.

Singh, Shubhr, Pankajakshan, Arjun and Benetos, Emmanouil (Oct. 2019). 'Audio Tagging using Linear Noise Modelling Layer'. In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, NY, USA, pp. 234–238.

Socoró, Joan Claudi, Alías, Francesc and Alsina-Pagès, Rosa Ma (2017). 'An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments'. In: *Sensors* vol. 17, no. 10, p. 2323.

Test, Tsafnat, Canfi, Ayala, Eyal, Arnona, Shoam-Vardi, Ilana and Sheiner, Einat K (2011). 'The influence of hearing impairment on sleep quality among workers exposed to harmful noise'. In: *Sleep* vol. 34, no. 1, pp. 25–30.

Vidaña-Vila, Ester, Duboc, Leticia, Alsina-Pagès, Rosa Ma, Polls, Francesc and Vargas, Harold (2020). 'BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset'. In: *Sustainability* vol. 12, no. 19, p. 8140.

Wang, Cuiping, Chen, Guoqiang, Dong, Rencai and Wang, Haowei (2013). 'Traffic noise monitoring and simulation research in Xiamen City based on the Environmental Internet of Things'. In: *International Journal of Sustainable Development & World Ecology* vol. 20, no. 3, pp. 248–253.

WHO/Europe | Noise - Data and statistics (n.d.). www.euro.who.int/en/health-topics/environment-and-health/noise/data-and-statistics. Accessed: 2020/09/06.

Wikipedia contributors (2020). *Eixample — Wikipedia, The Free Encyclopedia*. [Online; accessed 5-October-2020].

Zambon, Giovanni, Benocci, Roberto, Bisceglie, Alessandro, Roman, H. Eduardo and Bellucci, Patrizia (2017). 'The LIFE DYNAMAP project: Towards a procedure for dynamic noise mapping in urban areas'. In: *Applied Acoustics* vol. 124, pp. 52–60.

Zhang, Xiangyu, Zhou, Xinyu, Lin, Mengxiao and Sun, Jian (2017). 'ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices'. In: *CoRR* vol. abs/1707.01083. arXiv: 1707.01083.

Authors' addresses

Ester Vidaña-Vila GTM – Grup de Recerca en Tecnologies Mèdia, La Salle Campus Barcelona - Universitat Ramon Llull Quatre Camins, 30, 08022 Barcelona, Spain
ester.vidana@salle.url.edu

Paper III

Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors

III

Ester Vidaña-Vila, Joan Navarro, Dan Stowell, Rosa Ma Alsina-Pagès

Published in *Sensors*, November 2021, volume 21, issue 22, pp. 7470. DOI: 10.3390/s21227470.

Abstract

Many people living in urban environments nowadays are overexposed to noise, which results in adverse effects on their health. Thus, urban sound monitoring has emerged as a powerful tool that might enable public administrations to automatically identify and quantify noise pollution. Therefore, identifying multiple and simultaneous acoustic sources in these environments in a reliable and cost-effective way has emerged as a hot research topic. The purpose of this paper is to propose a two-stage classifier able to identify, in real time, a set of up to 21 urban acoustic events that may occur simultaneously (i.e., multilabel), taking advantage of physical redundancy in acoustic sensors from a wireless acoustic sensors network. The first stage of the proposed system consists of a multilabel deep neural network that makes a classification for each 4-second window. The second stage intelligently aggregates the classification results from the first stage of four neighboring nodes to determine the final classification result. Conducted experiments with real-world data and up to three different computing devices show that the system is able to provide classification results in less than 1 s and that it has good performance when classifying the most common events from the dataset. The results of this research may help civic organisations to obtain actionable noise monitoring information from automatic systems.

III.1 Introduction

Acoustic noise (or noise pollution) can be defined as any sound that is loud or unpleasant enough that causes some kind of disturbance (Moudon 2009). Noise pollution is one of

the major concerns for the European population, especially for citizens living in urban environments (Hurtley 2009), which is materialized in an ever-rising number of complaints to public administrations (Organization et al. 2019). This issue is further stressed on those residential areas located close to aggressive noise pollutants such as airports, railways, or highways (WHO 2011). In fact, according to the World Health Organization (WHO) (Hurtley 2009; WHO 2011), there is a worrying portion of the European population that is systematically exposed to harmful levels of noise pollution. Concretely, it is estimated that from all the citizens living in the European Union (EU), about 40% are exposed to road traffic noise levels above 55 dB(A), about 20% are exposed to levels above 65 dB(A) in daytime, and over 30% are exposed to noise levels exceeding 55 dB(A) at nighttime (*Data and statistics* n.d.).

Continuous exposure to environmental noise pollution may result in adverse effects on health, ranging from moderate disturbances such as difficulties in understanding a voice message to chronic illnesses such as cardiovascular diseases (e.g., myocardial infarction), cognitive impairment in children, psychological disorders derived from lack of rest or sleep, or tinnitus (Test et al. 2011; WHO 2011). In fact, according to the European Environment Agency (EEA), it is estimated that, only in Europe, 12,000 premature deaths are associated with long-term noise exposure each year. Moreover, in their latest report (*Noise* n.d.), it is estimated that more than 28 million people suffer from the aforementioned health effects derived from overexposure to noise.

In order to overcome this situation, a set of recommendations have been established (e.g., Environmental Noise Directive 2002/49/EC (*Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002* n.d.) from the European Commission or the Environmental Noise Guidelines for the European Region from the WHO (Guski et al. 2017)) to define the thresholds on the maximum amount of noise that should be perceived by citizens. For instance, the WHO distinguishes up to five different types of noise sources (i.e., road traffic noise, railway noise, aircraft noise, wind turbine noise, and leisure noise) and recommends different noise thresholds for each source depending on the time of the day (i.e., day, night) (Organization et al. 2018). From these recommendations, it can be inferred that not all sound sources have the same impact on human disturbance. In fact, the sound level is not the only parameter that indicates the extent and intensity of noise pollution (Abbaspour et al. 2015). 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 private and public entities on identifying acoustically polluted environments in urban areas (Bello et al. 2019). Typically, this is done by either analyzing the distribution of noise-related complaints in a certain area or by deploying a Wireless Acoustic Sensor Network (WASN) to automatically monitor the soundscape (Bello et al. 2019). Both approaches entail the same main underlying challenges:

1. Identifying multiple concurrent noise sources that populate a given soundscape. Typically, in real-world environments, several sounds occur simultaneously. This complicates the task of building a reliable automatic sound classifier system specialized in identifying a

predefined set of acoustic events (Fonseca et al. 2019).

2. Monitoring large-scale urban areas in a cost-effective way. Populating (with either automatic devices or human resources) extensive urban environments requires a considerable amount of resources. For instance, it has been reported (Bello et al. 2019) that the Department of Environmental Protection from New York City employs about 50 highly qualified sound inspectors. In addition, the starting price of autonomous nodes to continuously monitoring sound is usually around EUR 1000 (Mejvald and Konopa 2019).

3. Real-time processing. Although continuous exposure to noise is harmful, short-term exposure to sporadic noise shall not be neglected. In fact, sometimes noise violations are sporadic (i.e., they last a few minutes or hours at most). Therefore, human-based noise complaint assessment systems result in being ineffective due to the fact that technicians may arrive way after the disturbance has finished (Bello et al. 2019). Furthermore, the large amount of data to be processed by autonomous acoustic sensors may make this kind of approach challenging.

The purpose of this paper is to present an automatic classification system for acoustic events in urban environments able to address the aforementioned challenges. The proposed approach combines and improves (1) the advances of our previous work in the conception of a WASN architecture for single-label classification using physical redundancy of low-cost sensors and synthetically generated audio files (Vidaña-Vila et al. 2020c), and (2) the outlined automatic multilabel classification system for acoustic events that the authors presented in (Vidaña-Vila et al. 2021). The resulting system presented in this work features a two-stage classifier that analyzes real-world acoustic frames in real time to distinguish all the events that appear in them——not only on the foreground soundscape but also on the background. It is understood that events in the foreground are those with more saliency than the average noise. Similarly, events in the background are those events with similar saliency to the average noise.

The first stage is composed of a deep neural network that has been trained to identify different events that may occur concurrently (also referred to as a multilabel classification). The second layer is aimed at aggregating the first-stage classification results from neighboring nodes (i.e., exploiting physical redundancy) to increase the classification reliability of individual sensors. Additionally, the whole system has been designed so it can meet the computing constraints typically found in the potential application domain of this system (i.e., low-cost WASN (Vidaña-Vila et al. 2020c)). In order to assess the classification performance of the presented approach, real-world data have been collected simultaneously at four corners of a traffic intersection in Barcelona.

Overall, the contributions of this paper are the following:

- A new real-world 5 h length dataset (containing concurrent events) recorded simultaneously at four spots from a street intersection. This results in 4 $\times$ 5 h of acoustic data. A total of 5 h of audio data corresponding to 1 spot have been manually annotated. To the best of our knowledge, this is the first dataset with these characteristics.

- A software-assisted strategy to reduce the number of user interactions when labelling acoustic data to reduce the amount of time spent on this task.

- A two-stage acoustic classifier aimed at increasing the local classification robustness by taking into consideration the classification results of neighboring nodes (i.e., exploiting the nodes' physical redundancy).

Conducted experiments over different low-cost architectures (Raspberry Pi 2B, 3B+, and 4) endorse the feasibility of our approach and encourage practitioners to extend this work in a large-scale real-world deployment.

The remainder of this paper is organized as follows. Section III.2 reviews the related work on the identification of acoustic events in urban environments. Section III.3 describes the real-world data collection and annotation processes that have led to the training and test sets used to assess the classification performance. Section III.4 details the proposed two-stage multilabel classifier system. Section III.5 evaluates the proposed approach. Section III.6 discusses the main findings of this work. Finally, Section III.7 concludes the paper and proposes potential future work directions.

III.2 Related Work

There is an increasing demand for automatic monitoring of noise levels in urban areas, especially if this monitoring can give information about the noise source of the measured levels (Guski et al. 2017; Organization et al. 2018). In this sense, several WASN-based projects are being developed in several parts of the world, mainly adapted to their requirements, some of them identifying types of noise source and others giving equivalent levels L_{Aeq}. Following this idea, some projects have to develop their own sensors to meet the requirements of the measurements, and others operate in the real world with commercial sensors. Additionally, there are some projects that do not only concentrate on noise monitoring but also on air pollution.

III.2.1 Commercial Sensor Networks

Commercial sound level meters or sensor networks are usually connected to a central server, which collects all the L_{Aeq} values gathered by the nodes. One of the first projects developed for this purpose is the Telos project (Polastre et al. 2005), which was one of the first experiences in this WASN design by means of an ultra-low-power wireless sensor module designed by the University of California (Berkeley). Some years later, a WASN was used in a large variety of environmental monitoring applications, with a central focus on urban sound, as we can find in (Santini and Vitaletti 2007; Santini et al. 2008).

In Xiamen City (China), authors deployed a traffic noise monitoring network covering 35 roads and 9 green spaces in the city (C. Wang et al. 2013). Data from the environmental monitoring stations were used to model the traffic of more than 100 roads in the city. Similarly, the FI-Sonic Project is focused on noise monitoring in a surveillance mode (Paulo et al. 2015).

Its main goal is to develop the artificial intelligence algorithms required to identify the location of sound events (Paulo et al. 2016) based on a FIWARE platform. The RUMEUR project, standing for Urban Network of Measurement of the sound Environment of Regional Use, is a hybrid wireless sensor network deployed by BruitParif (F. Mietlicki et al. 2015) in Paris and its surrounding cities. It has been designed to have high accuracy in critical places, such as airports, where the WHO directive has defined stringent thresholds (*Data and statistics* n.d.), while other locations have less precise measurements. Years after, the RUMEUR project has evolved to Medusa (C. Mietlicki and F. Mietlicki 2018), a new network combining four microphones and two optical systems with the goal of identifying the sound source location. Its computational load is high, and therefore it cannot be resolved by most of the low-cost acoustic sensor systems.

III.2.2 Ad Hoc Developed Acoustic Sensor Networks

Other projects have the goal of developing a custom WASN in order to meet the requirements of specific applications, mainly of particular analysis over the acoustic data. The IDEA project (Intelligent Distributed Environmental Assessment) (Botteldooren et al. 2011) seeks to analyze air and noise pollutants in several urban areas of Belgium. It integrates a sensor network based on a cloud platform, and it measures noise and air quality (Domínguez et al. 2014). The CENSE project, which stands for characterization of urban sound environments, is committed to conceiving noise maps in France (*Cense - Characterization of urban sound environments* n.d.), integrating both simulated and measured data collected from a cost-affordable WASN. The MESSAGE project, which stands for Mobile Environmental Sensing System Across Grid Environments, (Bell and Galatioto 2013) not only monitors noise, carbon monoxide, nitrogen dioxide, and temperature, but also goes further and gathers real-time humidity and traffic occupancy in the United Kingdom. Moreover, the MONZA project (Bartalucci et al. 2018; Bartalucci et al. 2020) follows both the idea of monitoring urban noise real-time together with other air pollutants in the Italian city of Monza.

A more recent approach when working with WASN and noise sources is the hybrid approach of combining the acoustic information with subjective perception surveys that are specially focused on the typology of events affecting everyday life activities, such as sleeping or studying (De Coensel and Botteldooren 2014). A noise identification system is applied to provide information about the detected sounds and establish a relationship between the perception surveys and the identified events related to road traffic noise (Brown and Coensel 2018).

One of the projects that faces the challenge of urban sound classification is Sounds of New York City Project (SONYC), which monitors the city using a low-cost static acoustic sensor network (Mydlarz et al. 2017). The goal of this project is to monitor noise pollution in real time by identifying the different noise sources that populate an acoustic environment. In this regard, it uses acoustic event detection (Bello et al. 2019; A. Cramer et al. 2021) over all the collected (and annotated) urban acoustic data (Cartwright et al. 2020).

Another project with a similar conceptual background is the DYNAMAP project (Sevillano

et al. 2016), which deployed two pilot areas in Italy, located in Rome (Bellucci et al. 2017) and Milan (Zambon et al. 2017), with the idea of computing and comparing the noise impact of road infrastructures in suburban and urban areas, respectively. The two WASNs monitored road traffic noise by reliably collecting data at a frequency of 44,100 Hz, managing to remove specific non-traffic audio events (Socoró et al. 2017; Alsina-Pagès et al. 2018) in order to build a more accurate road traffic noise map (Bellucci and Cruciani 2016).

Deep learning has been applied to urban audio datasets, obtaining encouraging results (Vidaña-Vila et al. 2020c; Gontier et al. 2021). However, many research studies are limited to datasets that are unrealistic because they are curated from audio libraries rather than real-world urban monitoring, and/or are single-label annotated, neglecting the simultaneous occurrence of sounds (Salamon et al. 2014). Recent work suggests that using multilabel data can enable practitioners to obtain more realistic results (Gontier et al. 2021). In this regard, edge intelligence is envisaged as a powerful alternative to address the typical computation overhead associated with multilabel classification systems (Srivastava et al. 2021).

III.2.3 Sensor Deployment Strategies

In addition to the sensors, an important design parameter for wireless (acoustic) sensor networks is the physical topology in which sensing units are deployed in a specific scenario. This section reviews the impact of the sensor deployment strategy on (1) the maximum size of the area of interest to be covered, (2) the power consumption of each node, and (3) the communication robustness.

As far as the area of interest is concerned, in (Biagioni and Sasaki 2003), the authors study different node placements for a wireless sensor network able to sense environmental parameters (e.g., sunlight, temperature, humidity, rainfall, or images) that are delivered to different base stations by means of ad hoc wireless communication links. Concretely, the authors propose an analytical model to come up with the optimal position of nodes according to the desired node arrangement (e.g., ring, star, triangle, square, and hexagon). Alternative sensor location strategies have been studied for Underwater Acoustic Sensor Networks (UASN) applications as well. For instance, in (Han et al. 2014), the authors study and compare the impacts of node deployment strategies in a 3-D environment. Their results show that a regular tetrahedron deployment scheme outperforms other topologies such as a random or cube topology. Concretely, the metrics that they use to compare the different schemes are the reduction of localization error and the optimization of localization ratio while maintaining the average number of neighbouring anchor nodes and network connectivity. Similarly, in (Murad et al. 2015), the effects of deploying UASNs together with the most well-known research projects in this field are reviewed.

Another way of extending the area size consists of using mobile nodes (e.g., robotic vehicles). For instance, in (Kim et al. 2016), the authors consider a dynamic topology in which nodes are constantly moving and study the best way to optimize power consumption.

Finally, in (Ding et al. 2017), the authors propose an advanced strategy for sensor placement that aims to maximize the connectivity robustness of the nodes for sparse networks. Concretely,

they explore an analytical topology composed of hexagonal clusters and develop an algorithm for geometric distance optimization to improve the overall robustness of the system.

III.3 Collection and Annotation of a Real-World Dataset

To evaluate the results of a multilabel classifier, the first step is to have available a dataset with multilabel data. This section (1) describes the procedure that we followed to collect these data from a real-world environment, (2) details how data were labeled, and (3) exhibits the number of events for each class that were identified in the dataset.

III.3.1 Recording Campaign

In order to obtain a suitable real-world dataset to validate the proposed approach (i.e., multilabel classification of urban sounds taking advantage of physical redundancy in sensor nodes), two recording campaigns were conducted in the metropolitan area of the center of Barcelona (Spain). To have a wider variety of data, the two recording campaigns took place in different seasons of the year. The first one was conducted during autumn 2020 (17 November 2020) and the second one was conducted during spring 2021 (31 May 2021). Another substantial difference is that the first recording campaign was conducted under mobility restrictions (Bonet-Solà et al. 2021) due to the COVID-19 pandemic, while during the second recording campaign, those restrictions were significantly softer. 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 location where the recording campaign was conducted is a specific crossroad in the Barcelona city center: the crossroad between Villarroel Street and Diputació Street (plus code 95M5+H9). This place is located in the Eixample area of Barcelona, which is the wide expansion district of the city. This place was chosen in order to validate the architecture proposed in (Vidaña-Vila et al. 2020c), as its shape follows strictly regular symmetry. From now on, these recordings will be referred to as Eixample Dataset.

Four Zoom H5 recorders (*H5 Handy Recorder - Operation Manual* 2014) (see Figure III.1) were used to record data, with one placed on the middle of each corner of the street intersection. Concretely, the devices stood over tripods at a distance of at least 4 m from the closest wall and 1.5 m from the floor to avoid undesired sound reflections. Furthermore, the inclination of the device with respect to the floor was 45°. This will enable us to have simultaneous audio recordings in order to assess with real-world data whether physical redundancy helps increase the robustness of the classification results of the end-to-end architecture proposed in (Vidaña-Vila et al. 2020c), as in that work we used synthetically generated audio files.

The two recording campaigns resulted in about 2 h and 30 min of acoustic data per sensor per campaign. Due to technical problems with the batteries of the recorders during the second campaign, the files were fragmented into two audio files. The time dedicated to changing the batteries was of about 5 min, in which we were not able to record data.

Figure III.1: Recording campaign and Zoom recorder.

III.3.2 Data Labeling

Existing approaches to automatically label acoustic data (Fonseca et al. 2019) inspired by semi-supervised learning techniques shall not provide the necessary high level of accuracy and precision to train and test a reliable model to be considered as ground truth in order to assess our proposed approach. Therefore, data collected from the recording campaign have to undergo the manual labeling process described below.

Our previous experiences on (manually) labeling real-world acoustic datasets (Vidaña-Vila et al. 2020a; Vidaña-Vila et al. 2020b; Vidaña-Vila et al. 2017) taught us that assigning a tag to an acoustic sample is a time-consuming process: contrary to other types of datasets (e.g., images) in which a label can be assigned as soon as the sample is shown, in acoustic data labeling one has to wait for the whole acoustic record to be reproduced before assigning it a label. Typically, this is done with off-the-shelf software alternatives such as Audacity (Audacity 2014) that provide end-users with a spectrogram of the full audio record; thus, it becomes easier to visually identify those time frames in which something anomalous (i.e., potential events of interest) might be happening. However, as the purpose of this work is to identify multiple events (i.e., classes) that occur concurrently not only in the foreground but also in the background, and thus potentially overlap in a given acoustic sample, all of the collected acoustic samples—coming from the aforementioned two 2.5 h length campaigns—must be systematically heard and labeled.

In this situation, off-the-shelf software alternatives come at little ease due to the fact that they require, from the user point of view, several sequential and time-consuming interactions with the mouse (e.g., dragging and selecting the desired part of the spectrogram, clicking to add the label) and keyboard (e.g., typing the labels for the selected area). Additionally, as far as keyboard interactions are concerned, we have found that it is very common to make typos when writing the labels (e.g., typing *rnt* instead of *rtn*), which often require an additional review stage before feeding the labels to the machine learning system. Obviously, all these interactions add a significant delay on the overall data labeling process.

To address these shortcomings, we decided to develop a simple, yet custom, python script aimed to ease the manual acoustic data labeling process. The behavior of the script is described in the following points:

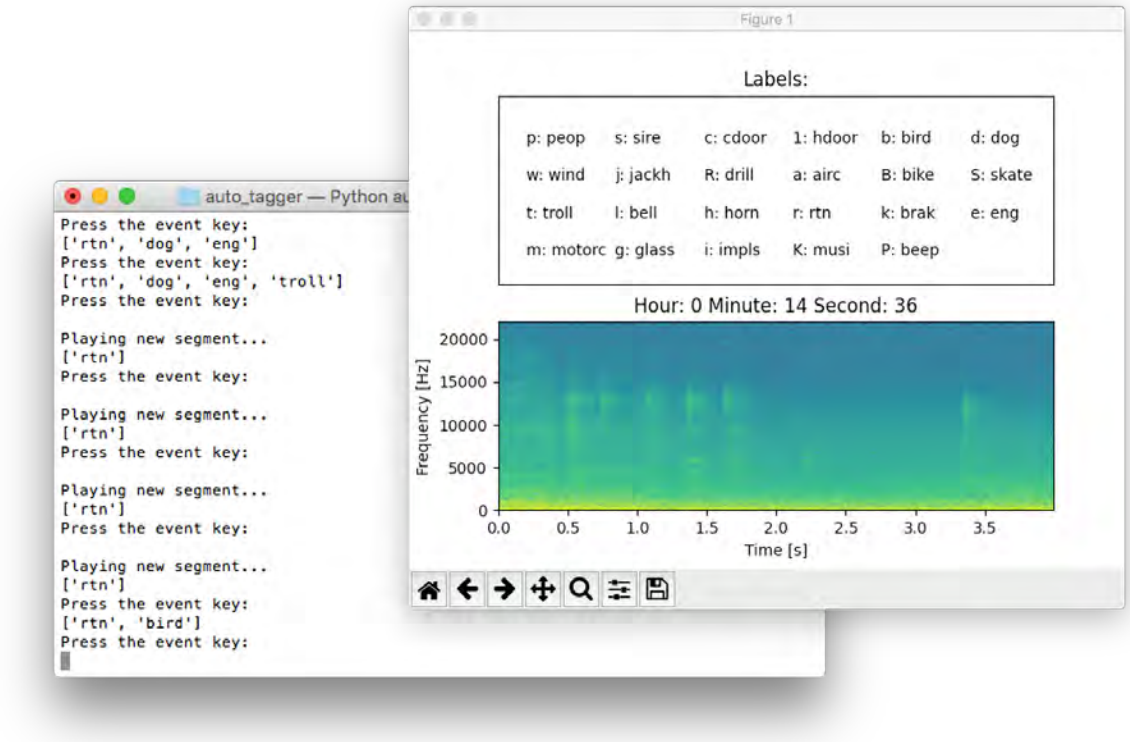


Figure III.2: Screenshot of the developed python script. The screen on the background (**left**) records the keystrokes. The screen on the foreground (**right**) shows the information of the current window and a legend with the correspondences between keys and labels.

- Input. The script reads the `.wav` files coming raw from the Zoom H5 recorders. This done with the module `AudioSegment` of the `pydub` library. This module loads the whole audio is input into a vector, which results in a very convenient solution when windowing it.

- Configuration file. Moreover, the script reads a configuration JSON file specifying (1) the window size in which the `.wav` file will be split, (2) all the possible labels that may appear in the recording, and (3) a key (one letter long) associated with each possible label.

- User interaction. As shown in Figure III.2, the script (1) displays a screen with the spectrogram—using the `pyplot` module of the `matplotlib` library—of the current window together with its start and finish times, (2) continuously reproduces the audio associated to the current window using the `pyaudio` library, and (3) shows the possible labels together with their associated keys in another screen. Then, each time the user presses a key corresponding to a label, the label is aggregated to the vector of labels associated with the current window. If the same key was pressed again, that event would be removed from the vector. Furthermore, the user can go to the following or previous acoustic window by using the arrow keys. Note that in this way, the user has a single interaction device (i.e., keyboard) and typos in labels are not possible.

- Output. The script writes a `.csv` file with (1) the start time of the window, (2) the finish time of the window, and (3) all the tags that have been selected for that window. For instance, a line in this `.csv` file would appear as follows:

```
276.000000 280.000000 bike+dog+troll
280.000000 284.000000 bike+glass
284.000000 288.000000 dog+peop+drill
```

  As a result, each line of the labels file derived from a recording contains the starting and ending time of the window and the different labels assigned to (i.e., appearing) that fragment.

Thanks to this software, we experienced that roughly, on average, the labeling process took us 30% less time than what it took in other works where we used off-the-shelf alternatives.

As far as the data labeling process is concerned, we labeled the acoustic data (i.e., two 2.5 h length recordings) from one of the four corners of the recording campaign. Concretely, it took about 12 h to annotate all these acoustic data using the aforementioned method. These data were then used as a reliable ground truth for the experimental evaluation. The other audio files were not manually labeled as they were only used as a complement to the selected sensor to check if the accuracy improves when joining together the classification results from neighboring nodes. In order to use a classification algorithm based on a deep neural network able to classify the spectral information of acoustic data (Vidaña-Vila et al. 2020c), we decided to directly label the audio files in windows of 4 seconds to keep compatibility with previous experiments (Vidaña-Vila et al. 2020c). Hence, as it can be seen in the spectrogram depicted in Figure III.2, the script sequentially split audio files in windows of 4-second length.

III.3.3 Obtained Dataset

The manual labeling task led the team to this taxonomy, with the number of classes and the number of labeled events shown in Table III.1.

III.3.4 Train/Validation/Test Split

As can be seen in the last column of Table III.1, the dataset is highly imbalanced: whereas the top events of the table are present in both recording campaigns with a considerable number of samples, there are some sounds that are poorly represented in the dataset. Actually, there are acoustic events that are only present in one of the two recording campaigns. For example, the *drill* sound is present only in the second campaign. Moreover, as the events were labeled in 4-second windows, the fact that there are 14 events labeled with the *drill* tag does not mean that there are 14 independent drilling sounds, as a long drilling sound lasting (for example) 8 seconds would be counted as two different windows. This phenomenon would happen with all the acoustic events that last more than the 4-second window (presumably, categories such as *sire*, *musi*, *eng*, or *motorc* among others). We are took this into account when splitting the dataset in Train/Validation and Test sets (see Section III.3.4).

Table III.1: Number of events annotated on the dataset.

Label	Description	Number of Occurrences		
		1st Campaign	2nd Campaign	Total
rtn	Background traffic noise	2177	2118	4295
peop	Sounds or noises 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 labeler 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 in 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

III.4 Two-Stage Multilabel Classifier

After obtaining the labeled dataset, this section details the whole classification procedure. First, it discusses the feature extraction process of the acoustic data. Next, it shows how the dataset is split into Train, Validation, and Test sets. Then, it describes how the problem of class imbalance has been addressed by using data augmentation. Finally, it details the two-stage classification process.

III.4.1 Feature Extraction

As features, and to maintain compatibility with (Vidaña-Vila et al. 2020c), 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 labeled events, we realized that the *brak* event 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 generated with a Fast Fourier Transform (FFT) (Cooley and Tukey 1965) window of 1024 points and using the `librosa` python library (McFee et al. 2015). 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.

The audio files obtained on the recording campaign had to be divided into Train, Validation, and 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 moments of the day, so that one single event is not split into different groups. 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 sets.

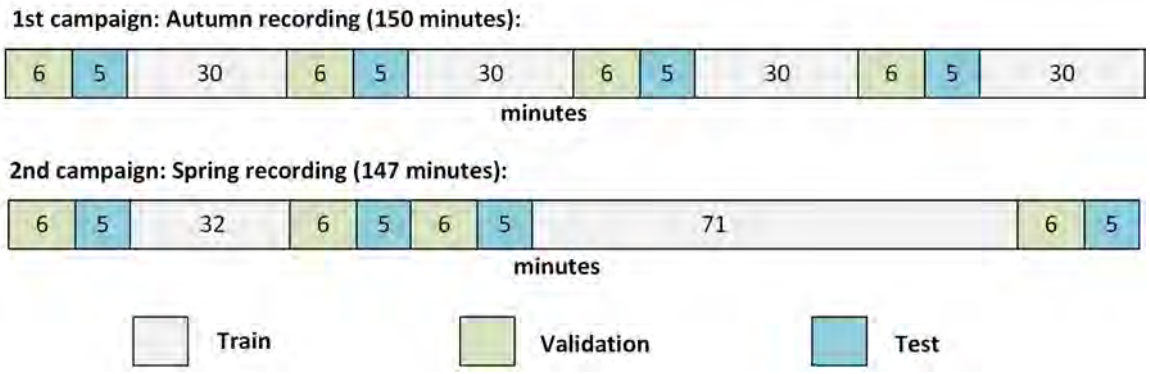


Figure III.3: Duration and temporal splitting of the Train, Validation, and Test sets of the dataset.

Concretely, the division was done as shown in Figure III.3: with divisions into contiguous regions ranging from 5 to 71 min length.

This division left the dataset with 209 min for Training, 40 min for Validating, and 48 min 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 can be appreciated in Table III.2, the three sets 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 properly. We discarded the option of splitting the 14 events into the Train and Test sets as they belonged to the same drilling machine recorded in the same location, which may have generated biased results. Moreover, we decided to remove the *cmplx* sounds from the dataset. As we could not identify the specific source of those sounds when labeling them, we arrived at the conclusion that they may confuse the system.

III.4.2 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 were obtained from the BCNDataset (Vidaña-Vila et al. 2020a), which is a dataset containing real-word urban and leisure events recorded at night in Barcelona. As the BCNDataset was labeled differently than the Eixample Dataset, labels from both datasets were unified.

Table III.2: Number of events on the Train, Validation, and Test 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

More concretely, per each of the acoustic events, on the BCNDataset the labels are provided as:

```
start_second end_second label
```

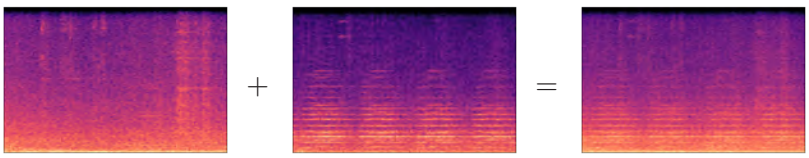


Figure III.4: Example of mixup data augmentation using two random 4-second fragments containing several acoustic events.

Note that in BCNDataset, the difference between the starting time and the ending time of each acoustic label is variable (not as in Eixample Dataset, where the ending time is always 4 s later than the starting time), and only one label is provided per each row of the text file. However, the format of the file is the same as the one presented in Section III.3.2, which eased the merging process of both datasets. To merge both datasets, the labels of the BCNDataset 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 of audio mixing, sometimes known as mixup (Stowell et al. 2019). As shown in Figure III.4, 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.

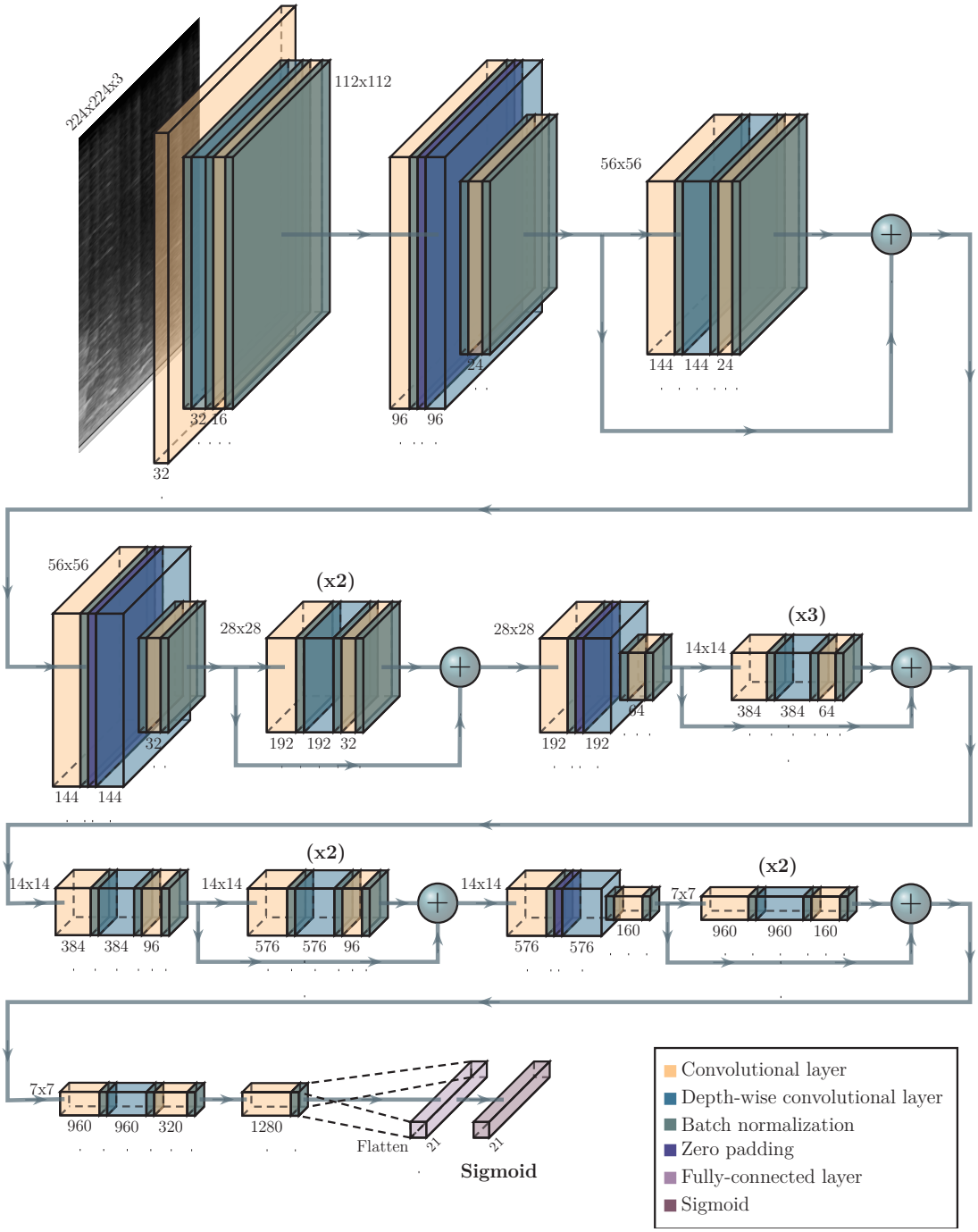


Figure III.5: Architecture of the MobileNet v2 (Howard et al. 2017) deep neural network used at the first stage of the classification process.

III.4.3 Multilabel Classification

The classification process consists of two layers:

1. The first layer (Section III.4.3.1) is a Deep Neural Network (DNN) that classifies 4-second fragments in a single node.

2. The second layer (Section III.4.3.2) aggregates the classification results of the deep neural networks running on the four corners of the intersection and makes a final decision on what events are actually happening on each corner by means of an ensemble of classifiers.

III.4.3.1 First Stage: Classification in One Node

The classification of the events on each of the nodes was carried out using a deep neural network with a MobileNet v2 architecture (Howard et al. 2017) with a size of 8.8 MB—which should fit on a low-cost computing node for a WASN. As shown in Figure III.5, the last layer of the neural network was replaced by a fully connected layer with one neuron per class and a Sigmoid activation function on each of them to allow multilabel classification. As a result, for each input datum, the output neurons showed the probability of that class being present on the input spectrogram. Once the probabilities were obtained, to evaluate whether the deep neural network was able to classify correctly without taking into account the decisions made by neighboring nodes, custom thresholds for each class were applied to determine if the event was actually present on the 4-second fragment. The thresholds were obtained by maximizing the F1-measure of each class on the validation set. As hyperparameters, an ADAM optimizer (Kingma and Ba 2017) was used with a learning rate of 1×10^{-4} and a weight decay regularization of 1×10^{-5}.

III.4.3.2 Second Stage: Classification Using Physical Redundancy

The aim of the second classification stage is to increase the robustness of the classification conducted at the previous stage by exploiting the physical redundancy of the nodes (i.e., nodes are physically deployed in such a way that the same event can be listened to by more than one node). Robustness in this context refers to the ability of the classifier to perform correctly when the output probabilities of the deep neural network for a given class are low but the event is actually happening. In this regard, our proposed system takes into consideration the classification results of neighboring nodes in order to strengthen (or weaken) its own results. For instance, if in the same frame Node A classified a *bell* with probability 0.3 and nodes B, C, and D classified *bell* with probability 0.8, then Node A should infer that a *bell* event actually happened. As manually defining these thresholds (or rules) might disregard some of the internal dynamics of the system, we propose to use a classifier to automatically generate them.

The process followed to train the automatic second stage classifier is detailed in what follows:

1. Once the deep neural network was trained, we used it to obtain a 21-component classification vector per each of the 4-second fragments of the original Eixample Dataset (see Section III.3). Each component of the vector indicated the likelihood of an acoustic event being present on the fragment. The labels from the dataset associated with each fragment were kept as ground truth.

2. The previous stage was done with the simultaneous audio of the remaining three neighboring locations. Therefore, for each 4-second fragment of the Eixample Dataset, we obtained four 21-component vectors together with the ground-truth labels.

3. The four vectors were concatenated horizontally, thus obtaining a single 84-component vector.

4. The 84-component vector and ground truth labels were used to fit a machine learning model that would output the final classification results.

For more clarification, this procedure is illustrated in Figure III.6.

III.5 Experimental Evaluation

To assess the classification performance of the proposed system, each one of both classification stages was evaluated.

III.5.1 Classification Performance at the First Stage

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

- **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 the data augmentation techniques detailed in Section III.4.2 to have around 500 samples for each class.

- **Experiment 2**: We used the same data as in Experiment 1 and we also added data from the UrbanSound 8K dataset (Salamon et al. 2014). 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., 44,100 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 were only assigned between a random 10% to 30% on the final weight of the spectrogram.

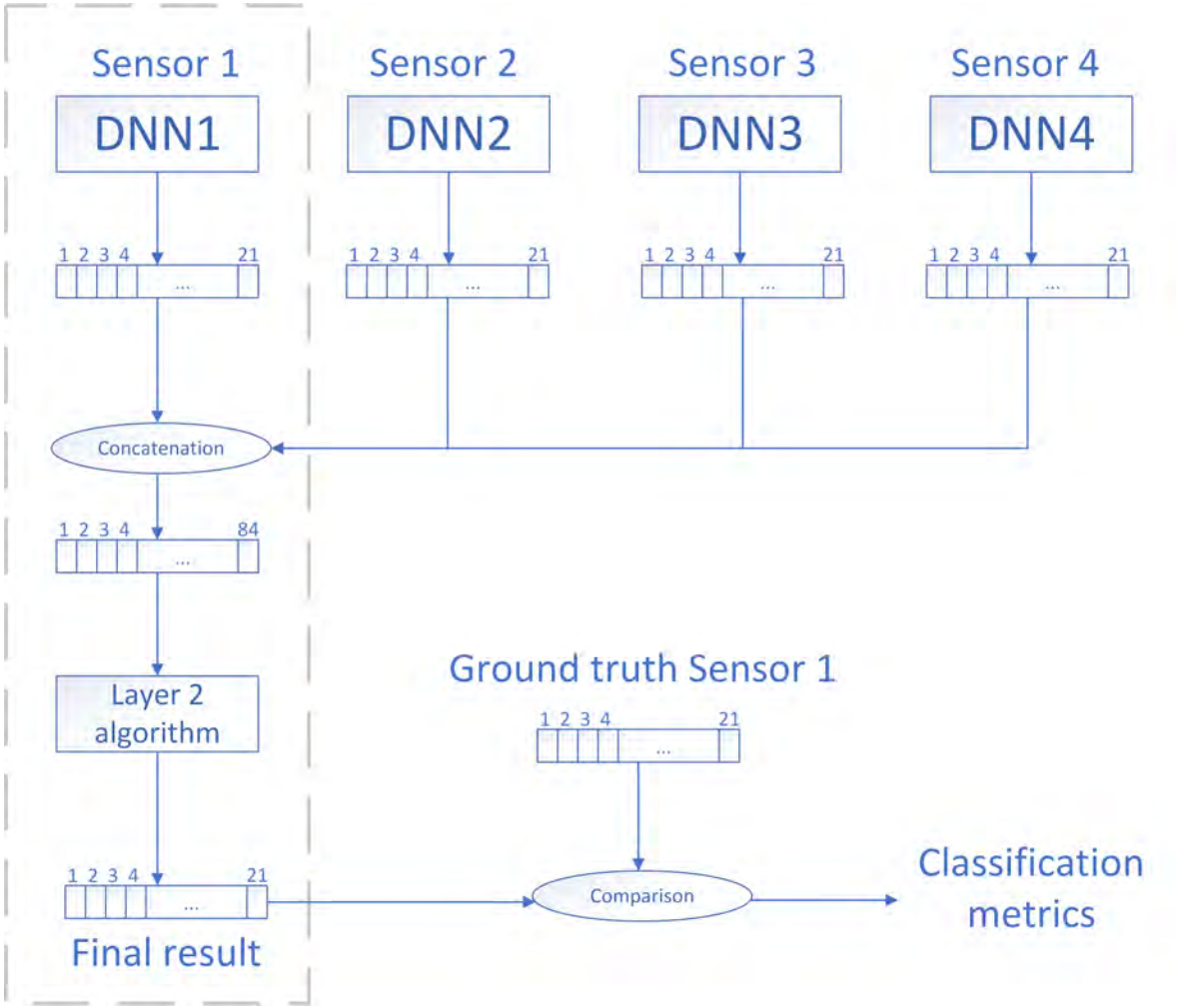


Figure III.6: Proposed system architecture with two classification stages. The first deep neural network of the first level outputs a 21-component vector that is later concatenated with the vectors from neighboring nodes. The resulting 84-component vector is examined by the second classification stage to obtain the final classification result. This scheme is replicated on each of the sensors of the system.

- **Experiment 3**: We used the same data as in Experiment 2, but on this occasion, each audio file from the UrbanSound dataset was used 10 times to combine it with a different audio file randomly selected from the BCNDataset or the Eixample dataset. This way, we increased the size of the Training data.

Table III.3: Macro and micro average F-1 scores for the experimental evaluation obtained at the first classification stage.

Dataset Used	F1- Macro Average	F1- Micro Average
Experiment 0	12%	46%
Experiment 1	39%	70%
Experiment 2	36%	75%
Experiment 3	33%	67%

The metrics that we used to compare the results are the Macro and Micro average F1-scores (Mesaros et al. 2016). 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 because, 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); 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 III.3 shows the classification results for each of the experiments. To compute the classification metrics, the *drilling* class was not taken into consideration as there are no events from that class on the Test set.

As can be seen in Table III.3, using the imbalanced data from the BCNDataset and the Eixample dataset without using any data augmentation techniques (i.e., Experiment 0) results in poor classification results. Concretely, the 12% on the Macro average F1 score tells us that the algorithm has problems in classifying most of the categories. In addition, having a Micro average F1 score higher than the Macro average F1 score tells us that the system performs better when classifying those categories with more samples than when classifying those categories with few instances. This phenomenon can be appreciated in all the experiments.

Generally speaking, as shown in Table III.3, the data augmentation techniques that have been used in this work (i.e., Experiments 1, 2, and 3) have helped build a more robust system. However, we think that using the UrbanSound samples has not actually helped to improve the performance of the overall system at all due to the following reasons. First, even if UrbanSound is a balanced dataset, it has fewer categories than the Eixample dataset. In addition, the difference between the original sampling rate of the Eixample dataset or the BCNDataset and the audio files from the UrbanSound dataset resulted in less realistic audio files than if we used two real-world datasets recorded with similar conditions (as we did in Experiment 1). Actually, comparing the classification metrics from Experiment 2 and Experiment 3, we can see that Experiment 2 has a better performance. We think that this is because when doing data augmentation, only a random 10% to 30% of data belong to the UrbanSound dataset, which means that most of the information belongs to the spectrograms from the other two datasets. As we are augmenting data 10 times using the same base spectrograms, the deterioration of the classification results may indicate that we are biasing the deep neural network with mild overfitting towards these base spectrograms when training.

To sum up, we propose that the data used in Experiment 1 offer the fairest trade-off between the performance of the system on large and small classes. Hence, from now on, for the experiments performed on Stage 2, we will use the model trained with Experiment 1 data.

Table III.4: Experiment results obtained at the second classification stage.

Algorithm Used	Micro Precision	Micro Recall	Micro F1	Macro F1
DT	71.6%	69.5%	70.5%	30.6 %
RF	81.8%	68.1%	74.3%	26.7 %
LR	77.3%	72%	74.6 %	37.8%
XGB	78.5%	70.2 %	74.1 %	39.3 %

III.5.2 Classification Performance at the Second Stage

As in this work the only labeled data that we had available were the ones recorded on one specific sensor, the experiments were conducted over that reference sensor. To discover the most suitable machine learning algorithm for the second classification stage, four different classification algorithms were evaluated:

1. Decision Tree (DT): The size of the model after training was 617 KB.
2. Random Forest (RF): The size of the model after training was 121 MB.
3. Logistic Regressor (LR): The size of the model after training was 20 KB.
4. XGBoost (XGB): The size of the model after training was 2.3 MB.

It is worth noting that the lighter classification algorithms from a computing point of view (i.e., the ones that require less RAM) are the DT and the LR, followed by XGB and, finally, the RF. In this case, to build the models, the only data that we could use were the ones belonging to the Eixample dataset, as this is the only one that has four simultaneous recordings. The algorithms resulted in the classification results shown in Table III.4. As classification metrics, apart from the metrics shown in Section III.4.3.1 (i.e., Micro F1 average and Macro F1 average), the Micro precision and Micro recall of the system are shown as well (Mesaros et al. 2016).

As can be seen in Table III.4, all the algorithms tend to have slightly higher values of Micro precision than Micro recall, which are emphasized in some of the classifiers (i.e., RF and XGB). Whereas the first metric illustrates what proportion of detected events were actually correct, the second one shows what proportion of actual events were correctly classified. The Macro F1 measure gives the same importance to both metrics.

Whereas the highest Micro precision result is achieved by using the RF algorithm (81.8%), the highest Micro recall result is obtained using the LR (72%). However, we believe that for the current context of this work (i.e., classification of urban sounds), we should also consider the F1 scores. In this sense, for the Micro F1 score, three classification algorithms present similar results (the RF, LR, and XGB with 74.3%, 74.6%, and 74.1%, respectively). However, when checking the Macro F1 average, XGB outperforms the other classification algorithms, obtaining a final score of 39.3%. Therefore, we believe that the algorithm that presents the fairest trade-off between all the classification metrics is XGB.

When comparing the classification results obtained at the first stage to the classification results obtained at the second stage, we can see that using physical redundancy allowed

for increasing the F1 Macro average from 70% to 74.1% (+4.1%). Regarding the F1 Micro average, the results change from 39% to 39.3% (+0.3%). These increments suggest that the second stage helps to improve classification results mainly on the classes that have more instances.

As this system is to be deployed in a low-cost device such as the one presented in (Vidaña-Vila et al. 2020c), it is not only the accuracy that matters but also the capability of the system in making real-time classifications within the 4-second selected window. Moreover, to make the classification process smoother, it would be desirable to use a sliding 4-second window with hops as small as possible (i.e., obtaining as many classification results as possible by sliding the 4-second window with overlap). The amount of overlap that can be used in the system depends on the classification speed of the system to output new data.

For this reason, to check the amount of time that it would take to the system to output a new classification result, experimental tests were carried out using three different computation units (i.e., Raspberry Pi Model 2B, Raspberry Pi Model 3B+, and Raspberry Pi Model 4) and a plug-and-play USB microphone.

The main hardware differences among these three models relevant to the research presented in this work are their computation capabilities (central processing unit and operating frequency) and their amount of RAM memory:

- Raspberry Pi Model 2B: Broadcom BCM2836 SoC (ARMv7), Quad-core ARM Cortex-A7, @ 900 MHz, 1GB LPDDR2 of RAM.
- Raspberry Pi model 3B+: Broadcom BCM2837B0 SoC (ARMv8), Cortex-A53, 64-bit @ 1.4GHz, 1GB LPDDR2 SDRAM.
- Raspberry Pi model 4: Broadcom BCM2711 SoC (ARMv8), Quad-core Cortex-A72 64-bit @ 1.5GHz, 4GB LPDDR4-3200 SDRAM.

For each experiment, we evaluated the timing performance of the processing units by making 100 test runs on each device. The obtained results can be seen in Table III.5. The times on the table start counting since a 4-second fragment is acquired by the microphone, and they include (1) the spectrogram computation, (2) the first stage classification (DNN), and (3) the second stage classification. As can be observed in the table, the device in which the experiments are conducted greatly affects the timing results.

Even though all the Raspberry Pi models are able to obtain a classification result within 4 seconds and would hence be suitable for a real-world deployment of the system, Raspberry Pi Model 2B offers a timing response that is at least about 1 s slower than its superior models. It can also be observed that Raspberry Pi Model 4B is, in general, about 0.5 s faster than Raspberry Pi Model 3B+. Concretely, when using Raspberry Pi Model 4B, the average response time of the system to perform a complete classification ranges from 0.66 s (when using the DNN + DT) to 0.78 s (when using the DNN + RF). Concretely, when using the aforementioned DNN + XGB algorithm, the classification would take on average 0.77 s. In this case, the system could use a 4-second length sliding window and a hop of 1 s (i.e., maximum

Table III.5: Time that it takes for the system to classify a 4-second audio fragment using three different sensor models. Results are shown in seconds after 100 runs.

Algorithms	RPi Model	Max. Time (seconds)	Min. Time (seconds)	Avg. Time (seconds)
DNN + DT	Model 2B	2.3	2.0	2.2
DNN + RF		2.9	2.4	2.6
DNN + LR		2.4	2.0	2.2
DNN + XGB		2.8	2.4	2.5
DNN + DT	Model 3B+	1.3	0.9	1.1
DNN + RF		1.5	1.2	1.3
DNN + LR		1.3	1.1	1.2
DNN + XGB		1.4	1.3	1.5
DNN + DT	Model 4B	0.7	0.6	0.6
DNN + RF		0.8	0.7	0.7
DNN + LR		0.8	0.6	0.6
DNN + XGB		1.0	0.7	0.7

classification time for DNN + XGB in Model 4B) and thus output a classification result in the next second.

Table III.6: Evaluation metrics of the system when combining the outputs of 4 local nodes by using the XGBoost algorithm.

Label	True Negative	False Positive	False Negative	True Positive	F1-Score
rtn	0	37	11	672	0.97
peop	495	44	96	85	0.55
brak	513	58	80	69	0.50
bird	485	27	59	149	0.78
motorc	469	54	79	100	0.60
eng	502	39	41	138	0.78
cdoor	652	15	40	13	0.32
impls	598	31	61	30	0.39
troll	670	18	18	14	0.44
wind	709	3	5	3	0.43
horn	709	1	7	3	0.43
sire	701	13	5	1	0.10
musi	714	2	4	0	0
bike	707	1	12	0	0
hdoor	705	7	8	0	0
bell	707	0	6	7	0.70
glass	707	0	2	1	0.50
beep	711	0	9	0	0
dog	718	1	1	0	0

Finally, to observe with detail the classification results obtained when using the selected parameters, Table III.6 shows the individual classification metrics per each class of the dataset based on the results obtained in Experiment 1 on Section III.4.3.1 and using the XGBoost classifier. As 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 III.6).

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.

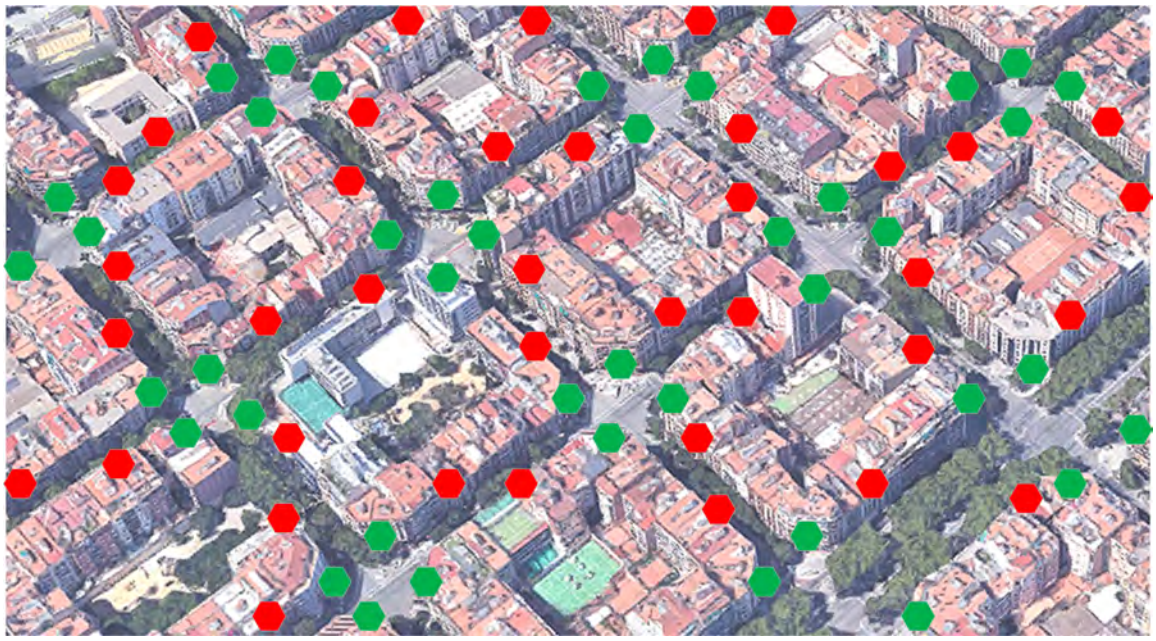


Figure III.7: Example of a possible future location of sensors. Green dots indicate the location used for the experiments conducted in this paper. Red dots indicate the new proposed locations.

III.6 Discussion

From our point of view, we believe that the obtained results in this research are encouraging in terms of covering the expected results. In fact, the proposed system has been shown to properly operate in a real-world environment. That is, the proposed system has been exposed to the real-operation conditions (in terms of audio) typically found in urban environments: appearance of sounds not previously recorded, various events happening simultaneously, etc. Using inexpensive commodity hardware (i.e., less than EUR 100 Raspberry Pi Model 4B), it has been able to produce classification outputs with reasonable accuracy in 1 s.

III.6.1 Location Perspective

The intrinsic Eixample topology makes the deployment of the sensors straightforward for this specific scenario. As all streets are totally symmetrical in this part of Barcelona, it is possible to deploy one sensor in each corner of the crossroads. If the proposed system were extended to the whole city, the symmetry for this low-cost sensor network would be still guaranteed for all places in the city center. However, up to now, we thought we should analyze, or at least

test, the results with other distributions, also taking advantage of the symmetry of the streets. As we have found that the most relevant sounds are detected in most of the four sensors in a crossroad, it might be interesting to discover what would happen if the location is slightly farther or if the sensor deployment strategy is different. In the latter case, we envisage a design trade-off between the advantages of physical redundancy in terms of accuracy, the cost of the WASN (i.e., number of sensors), power consumption, robustness, and size of the area under interest.

In Figure III.7, we show a possible future location deployment of the sensors with a wider distance between them, which despite reducing the effects of physical redundancy, may make the nodes more aware of what happens in the streets, instead of focusing on the crossroads. This could open further research on discovering the optimal distance between sensors according to the symmetry of the streets and balancing accuracy with the number of sensors to be deployed. A further step in this analysis would be to study the deployment of the proposed system in other parts of Barcelona without the Eixample symmetry: narrower streets, irregular crossroads, small squares, and other urban layouts that may make distribution of sensors difficult, in order to cover all the events happening on the street.

An accurate identification of sounds in an urban acoustic soundscape taking advantage of physical redundancy in the nodes could help to locate the sound source. While this might not be relevant in some cases/applications, it could be really helpful when assessing noise complaints (neighbors, dogs, etc.). In fact, this could help local authorities to identify those places (buildings, shops, bars, discos, etc.) where the noise is generated and conduct appropriate corrective measures. Additionally, this could provide crucial support to model the noise behavior in any city if the proposed system is deployed for long periods of time (i.e., months or years). In this case, it would be possible to discover recurrent patterns as certain types of noise would come always from the same places (e.g. ambulances, trucks, motorbikes, etc). Hence, a city noise model could be designed using the outputs of the proposed system.

III.6.2 Accuracy and Sample Availability

According to the conducted experiments, we have observed that the F1-Micro average score is consistently higher than the F1-Macro average. This means that the system has better performance on those classes that have a significant number of instances for train and test. For instance, while the *bird* class has 913 instances for train and 208 instances for test, obtaining an F1-score of 0.78, the *bike* class only has 55 instances for train and 12 instances for test, obtaining an F1-score of 0. To fight this situation, we believe that the individual detection may be improved by balancing and obtaining more data from recording campaigns in the same location or in other locations in Barcelona.

However, the number of instances is not the only relevant factor here: as it can be seen in Table III.6, the *peop* class has 954 instances for train and 181 instances for test, but only obtains an F1-score of 0.55. This drives us to think that the saliency of each event should be considered as well. In fact, we have observed that those events with low saliency are easily masked by other events occurring concurrently with higher saliency. This situation makes the

system obtain a higher number of false negatives than false positives for those specific events. Further experimentation with alternative features and/or distinguishing between foreground and background events at the annotation stage would be needed to validate this hypothesis.

III.7 Conclusions

In this work, progress has been made in the Training, Testing, and Validation of a two-stage classifier composed of a deep neural network and an XGBoost classifier with a very relevant focus on the use of real-world data. In our experiment, real-world data gathered at the city center of Barcelona have been used to validate the feasibility of a real-operation deployment of the algorithm. The data gathering process has been carried out in four simultaneous spots at a traffic intersection in order to assess up to what extent physical redundancy increases the robustness of the classifier. Furthermore, a new data labeling procedure aimed to reduce the amount of time spent on the task of manually labeling acoustic samples has been described. We have also shown which strategies we used to enrich the gathered data (i.e., data augmentation) to balance the corpus and thus improve the performance of the classifier.

From the experiments conducted, we can conclude that applying data augmentation techniques has helped the classifier to identify better those categories with few instances on the dataset. Moreover, physical redundancy of sensors has helped increasing the Micro and Macro F1-metrics. However, the improvement is mostly noticeable in those classes of the dataset that have more sample instances.

A real-world deployment of a WASN capable of detecting multiple acoustic events occurring simultaneously such as the one proposed in this paper would enable public administrations to have more information available about the types of sounds present in each area of the city in real time. This information may be helpful to assess neighbor complaints or detect the most acoustically polluted areas as well as to design policies to improve the quality of life of citizens of the more acoustically polluted areas.

As future work, we foresee that adding a memory layer to the system may increase the classifier performance (e.g. if there is a *siren* sound in a 4-second fragment, then it is likely that the next 4-second frame contains a *siren* sound as well). That is, we believe that knowing the probability of certain events in certain cases may help. Thus, this hypothesis will be further evaluated in future works. In addition, as it has been detected that the class imbalance of the dataset deteriorates the performance of the system on the poor classes, new training and testing data should be acquired. Finally, as the type of acoustic events present in urban environments are volatile, may vary day by day, and in some cases, only a few instances of each class might occur, it would be interesting to study the potential application of techniques that explicitly allow for new categories, such as few-shot learning or active learning.

Author's contributions

Conceptualization, D.S., R.M.A.-P., E.V.-V., and J.N.; methodology, E.V.-V., R.M.A.-P., D.S., and J.N.; software, E.V.-V., and J.N.; validation, D.S., J.N., and R.M.A.-P.; formal analysis,

E.V.-V.; investigation, E.V.-V; resources, J.N.; data curation, E.V.-V.; writing—original draft preparation, E.V.-V. and J.N.; writing—review and editing, J.N., E.V.-V., R.M.A.-P., and D.S.; visualization, E.V.-V. and J.N.; supervision, R.M.A.-P. and D.S.; project administration, R.M.A.-P.; funding acquisition, R.M.A.-P. All authors have read and agreed to the published version of the manuscript.

Funding

We 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. Additionally, we would like to thank La Salle Campus BCN - URL for partially funding the joint research with Tilburg University in the framework of Ms. Vidaña-Vila's PhD thesis.

Acknowledgements

We would like to thank Gerard Ginovart for his valuable assistance in the recording campaign in both seasons.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DNN	Deep Neural Network
DT	Decision Tree
EEA	European Environment Agency
EU	European Union
LR	Logistic Regressor
RF	Random Forest
UASN	Underwater Acoustic Sensor Networks
WASN	Wireless Acoustic Sensor Network
XGB	XGBoost
WHO	World Health Organization

References

Abbaspour, Majid, Karimi, Elham, Nassiri, Parvin, Monazzam, Mohammad Reza and Taghavi, Lobat (2015). 'Hierarchal assessment of noise pollution in urban areas–A case study'. In: *Transportation Research Part D: Transport and Environment* vol. 34, pp. 95–103.

Alsina-Pagès, Rosa Ma, Alías, Francesc, Socoró, Joan Claudi and Orga, Ferran (2018). 'Detection of Anomalous Noise Events on Low-Capacity Acoustic Nodes for Dynamic Road Traffic Noise Mapping within an Hybrid WASN'. In: *Sensors* vol. 18, no. 4, p. 1272.

Audacity, Team (2014). *Audacity*.

Bartalucci, Chiara, Borchi, Francesco and Carfagni, Monica (2020). 'Noise monitoring in Monza (Italy) during COVID-19 pandemic by means of the smart network of sensors developed in the LIFE MONZA project'. In: *Noise Mapping* vol. 7, no. 1, pp. 199–211.

Bartalucci, Chiara, Borchi, Francesco, Carfagni, Monica, Furferi, Rocco, Governi, Lapo, Lapini, Alessandro, Bellomini, Raffaella, Luzzi, Sergio and Nencini, Luca (2018). 'The smart noise monitoring system implemented in the frame of the Life MONZA project'. In: *Proceedings of the EuroNoise*, pp. 783–788.

Bell, Margaret Carol and Galatioto, Fabio (2013). 'Novel wireless pervasive sensor network to improve the understanding of noise in street canyons'. In: *Applied Acoustics* vol. 74, no. 1, pp. 169–180.

Bello, Juan P, Silva, Claudio, Nov, Oded, Dubois, R Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles and Doraiswamy, Harish (2019). 'Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution'. In: *Communications of the ACM* vol. 62, no. 2, pp. 68–77.

Bellucci, Patrizia and Cruciani, Francesca Romana (2016). 'Implementing the Dynamap system in the suburban area of Rome'. In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. 253 3. Institute of Noise Control Engineering, pp. 5518–5529.

Bellucci, Patrizia, Peruzzi, Laura and Zambon, Giovanni (2017). 'LIFE DYNAMAP project: The case study of Rome'. In: *Applied Acoustics* vol. 117, pp. 193–206.

Biagioni, Edoardo S and Sasaki, Galen (2003). 'Wireless sensor placement for reliable and efficient data collection'. In: *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 10–pp.

Bonet-Solà, Daniel, Martínez-Suquía, Carme, Alsina-Pagès, Rosa Ma and Bergadà, Pau (2021). 'The Soundscape of the COVID-19 Lockdown: Barcelona Noise Monitoring Network Case Study'. In: *International Journal of Environmental Research and Public Health* vol. 18, no. 11, p. 5799.

Botteldooren, Dick, De Coensel, Bert, Oldoni, Damiano, Van Renterghem, Timothy and Dauwe, Samuel (2011). 'Sound monitoring networks new style'. In: *Acoustics 2011: Breaking New Ground: Annual Conference of the Australian Acoustical Society*. Australian Acoustical Society, pp. 1–5.

Brown, AL and Coensel, B De (2018). 'A study of the performance of a generalized exceedance algorithm for detecting noise events caused by road traffic'. In: *Applied Acoustics* vol. 138, pp. 101–114.

Cartwright, Mark et al. (2020). 'SONYC-UST-V2: An Urban Sound Tagging Dataset with Spatiotemporal Context'. In: *arXiv preprint arXiv:2009.05188*.

Cense - Characterization of urban sound environments (n.d.). http://cense.ifsttar.fr/.

Cooley, James W and Tukey, John W (1965). 'An algorithm for the machine calculation of complex Fourier series'. In: *Mathematics of computation* vol. 19, no. 90, pp. 297–301.

Cramer, Aurora, Cartwright, Mark, Pishdadian, Fatemeh and Bello, Juan Pablo (2021). 'Weakly Supervised Source-Specific Sound Level Estimation in Noisy Soundscapes'. In: *arXiv preprint arXiv:2105.02911*.

Data and statistics (n.d.).

De Coensel, Bert and Botteldooren, Dick (Nov. 2014). 'Smart sound monitoring for sound event detection and characterization'. In: *Proceedings of the 43rd International Congress on Noise Control Engineering (Inter-Noise 2014)*. Melbourne, Australia, pp. 1–10.

Ding, Kai, Yousefizadeh, Homayoun and Jabbari, Faryar (2017). 'A robust advantaged node placement strategy for sparse network graphs'. In: *IEEE Transactions on Network Science and Engineering* vol. 5, no. 2, pp. 113–126.

Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 (n.d.).

Domínguez, Federico, Dauwe, Samuel, Cuong, Nguyen The, Cariolaro, Dimitri, Touhafi, Abdellah, Dhoedt, Bart, Botteldooren, Dick and Steenhaut, Kris (2014). 'Towards an environmental measurement cloud: Delivering pollution awareness to the public'. In: *International Journal of Distributed Sensor Networks* vol. 10, no. 3, p. 541360.

Fonseca, Eduardo, Plakal, Manoj, Font, Frederic, Ellis, Daniel PW and Serra, Xavier (2019). 'Audio tagging with noisy labels and minimal supervision'. In: *arXiv preprint arXiv:1906.02975*.

Gontier, Félix, Lostanlen, Vincent, Lagrange, Mathieu, Fortin, Nicolas, Lavandier, Catherine and Petiot, Jean-Francois (2021). 'Polyphonic training set synthesis improves self-supervised urban sound classification'. In: *The Journal of the Acoustical Society of America* vol. 149, no. 6, pp. 4309–4326.

Guski, Rainer, Schreckenberg, Dirk and Schuemer, Rudolf (2017). 'WHO environmental noise guidelines for the European region: A systematic review on environmental noise and annoyance'. In: *International journal of environmental research and public health* vol. 14, no. 12, p. 1539.

H5 Handy Recorder - Operation Manual (2014). Zoom Corporation.

Han, Guangjie, Zhang, Chenyu, Shu, Lei and Rodrigues, Joel JPC (2014). 'Impacts of deployment strategies on localization performance in underwater acoustic sensor networks'. In: *IEEE Transactions on Industrial Electronics* vol. 62, no. 3, pp. 1725–1733.

Howard, Andrew G, Zhu, Menglong, Chen, Bo, Kalenichenko, Dmitry, Wang, Weijun, Weyand, Tobias, Andreetto, Marco and Adam, Hartwig (2017). 'Mobilenets: Efficient convolutional neural networks for mobile vision applications'. In: *arXiv preprint arXiv:1704.04861*.

Hurtley, Charlotte (2009). *Night noise guidelines for Europe*. WHO Regional Office Europe.

Kim, Donghyun, Wang, Wei, Li, Deying, Lee, Joong-Lyul, Wu, Weili and Tokuta, Alade O (2016). 'A joint optimization of data ferry trajectories and communication powers of ground sensors for long-term environmental monitoring'. In: *Journal of Combinatorial Optimization* vol. 31, no. 4, pp. 1550–1568.

Kingma, Diederik P. and Ba, Jimmy (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

McFee, Brian, Raffel, Colin, Liang, Dawen, Ellis, Daniel PW, McVicar, Matt, Battenberg, Eric and Nieto, Oriol (2015). 'librosa: Audio and music signal analysis in python'. In: *Proceedings of the 14th python in science conference*. Vol. 8. Citeseer, pp. 18–25.

Mejvald, Pavel and Konopa, Ondrej (2019). 'Continuous acoustic monitoring of railroad network in the Czech Republic using smart city sensors'. In: *2019 International Council on Technologies of Environmental Protection (ICTEP)*. IEEE, pp. 181–186.

Mesaros, Annamaria, Heittola, Toni and Virtanen, Tuomas (2016). 'Metrics for polyphonic sound event detection'. In: *Applied Sciences* vol. 6, no. 6, p. 162.

Mietlicki, Christophe and Mietlicki, Fanny (2018). 'Medusa: a new approach for noise management and control in urban environment'. In: *Proceedings of the 11th European Congress and Exposition on Noise Control Engineering (Euronoise2018), Crete, Greece*, pp. 27–31.

Mietlicki, Fanny, Mietlicki, Christophe and Sineau, Matthieu (May 2015). 'An innovative approach for long-term environmental noise measurement: RUMEUR network'. In: *Proceedings of EuroNoise 2015*. Maastrich, Netherlands: EAA-NAG-ABAV, pp. 2309–2314.

Moudon, Anne Vernez (2009). 'Real noise from the urban environment: how ambient community noise affects health and what can be done about it'. In: *American journal of preventive medicine* vol. 37, no. 2, pp. 167–171.

Murad, Mohsin, Sheikh, Adil A, Manzoor, Muhammad Asif, Felemban, Emad and Qaisar, Saad (2015). 'A survey on current underwater acoustic sensor network applications'. In: *International Journal of Computer Theory and Engineering* vol. 7, no. 1, p. 51.

Mydlarz, Charlie, Salamon, Justin and Bello, Juan Pablo (2017). 'The implementation of low-cost urban acoustic monitoring devices'. In: *Applied Acoustics* vol. 117, pp. 207–218.

Noise (n.d.).

Organization, World Health et al. (2018). 'Environmental noise guidelines for the European region'. In: *World Health Organization. Regional Office for Europe*.

Organization, World Health et al. (2019). 'Environmental health inequalities in Europe: second assessment report'. In: *World Health Organization. Regional Office for Europe*.

Paulo, J, Fazenda, P, Oliveira, T, Carvalho, C and Félix, M (2015). 'Framework to monitor sound events in the city supported by the FIWARE platform'. In: *Proceedings of the 46o Congreso Español de Acústica*. Valencia, Spain, pp. 21–23.

Paulo, J, Fazenda, Pedro, Oliveira, Tiago and Casaleiro, Joao (June 2016). 'Continuos sound analysis in urban environments supported by FIWARE platform'. In: *Proceedings of the EuroRegio2016/TecniAcústica*. Porto, Portugal, pp. 1–10.

Polastre, Joseph, Szewczyk, Robert and Culler, David (2005). 'Telos: enabling ultra-low power wireless research'. In: *Proceedings of the 4th international symposium on Information processing in sensor networks*. IEEE Press, p. 48.

Salamon, Justin, Jacoby, Christopher and Bello, Juan Pablo (2014). 'A dataset and taxonomy for urban sound research'. In: *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044.

Santini, Silvia, Ostermaier, Benedikt and Vitaletti, Andrea (2008). 'First experiences using wireless sensor networks for noise pollution monitoring'. In: *Proceedings of the 2008 Workshop on Real-World Wireless Sensor Networks (REALWSN)*. ACM, pp. 61–65.

Santini, Silvia and Vitaletti, Andrea (2007). 'Wireless sensor networks for environmental noise monitoring'. In: *6. Fachgesprach Sensornetzwerke*, p. 98.

Sevillano, Xavier et al. (2016). 'DYNAMAP–Development of low cost sensors networks for real time noise mapping'. In: *Noise mapping* vol. 1, no. open-issue.

Socoró, Joan Claudi, Alías, Francesc and Alsina-Pagès, Rosa Ma (2017). 'An Anomalous Noise Events Detector for Dynamic Road Traffic Noise Mapping in Real-Life Urban and Suburban Environments'. In: *Sensors* vol. 17, no. 10, p. 2323.

Srivastava, Sangeeta, Roy, Dhrubojyoti, Cartwright, Mark, Bello, Juan P and Arora, Anish (2021). 'Specialized Embedding Approximation for Edge Intelligence: A Case Study in Urban Sound Classification'. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8378–8382.

Stowell, Dan, Petrusková, Tereza, Sálek, Martin and Linhart, Pavel (2019). 'Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions'. In: *Journal of the Royal Society Interface* vol. 16, no. 153, p. 20180940.

Test, Tsafnat, Canfi, Ayala, Eyal, Arnona, Shoam-Vardi, Ilana and Sheiner, Einat K (2011). 'The influence of hearing impairment on sleep quality among workers exposed to harmful noise'. In: *Sleep* vol. 34, no. 1, pp. 25–30.

Vidaña-Vila, Ester, Duboc, Leticia, Alsina-Pagès, Rosa Ma, Polls, Francesc and Vargas, Harold (2020a). 'BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset'. In: *Sustainability* vol. 12, no. 19, p. 8140.

Vidaña-Vila, Ester, Navarro, Joan and Alsina-Pagès, Rosa Ma (2017). 'Towards automatic bird detection: An annotated and segmented acoustic dataset of seven picidae species'. In: *Data* vol. 2, no. 2, p. 18.

Vidaña-Vila, Ester, Navarro, Joan, Alsina-Pagès, Rosa Ma and Ramírez, Álvaro (2020b). 'A two-stage approach to automatically detect and classify woodpecker (Fam. Picidae) sounds'. In: *Applied Acoustics* vol. 166, p. 107312.

Vidaña-Vila, Ester, Navarro, Joan, Borda-Fortuny, Cristina, Stowell, Dan and Alsina-Pagès, Rosa Ma (2020c). 'Low-cost distributed acoustic sensor network for real-time urban sound monitoring'. In: *Electronics* vol. 9, no. 12, p. 2119.

Vidaña-Vila, Ester, Stowell, Dan, Navarro, Joan and Alsina-Pagès, Rosa Ma (2021). 'Multilabel acoustic event classification for urban sound monitoring at a traffic intersection'. In: *Euronoise 2021*.

Wang, Cuiping, Chen, Guoqiang, Dong, Rencai and Wang, Haowei (2013). 'Traffic noise monitoring and simulation research in Xiamen City based on the Environmental Internet

of Things'. In: *International Journal of Sustainable Development & World Ecology* vol. 20, no. 3, pp. 248–253.

WHO (2011). *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*. World Health Organization. Regional Office for Europe.

Zambon, Giovanni, Benocci, Roberto, Bisceglie, Alessandro, Roman, H. Eduardo and Bellucci, Patrizia (2017). 'The LIFE DYNAMAP project: Towards a procedure for dynamic noise mapping in urban areas'. In: *Applied Acoustics* vol. 124, pp. 52–60.

Authors' addresses

Ester Vidaña-Vila GTM – Grup de Recerca en Tecnologies Mèdia, La Salle Campus Barcelona - Universitat Ramon Llull Quatre Camins, 30, 08022 Barcelona, Spain
ester.vidana@salle.url.edu

Capítol 4

Conclusions

4.1 Resum

Aquesta tesi ha investigat la classificació d'esdeveniments acústics en entorns urbans fent servir dispositius de baix cost. Concretament, s'han abordat dos reptes moderns. El primer repte ha estat el disseny d'una WASN escalable i de baix cost, creada a partir de sensors comercials programables i capacitats de computació limitades. A més, WASN és capaç de supervisar àrees a gran escala. El segon repte ha estat el desenvolupament d'un algorisme de classificació en temps real capaç de correr sobre els nodes de detecció que s'han dissenyat. D'aquesta manera, aquesta tesi ha completat el cicle d'un projecte típic de ML o DL juntament amb el disseny d'una topologia i arquitectura capaç de realitzar la classificació. Aquest cicle ha inclòs la selecció d'un escenari d'ús, l'estudi i l'anàlisi dels sons ambientals de la ubicació seleccionada, una definició de taxonomia per a tots els sons que s'han escoltat, una fase de prototipatge (amb dades en-línia) d'un algorisme de classificació juntament amb un estudi sobre quines són les característiques més convenients per parametritzar els senyals acústics, una millora d'aquest algorisme utilitzant dades del món real per donar suport a la polifonia (esdeveniments acústics múltiples que ocorren simultàniament), i la prova d'aquest algorisme sobre els sensors físics. A més, la tesi ha estudiat la possibilitat de desplegar els sensors amb redundància física per comprovar si això millora els resultats de classificació.

El propòsit d'aquest capítol és (1) resumir el treball analitzant si s'han complert els objectius de la tesi, (2) destacar les conclusions obtingudes en cadascuna de les etapes o processos de la dissertació i (3) proposa algunes línies de futur que es podrien considerar a la llum dels resultats obtinguts en aquesta dissertació.

4.2 Conclusions

En els últims anys, tant les organitzacions públiques com les privades han fet un esforç per controlar acústicament certs entorns urbans amb l'objectiu d'identificar les zones més contaminades d'una zona determinada. Aquest interès sorgeix a causa dels efectes secundaris que el soroll pot tenir sobre els éssers humans. No obstant això, alguns estudis revelen que no només el nivell de soroll és important per a la salut, sinó també el tipus de sons als quals estan exposats els ciutadans. Per aquesta raó, i veient que la majoria de les tecnologies actuals només permeten comprovar el nivell de soroll en una determinada ubicació en lloc d'identificar també la font de soroll (en la majoria dels casos, els tècnics han d'anar físicament al lloc on es genera el soroll per veure quina és la font que causa l'esdeveniment acústic), aquesta tesi s'ha dut a terme com un petit pas cap a la detecció automàtica de fonts sonores en entorns urbans utilitzant dispositius de baix cost i redundància física de sensors.

Una implementació en el món real del sistema proposat permetria monitorar quines són les àrees més contaminades de certs entorns urbans. A més, el sistema podria dir *quins* són els esdeveniments acústics que estan ocorrent en aquelles àrees i permetria tenir una visió global dels diferents *soundscapes* que hi ha a les diferents parts de la ciutat. Així doncs, el sistema es podria fer servir com una eina perquè les entitats públiques proposassin mesures per a mitigar situacions que són perjudicials per a la salut de la població (per exemple, redirigint el trànsit de diferents carrers o aplicant mesures restrictives relacionades amb el soroll).

Els dos reptes principals que s'han abordat en aquesta tesi són:

1. Dissenyar sensors de baix cost utilitzant maquinari comercial de manera que es puguin desplegar en àrees de monitorització de gran extensió. S'entén que un sensor de baix cost és un node amb un preu comercial inferior a 100€.

2. Desenvolupar un algorisme de classificació utilitzant tècniques DL que permeti realitzar la classificació automàtica d'esdeveniments acústics utilitzant els nodes de detecció dissenyats. A més, aquest algorisme de classificació ha de ser capaç de classificar esdeveniments que es produeixen simultàniament, ja que la polifonia és un fenòmen molt comú en entorns urbans.

A partir d'aquests dos reptes, es van definir tres qüestions de recerca i quatre objectius de tesi. La següent subsecció explica en primer lloc si s'han aconseguit els objectius de la tesi i, a continuació, la següent subsecció respon a les preguntes de recerca.

4.2.1 Acomplició dels objectius de la Tesi

Thesis Objective 1 o Objectiu de Tesi 1 (TO1): **Desenvolupar un sistema classificador automàtic capaç de detectar esdeveniments acústics en ambients urbans utilitzant informació espectral i temporal.**

Aquest primer objectiu de la tesi requeria fer ús d'informació espectral i temporal per a la classificació automàtica d'esdeveniments acústics. Per aconseguir aquest objectiu, es van analitzar diverses característiques acústiques per comprovar quines eren més convenients per a les dades acústiques donades. Al final, es van escollir com a característiques els espectrogrames, als quals se'ls va aplicar diversos càlculs i tècniques sobre com ara la normalització de l'espectrograma, l'estandardització, utilitzar espectrogrames log-mel o espectrogrames regulars o tècniques de processament com PCEN per comprovar quina representació s'adaptava millor a les dades a caracteritzar.

A més, es van analitzar diferents mides de finestra per comprovar quina era la mida més convenient per a les dades. Com que l'algorisme de classificació era una CNN, es va observar que la millor finestra era la que contenia prou informació per veure patrons dels esdeveniments acústics definits en la taxonomia. Per exemple, les sirenes tenen un clar patró en el temps que és útil per al classificador.

A més, aquest sistema automàtic havia de ser prou lleuger (en termes de requisits de càrrega computacional) per poder córrer sobre un sensor de baix cost. Per aquesta raó,

es van provar diferents arquitectures CNN. Al final, la arquitectura que es va seleccionar permetia fer la classificació en les unitats de computació proposades.

Podem concloure que, després dels experiments realitzats, aquest primer objectiu s'ha aconseguit amb èxit.

Thesis Objective 2 o Objectiu de Tesi 2 (TO2): **Dissenyar una plataforma de maquinari de baix cost capaç de classificar esdeveniments acústics en temps real.**

Aquest segon objectiu s'aconsegueix per mitjà de dissenyar i provar uns WASN de baix cost. Específicament, els nodes de detecció conténen una Raspberry Pi i un micròfon USB.

La Raspberry Pi va ser seleccionada com a plataforma de computació donat el seu balanç entre cost i característiques. A més, la plataforma té una àmplia comunitat de suport que pot ser útil per resoldre problemes. Es van provar diferents models de Raspberry Pi, però el model Raspberry Pi 4B va obtenir els millors resultats de classificació, sent capaç de completar el cicle de l'adquisició de dades, el seu processament i la classificació (incloent una DNN i un sistema intel · ligent ML que té en compte les sortides de diferents nodes veïns) en aproximadament 0,6 segons. Els altres models de Raspberry Pi avaluats (Model 2B i 3B+) van trigar uns 1,3 o 2,5 segons respectivament. També cal tenir en compte que els preus dels models de Raspberry Pi 2 i 3B+ són inferiors al preu del model de Raspberry Pi 4. Per tant, depenent dels requisits de diferents aplicacions particulars, els models Raspberry Pi 2 o 3B+ ja podien satisfer els adequades per certs problemes. El OS que s'executa als Raspberries és el Raspbian Lite.

En quant als micròfons, es va seleccionar el micròfon USB ja que no necessitava cap maquinari addicional per poder adquirir dades acústiques amb un Raspberry Pi. No s'ha dut a terme una anàlisi completa del micròfon seleccionat o una comparació entre diferents micròfons USB en aquesta dissertació, s'ha suposat que l'escollit té una resposta en freqüència aproximadament plana en les freqüències d'interès tal i com indica el seu fabricant. S'han descartat altres tipus de micròfon (com ara Micro Electro Mechanical System o Micro Electret Sistema Mecànic (MEMS)) ja que no tenen una resposta de freqüència plana o requereixen d'un convertidor extern Analogue-to-Digital Converter o Convertidor Analògic-Digital (ADC), cosa que els fa menys adequats per a un desplegament a gran escala.

Els nodes finals, disposats en una topologia distribuïda, són capaços de recopilar dades acústiques a un ritme de 44 100 Hz i processar-les localment utilitzant el paradigma de computació de *edge computing*, assegurant la privadesa dels ciutadans i evitant enviar fluxos de dades crues a un node centralitzat. Per tant, es pot confirmar que l'objectiu TO2 també s'ha complert.

Thesis Objective 3 o Objectiu de Tesi 3 (TO3): **Utilitzar dades del món real per entrenar i avaluar la plataforma de classificació (programari i maquinari) per estudiar la viabilitat d'un desplegament en el món real.**

Per estudiar la viabilitat d'un desplegament en el món real, es van dur a terme diferents campanyes de gravació a l'escenari d'ús seleccionat (una cruïlla al centre del *L' Antiga Esquerra de l'Eixample* de Barcelona). Una primera campanya d'enregistrament va permetre analitzar el paisatge sonor del barri i definir una taxonomia, i dues campanyes d'enregistrament més van permetre recopilar dades en una topologia amb redundància física.

A causa de les troballes de l'anàlisi de la primera campanya de gravació, es va poder observar que la zona incloïa sons tant de trànsit com d'oci. Aquest paisatge sonor va canviar durant la pandèmia de COVID-19 (especialment durant les restriccions més estrictes de març de 2020), però el treball va servir com a base per seleccionar una taxonomia i veure que els esdeveniments acústics estaven ocorrent constantment i, més important, simultàniament. Per aquesta raó, es va implementar un classificador polifònic (és a dir, multietiqueta).

En la segona i tercera campanya d'enregistrament (les que composen el *Eixample dataset*), tots els esdeveniments acústics van ser etiquetats: tant els que es senten en primer pla i destaquen per sobre el soroll de fons com els de menys nivell acústic que es podrien considerar com a soroll de fons. La hipòtesi inicial que volíem confirmar era que, si un humà és capaç d'escoltar el soroll de fons, un classificador automàtic també pot ser capaç de classificar-lo. No obstant això, a causa dels diferents sons que se superposen en primer pla i fons, la hipòtesi es va validar parcialment, ja que el classificador tenia problemes a l'intentar classificar el soroll de fons.

El problema principal que es va trobar quan es van fer servir dades del món real per a la classificació va ser el desequilibri de classes del conjunt de dades. Aquest fenomen de desequilibri de dades és normal donada la naturalesa dels esdeveniments (esdeveniments que estan ocorrent en un carrer concorregut de la ciutat), en els quals el soroll de trànsit, incloent el pas de cotxes o motocicletes, tendeix a estar més present que, per exemple, altres sons com sirenes d'ambulàncies o sons produïts per persones que caminen. Per mitigar aquest problema, es van provar diferents tècniques d'augment de dades, utilitzant principalment *mix-up augmentation* i combinant tant dades del món real de Barcelona com dades en línia (UrbanSound dataset). En la versió final del classificador, el sistema encara classifica millor les classes més comunes que les que tenen menys instàncies.

En aquest sentit, tot i no ser el rendiment més ideal (idealment, ens agradaria tenir un sistema que pogués classificar per igual els sons que es produeixen en el fons i en primer pla i que, a més, pogués fins i tot classificar els esdeveniments més inesperats), l'algorisme proposat és capaç de detectar amb una precisió raonable (comparada amb els resultats de la classificació que es poden trobar en la literatura) alguns esdeveniments que ocorren simultàniament si tenen prou energia acústica. Per tant, podem concloure que l'objectiu TO3 s'ha complert.

Thesis Objective 4 o Objectiu de Tesi 4 (TO4): **Quantificar fins a quin punt la redundància física dels sensors millora la precisió del classificador.**

Per investigar i quantificar si la redundància física dels sensors millora els resultats d'un classificador automàtic d'esdeveniments acústics, es van dur a terme dues campanyes d'enregistrament tenint en compte una topologia de sensors específica i utilitzant quatre gravadors diferents Zoom H5 simultàniament (el conjunt de dades *Eixample dataset*). Concretament, la topologia seleccionada va considerar quatre sensors, cada sensor es col · locat en una cantonada d'una intersecció de trànsit al mig del *Eixample* de Barcelona. La ubicació escollida va ser la cruïlla entre el carrer Villarroel i el carrer Diputació (codi 95M5+H9). Cada campanya d'enregistrament contenia unes 2 hores i 30 minuts de dades acústiques captades a cada sensor, amb diferents esdeveniments acústics depenent de la naturalesa del soroll generat al carrer en el moment de les gravacions.

La redundància física dels sensors es va tenir en compte quan s'executava una capa intel · ligent sobre els resultats de la classificació inicial de cada node. Més concretament, es van avaluar quatre algoritmes diferents (un Arbre de Decisió, un Bosc Aleatori, un Regressor i un XGBoost). No obstant això, i com va passar igualment quan s'utilitzava un algorisme de classificació basat en DL en cada sensor, el sistema va tenir problemes en classificar els esdeveniments que estaven mal representats en el conjunt de dades a causa del desequilibri de classes. Això va resultar en una mitjana F1-Micro més alta que la mitjana F1-Macro.

No obstant això, els experiments realitzats han permès quantificar fins a quin punt la redundància física pot ajudar a millorar les mètriques de classificació. En el cas del conjunt de dades utilitzat (el conjunt de dades *Eixample dataset* amb tècniques d'augment de dades per a l'entrenament), la segona capa intel · ligent ha permès passar d'una mitjana de F1-Micro del 70% a un 74,1%, el que significa que un 4,1% més de mostres es van classificar correctament. En termes de micro-mitjana, la redundancia física ha permès passar d'un 39% a un 39,3%. Aquesta mètrica indica que la redundància física només ha estat útil en aquells casos en què el conjunt d'entrenament contenia una quantitat substancial de mostres de la categoria que s'està avaluant. Per aquesta raó, podem concloure que l'TO4 s'ha complert.

4.2.2 Respostes a les preguntes de recerca

Research Question 1 o Pregunta de Recerca 1 (RQ1): **Podem detectar i identificar esdeveniments acústics en un univers predefinit usant informació espectral i temporal encara que els esdeveniments es produeixin simultàniament?**

Després de l'anàlisi dels resultats obtinguts en aquesta dissertació, es pot confirmar que, quan s'utilitzen característiques espectro-temporals com els espectrogrames, és possible detectar i classificar esdeveniments acústics (almenys, en l'univers predefinit de l'entorn urbà analitzat) sempre que aquests esdeveniments es produeixin prou a prop del sensor que els monitoritza. En aquest sentit, prou a prop significa que l'esdeveniment destaca per sobre del soroll de fons en el sensor.

En els experiments realitzats, aquells esdeveniments que estaven emmascarats per esdeveniments més forts eren més difícils d'identificar, i el mateix va ocórrer amb aquells esdeveniments que no van ocórrer molt sovint en l'entorn predefinit (les classes que estaven mal representades en el conjunt de dades).

Research Question 2 o Pregunta de Recerca 2 (RQ2): **És possible encabir un algorisme classificador d'àudio en un dispositiu de baix cost per tal que la classificació doni resultats en temps real?**

Els experiments realitzats en aquesta tesi han confirmat que és possible encabir en un sistema de classificació acústic en un dispositiu de baix cost, fins i tot si el classificador està format per una DNN, que normalment es considera com un algorisme pesat. Diferents proves realitzades en certes unitats de computació determinades (models de Raspberry Pi) han demostrat que es poden obtenir resultats en menys de 4 segons. Com que la finestra seleccionada utilitzada en aquest treball és de 4 segons, és un requisit obligatori que el sistema classificador doni un resultat de classificació en menys d'aquesta quantitat de temps, per permetre un sistema fluid en temps real. En realitat, utilitzant el dispositiu més potent —tot i que encara es considera de baix cost— (és a dir, Raspberry Pi Model 4), el sistema va ser capaç de generar un resultat en menys d'1 segon (mitjana de 0,6 segons).

Research Question 3 o Pregunta de Recerca 3 (RQ3): **Fins a quin punt la redundància física dels sensors pot ajudar a millorar un algorisme classificador d'esdeveniments acústics?**

D'acord amb les conclusions obtingudes per a l'objectiu TO4, la redundància física de sensors pot ajudar a classificar aquelles categories que es representen amb un gran nombre d'instàncies en el conjunt de dades i també pot ajudar a classificar aquells esdeveniments que no estan completament emmascarats per altres sons en tots els nodes.

En resum, les contribucions d'aquesta tesi validen la viabilitat d'un desplegament en el món real d'una Wireless Acoustic Sensor Network o Xarxa de Sensors Acústics sense Fils composada per nodes de detecció de baix cost que generarien —en temps real— un resultat de classificació. Aquest resultat de classificació especificaria quins esdeveniments acústics estan ocorrent en un entorn urbà. No obstant això, una cosa que s'ha de tenir en compte és que l'equilibri de classes esdevé crucial quan s'entrena un classificador basat en Deep Learning o Aprenentatge Profund. Una altra qüestió que ha de tenir-se en compte és que, quan s'utilitzen espectrogrames com a entrades, els esdeveniments que estan emmascarats pel soroll de fons no són detectats o classificats adequadament.

4.3 Línies de futur

Aquest treball abasta diversos temes en el cicle de vida d'un sistema de classificació d'esdeveniments acústics. Per a cada tema, s'han identificat algunes qüestions obertes que podrien requerir de recerca addicional per millorar les idees presentades en aquesta dissertació.

Definició de taxonomia: En aquesta dissertació, la taxonomia s'ha definit d'acord amb les dades que s'han recopilat en diferents dies en llocs concrets del centre de la ciutat de Barcelona. No obstant això, aquestes dades poden estar esbiaixades per les hores o el moment de l'any de les campanyes de gravació i els punts de gravació seleccionats. Per tenir una major varietat d'esdeveniments i, per tant, poder parametritzar millor el paisatge sonor de la ciutat, es podrien dur a terme més campanyes d'enregistrament. Es suggereix que, en lloc de campanyes de gravació llargues (de més d'1 hora cadascuna) com les realitzades en aquesta tesi, es gravéssin campanyes més curtes i en diferents llocs de la ciutat (com carrers amb alta densitat de trànsit, parcs, hospitals, escoles, etc.). D'aquesta manera, seria possible validar si les campanyes d'enregistrament realitzades en aquesta tesi (és a dir, el BCNDatset i el conjunt de dades d'Eixample) representen plenament el paisatge sonor de la ciutat. A més, els resultats obtinguts en aquesta tesi poden estar esbiaixats també a causa del canvi del paisatge sonor en diferents condicions (com la pandèmia COVID-19 i les seves conseqüències intrínseques com les restriccions de mobilitat; estacions de l'any diferents o, senzillament, hores diferents al llarg del dia). Per aquesta raó, seria convenient tenir en compte tots aquests paràmetres per a les diferents campanyes d'enregistrament i realitzar un estudi profund per a comprovar com es relacionen amb el paisatge sonor i explorar si la taxonomia hauria d'ampliar-se.

Millora de l'algorisme de classificació: Actualment, els algoritmes desenvolupats només tenen en compte quins esdeveniments s'estan produint en un moment determinat i en una ubicació determinada. En aquest sentit, l'algorisme és capaç de determinar quins esdeveniments acústics s'estan produint, però no el seu nivell de soroll equivalent o si els sorolls compleixen les normes actuals a la zona. En aquest sentit, un treball que es podria duur a terme en el futur consistiria a afegir una capa superior a l'algorisme de classificació que tindria en compte: (1) les característiques de la zona que es monitoritza (per exemple, si és una àrea residencial), (2) el nivell de soroll equivalent que perceb el sensor, (3) el tipus de so que es produeix en temps real i (4) l'hora del dia en què es produeixen els sons. D'aquesta manera, aquesta capa superior seria capaç de detectar si el soroll compleix les normes acústiques de la zona i facilitaria la tasca dels tècnics o experts que gestionen les queixes relacionades amb el soroll a les ciutats.

Afegir memòria al sistema: Una altra manera de millorar el sistema de classificació consistiria a afegir una capa de memòria al sistema. És a dir, afegir una capa de programari que tindria en compte els esdeveniments acústics que van ocórrer en fotogrames passats per predir o validar els esdeveniments que estan ocorrent en el present o en un futur pròxim. Aquesta capa de memòria hauria de tenir en compte la naturalesa intrínseca dels esdeveniments acústics de la taxonomia, ja que hi ha esdeveniments amb més probabilitats de repetir-se que altres. Per exemple, si en els últims 4 segons una sirena d'una ambulància estava present en el paisatge sonor, és molt probable que la sirena encara estigui present en el següent fragment de 4 segons. Això

es deu a la naturalesa de l'esdeveniment de sirena, que sol ser llarg en el temps. No obstant això, aquesta naturalesa no es comparteix amb altres esdeveniments com ara clàxons de cotxes, que poden ocórrer independentment i que solen ser més curtes en el temps que els sons de sirenes. Una futura direcció de treball consistiria en avaluar quins esdeveniments són més propensos a persistir en el temps i comprovar si una capa de memòria en el sistema permetria aconseguir millors resultats de classificació.

Avaluar més tècniques d'augment de dades: En aquest treball, la tècnica principal d'augment de dades que s'ha utilitzat és l'augment de la mescla (*mix-up augmentation*), que consisteix en combinar dos senyals d'àudio en un per tenir una varietat més àmplia d'esdeveniments (es combinen diferents esdeveniments per tal de balancejar les classes). Concretament, les dades de tres conjunts de dades diferents —dos conjunts de dades del món real i un conjunt de dades en línia— s'han combinat, i s'ha seguit una estratègia per fer coincidir totes les etiquetes. Es van trobar problemes quan s'utilitzava el conjunt de dades en línia, ja que les dades d'aquell conjunt es van enregistrar a un ritme de mostreig més baix i no tenien el soroll de fons típic que es podia escoltar a les ciutats, la qual cosa el feia menys realista.

Com a futura línia de treball, seria científicament interessant desenvolupar un sistema que generi mostres acústiques realistes per enriquir el conjunt de dades d'entrenament. La generació d'àudio es podria aconseguir mitjançant un Variational AutoEncoder o AutoCodificador Variacional (VAE) o una Generative Adversarial Network o Xarxa Generativa Antagònica (GAN).

Avaluar el *hardware* seleccionat: Per a aquest projecte, el maquinari dels nodes de la xarxa s'ha triat d'acord amb les decisions adoptades en llegir les especificacions de les unitats de computació i micròfons comercials, però no s'ha dut a terme cap prova en el món real amb altres models. Una futura línia de treball seria provar diferents unitats de computació d'altres marques comercials (com Banana Pi, Jaguar One o Hummingboard) per poder fer una comparació objectiva amb els algorismes ja desenvolupats.

En aquesta línia d'investigació, també seria interessant avaluar el sistema amb diferents tipus de micròfons, utilitzant o bé micròfons USB de diferents marques, o bé MEMS o micròfons d'alta precisió. La idea seria caracteritzar-los físicament i veure els punts forts i febles de cada model.

Provar altres topologies de sensors: La topologia seleccionada per a aquest projecte implicava tenir un sensor situat en cada una de les cantonades d'una intersecció de trànsit, tenint anells de 4 sensors en cada cruïlla. Tenir aquests quatre nodes prou propers han permès estudiar l'efecte de la redundància física dels sensors al carrer. Tanmateix, altres configuracions podrien ser estudiades en un futur treball d'investigació. Per exemple, els nodes es podien col · locar al mig de cada bloc de la ciutat en lloc de a la cantonada, canviant radicalment la topologia. Aquesta ubicació potencial s'il · lustra a Figura 4.1 (triangles verds). Una altra ubicació potencial seria col · locar el sensor al mig

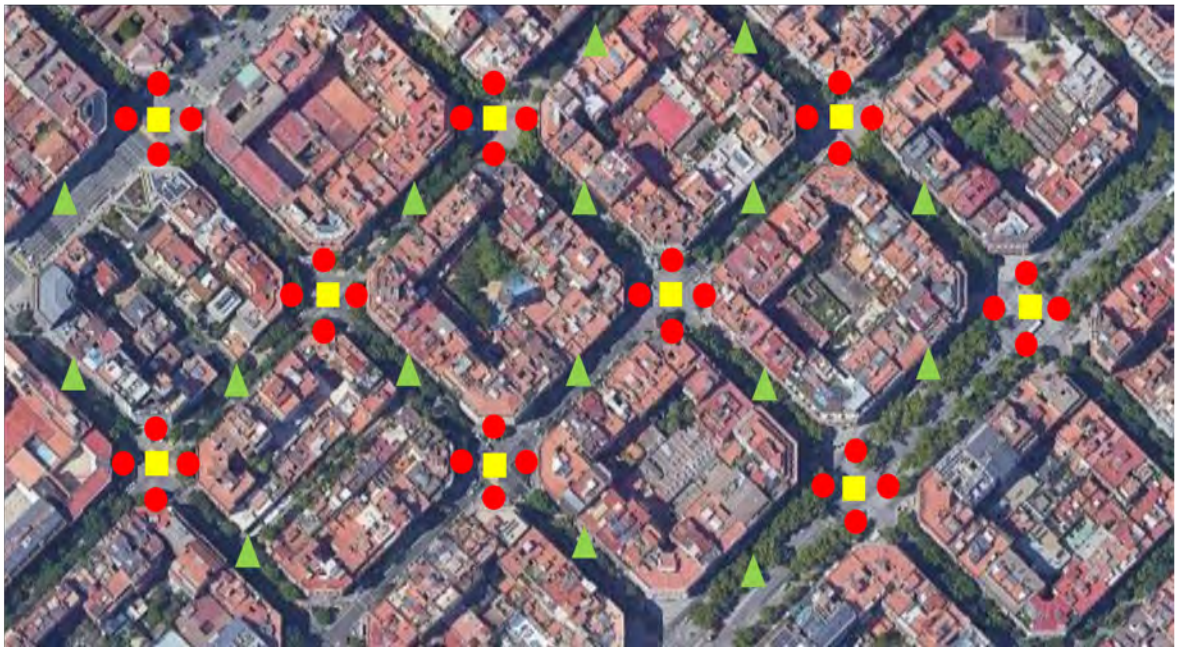


Figura 4.1: Posicions potencials per als sensors. Els punts vermells indiquen la posició actual dels sensors en la topologia proposada, els quadrats grocs i els triangles verds indiquen les ubicacions potencials que podrien ser estudiades en un futur treball.

de la cruïlla (quadrats grocs). D'aquesta manera, només es necessitaria un sensor per carrer, reduint el preu de la xarxa. Tanmateix, això també reduiria la redundància física, de manera que molt pocs esdeveniments (només els que tenen el volum més alt) serien percebuts o detectats per diferents sensors. Això tindria un efecte clar en els resultats de la classificació, que hauria d'analitzar-se. Qualsevol canvi en la topologia implicaria un canvi en la configuració dels anells de la xarxa, però en termes de programari això no seria un gran inconvenient.

Una altra línia d'investigació futura podria ser l'estudi de l'efecte de l'alçada en els nodes. Ara mateix, tots els estudis s'han dut a terme utilitzant els sensors a l'alçada d'1,5 metres i una inclinació de 45 metres respecte al terra. Posicionar els sensors, per exemple, més a prop de les façanes de l'edifici o a l'alçada d'un semàfor, els faria capturar informació acústica diferent. Es podria estudiar l'efecte de canviar la posició dels nodes en el sistema de classificació.

Desplegar el sistema en diferents escenaris d'ús: Finalment, un estudi interessant seria replicar els experiments que s'han realitzat en aquesta tesi en ciutats diferents de Barcelona. Malgrat ser escollida per les seves característiques interessants (una de les ciutats més sorolloses del món i amb arquitectura simètrica), Barcelona no és l'única ciutat en la qual es podria desplegar la WASN. El sistema proposat podria adaptar-se a qualsevol ciutat moderna del món, amb l'única diferència que la topologia hauria de modificar-se lleugerament. Es podrien mantenir els anells de fitxes, tot i que la distància entre els sensors probablement canviaria. Com a línia futura, es suggereix provar el sistema en altres escenaris d'ús amb diferents característiques (per exemple, Girona, que

és una ciutat més petita, Madrid, París, etc.).

A més, el comportament del sistema en el temps també s'ha d'estudiar. Fins ara, tots els experiments s'han provat en períodes curts de temps. Seria d'interès científic desplegar el sistema i deixar-lo connectat durant uns mesos, o fins i tot anys, per estudiar les febleses del maquinari com ara les derivacions potencials de la resposta de freqüència del micròfon o els problemes de sincronització entre les diferents unitats de computació, així com la possibilitat de mesurar els canvis estacionals en l'entorn acústic de la ciutat.

Per acabar, una vegada el sistema hagués estat desplegat durant llargs períodes de temps, es podria estudiar fins a quin punt aquest podria ser utilitzat com a eina per a millorar la salut de la població a força d'intentar modificar el *soundscape* de les àrees més contaminades acústicament. S'ha de tenir en compte, però, que aquesta última línia de futur és molt ambiciosa, ja que implicaria (1) tenir un acord amb els ajuntaments de les ciutats en les quals s'hagués desplegat el sistema i (2) desenvolupar una interfície d'usuari amigable per a què els treballadors dels departaments de qualitat de soroll el poguessin fer servir.

Chapter 4

Conclusions

4.1 Summary

This thesis has investigated acoustic event classification in urban environments using low-cost devices. Specifically, two modern challenges have been addressed: (1) the design of a low-cost, scalable WASN based on programmable commercial sensors and limited computing capabilities able to monitor large-scale areas and (2) the development of a real-time classification algorithm able to run over the designed sensing nodes. This way, this thesis has completed the full cycle of a typical ML or DL project together with the design of a topology and architecture able to perform classification. This cycle has included the selection of a use-case scenario, the study and analysis of the environmental sounds of the selected location, a taxonomy definition for all the sounds that have been heard, a prototyping stage (with online data) of a classification algorithm together with a study on which are the most convenient features to parametrize the acoustic signals, an enhancement of that algorithm using real-world data to support polyphony (multiple acoustic events occurring simultaneously), and the testing of that algorithm over physical sensors. Moreover, the thesis has studied the possibility of deploying the sensors with physical redundancy to check whether this improves the classification results. The purpose of this chapter is to (1) summarize the work by analyzing whether the thesis objectives have been accomplished, (2) highlight the obtained conclusions on each of the stages or processes of the dissertation and (3) provide some future work directions that could be considered in light of the outcomes provided by this dissertation.

4.2 Conclusions

Over the last years, both public and private organizations have put an effort on acoustically monitoring urban environments with the aim of identifying the most polluted zones of a given area due to the side health effects that noise may have on human beings. However, some studies reveal that it is not only the level of noise that matters but also the type of sounds the citizens are exposed to. For this reason, and seeing that most of current technologies just allow to check what is the noise level in a given spot instead of detecting the noise source as well (in most cases, technicians have to go to the place in which noise is generated to check what source is causing the acoustic event), this thesis has been conducted as a small step towards automatic noise source detection in urban environments using low-cost devices and physical redundancy of sensors.

A real-world implementation of the proposed system would allow to monitor which are the most polluted zones in a certain urban environment. Moreover, the system would tell *which* are the acoustic events that are occurring in that area and would allow to have a global

picture of the different soundscapes of different parts of the city. Hence, the proposed system could be used as a tool for public entities to propose measures to mitigate situations that are perjudicial to the health of the population (e.g. redirecting the traffic of different streets or applying noise-related restrictions).

The two main challenges that have been addressed in this dissertation are:

1. Designing low-cost sensors using commercial hardware so they can be deployed over wide areas to be surveyed. It is understood that a low-cost sensor is a sensing node with a commercial price lower than 100€.

2. Developing a classification algorithm using DL techniques that allows to perform automatic acoustic event classification using the designed sensing nodes. Also, this classification algorithm must be capable of classifying events that occur simultaneously, as polyphony is a very common phenomenum on urban environments.

From these two challenges, three research questions and four thesis objectives were defined. The next subsection will first explain whether the thesis objectives have been achieved, and then, the following subsection will answer the research questions.

4.2.1 Achievement of the thesis objectives

Thesis Objective 1 o Objectiu de Tesi 1 (TO1): **Develop an automatic classifier system capable of detecting acoustic events occurring in urban environments using spectral and temporal information.**

This first thesis objective aimed at using both spectral and temporal information for automatic acoustic event classification. To achieve this objective, several acoustic features were analyzed to check which ones were more convenient for the given acoustic data. At the end, spectrograms were chosen, and several spectrograms calculations and techniques (such as spectrogram normalization, standardization, log-mel spectrograms vs. regular spectrograms or processing techniques such as PCEN) were applied to check which representation suited best the data to be characterized.

Also, different window sizes were analyzed by means of a grid search to check which was the most convenient size for the data. As the classification algorithm was a CNN, it was observed that the best window was the one containing enough information to see patterns of the acoustic events defined in the taxonomy. For example, sirens have a clear pattern on time that is useful for the classifier.

Also, this automatic system had to be light enough (in terms of computational load requirements) to be able to run over a low-cost sensor. For this reason, different CNN architectures were tested. At the end, the one that was selected could fit in the proposed computing units.

We can conclude that, after the conducted experiments, this first objective was successfully accomplished.

Thesis Objective 2 o Objectiu de Tesi 2 (TO2): **Design a low-cost hardware platform capable of classifying acoustic events in real-time.**

This second objective is achieved by means of designing and testing a low-cost WASN. Specifically, the sensing nodes have been chosen to be Raspberry Pi computing units and USB plug-and- play microphones.

Raspberry Pi was selected as the computing platform given its fair trade-off between cost and features. Moreover, the platform has a wide support community that may be useful for troubleshooting. Different models of Raspberry Pi were tested, but Raspberry Pi model 4B obtained the best classification results, being able to complete the cycle of data acquisition, data processing and classification (including a DNN and an intelligent ML system that takes into account the outputs of different neighboring nodes) in about 0.6 seconds. The other models of Raspberry Pi evaluated (Model 2B and 3B+) took about 1.3 or 2.5 seconds respectively. It must also be taken into account that the prices of Raspberry Pi models 2 and 3B+ are lower than the price of Raspberry Pi model 4. Hence, depending on the requirements of particular applications, Raspberry Pi models 2 or 3B+ could already satisfy the requirements. The OS running on the Raspberries is Raspbian Lite.

Regarding the microphones, the USB plug-and-play microphone was selected as it did not need any additional hardware to be able to acquire acoustic data with a Raspberry Pi. A complete analysis of the selected microphone or a comparison between different USB microphones has not been carried out in this dissertation, it has been assumed to have a flat frequency response in the frequencies of interest. Other microphone types (such as Micro Electro Mechanical System o Micro Electret Sistema Mecànic (MEMS)) that do not have a flat frequency response or require from an external Analogue-to-Digital Converter o Convertidor Analògic-Digital (ADC) converter, which make them less suitable for a wide deployment, have been discarded.

The final nodes, arranged in a distributed topology, are able to collect acoustic data at a rate of 44 100 Hz and process them locally using the edge computing paradigm, ensuring citizens' privacy and avoiding to send raw data streams to a centralized node.Thus, it can be confirmed that TO2 has successfully been achieved.

Thesis Objective 3 o Objectiu de Tesi 3 (TO3): **Use real-world data to train and evaluate the classification platform (hardware and software) in order to study the feasibility of a real-world deployment.**

To study the feasibility of a real-world deployment, different recording campaigns took place in the selected use-case scenario (a cross-road in the center of the *L'Antiga Esquerra de l'Eixample* of Barcelona). A first recording campaign enabled to analyze the soundscape of the neighborhood and define a taxonomy, and then two more recording campaigns enabled to gather data in a physical redundancy topology.

Due to the findings of the analysis of the first recording campaign, it could be observed that the zone included both *traffic* and *leisure* sounds. This soundscape had for sure

changed during the COVID-19 pandemic (specially during the most strict lockdown of March 2020), but the work served as a baseline to select a taxonomy and see that the acoustic events were happening constantly and, more important, simultaneously. For this reason, a polyphonic (i.e., multi-label) classifier was implemented.

In the second and third recording campaigns (the ones that compose the *Eixample dataset*), all the acoustic events were labelled: both the ones on the foreground standing over the background noise and the ones with less acoustic level that could be considered as background noise. The initial hypothesis that we wanted to confirm was that, if a human is able to hear the background noise, an automatic classifier may be able to classify it as well. However, due to different sounds overlapping in foreground and background, the hypothesis was just partially validated as the classifier struggled when trying to classify the sounds in the background.

The main problem encountered when using real-world data for classification purposes was the high class imbalance encountered in the dataset. This data imbalance phenomenum is normal given the nature of the events (events that are happening in a crowded street of the city), in which traffic noise including the by-pass of cars or motorcycles tends to be more present than, for example, other sounds such as siren sounds from ambulances or sounds produced by people walking. To mitigate this issue, different data augmentation techniques were tested, using mainly mix-up augmentation combining both real-world from Barcelona and online data. In the final version of the classifier, the system still classifies better the most common classes than the ones that have less instances.

In this sense, despite not being the most ideal performance (a system that could classify equally the sounds that occur in background and in foreground and a system that can even classify the events that are least likely to occur), the proposed algorithm is able to detect with a reasonable accuracy (compared to the classification results that can be found in the literature) some events occurring simultaneously if they have enough acoustic energy. Hence, we can conclude that TO3 has been accomplished.

Thesis Objective 4 o Objectiu de Tesi 4 (TO4): **Quantify up to what extent physical redundancy of sensors improves the accuracy of the classifier.**

To investigate and quantify wether physical redundancy of sensors improves the results of an automatic acoustic event classifier, two recording campaigns were carried out taking into consideration a specific sensors topology and using four different Zoom H5 recorders simultaneously (the *Eixample* dataset). Specifically, the selected topology considered four sensors, each sensor being placed in a corner of a traffic intersection in the middle of the *Eixample* of Barcelona. The location was the crossroad between Villarroel Street and Diputació Street (plus code 95M5+H9). Each recording campaign contained about 2 hours and 30 minutes of acoustic data, with different acoustic events depending on the nature of the noise generated in the street at that time.

The physical redundancy of sensors was taken into consideration when running an intelligent layer over the initial classification results of each sensing node. More concretely,

four different algorithms were evaluated (a Decision Tree, a Random Forest, a Learning Regressor and an XGBoost). However, and as it equally happened when using a single DL-based classification algorithm in each sensor, the system struggled when classifying events that were poorly represented in the dataset due to class imbalance. This resulted in a F1-Micro average higher than the F1-Macro average.

However, conducted experiments have enabled to quantify up to what extend physical redundancy could help to improve the classification metrics. In the case of the used dataset (the *Eixample* dataset with data augmentation techniques for training), the second intelligent layer has allowed to pass from a 70% F1-Micro Average metric to a 74.1%, meaning that a 4.1% more of samples were correctly classified. In terms of Micro-averaging, physical redundancy has enabled to pass from a 39% to a 39.3% of accuracy. These metrics indicate that physical redundancy has been helpful only in those cases in which the training set contained a substantial amount of samples from the category being evaluated. For this reason, we can conclude that TO4 has been accomplished.

4.2.2 Answers to the research questions

Research Question 1 o Pregunta de Recerca 1 (RQ1): **Can we detect and identify acoustic events in a predefined universe using spectral and temporal information even if they occur simultaneously?**

After the analysis of the results obtained in this thesis, it can be confirmed that, when using spectro-temporal features such as spectrograms, it is possible to detect and classify acoustic events (at least, in the predefined universe of the analyzed urban environment) as long as those events are occurring close enough to the sensor that is monitoring them. In this sense, close enough means that the event stands out of the background noise in the sensor.

In the conducted experiments, those events that were masked by louder events were more difficult to identify, and the same occurred with those events that did not happen very often in the predefined environment (those classes that were poorly represented in the dataset).

Research Question 2 o Pregunta de Recerca 2 (RQ2): **Is it possible to fit an audio classifier algorithm in a low-cost device so it outputs the classification results in real-time?**

The experiments carried out in this thesis have confirmed that it is possible to fit an acoustic classifier system in a low-cost device, even if the classifier is composed of a DNN, which is usually thought to be a heavy algorithm. Different tests conducted over a few computation units (Raspberry Pi models) have proved to be able to output a result in less than 4-seconds. As the selected window used in this work is 4 seconds, it is a compulsory requirement that the classifier system outputs a classification result in

less than that amount of time, to allow a fluent real-time system. Actually, using the most powerful —and yet, low-cost— device (i.e., Raspberry Pi Model 4), the system was able to output a result in less than 1 second (average of 0.6 seconds).

Research Question 3 o Pregunta de Recerca 3 (RQ3): **Up to what extent physical redundancy of sensors can help improving an acoustic classifier algorithm?**

According to the conclusions obtained for TO4, physical redundancy of sensors can help when classifying those categories that are represented with a large number of instances in the dataset and also for those events that are not completely masked by other sounds in all of the sensing nodes.

To sum up, the contributions of this thesis validate the feasibility of a real-world deployment of a Wireless Acoustic Sensor Network o Xarxa de Sensors Acústics sense Fils composed of low-cost sensing nodes that would output —in real-time— a classification result. This classification result would specify which acoustic events are happening in a urban environment event if they were occurring simultaneously. Something that must be taken into account, though, is that class balancing becomes crucial when training a Deep Learning o Aprenentatge Profund-based classifier. Another issue that must be taken into account is that, when using spectrograms as inputs, those events that are masked by the background acoustic noise may not be properly detected or classified.

4.3 Future work

This work embraces several topics in the life cycle of an acoustic event classification system. For each topic, there have been identified some open issues that could require additional research to improve the ideas presented in this dissertation.

Taxonomy definition: In this dissertation, the taxonomy has been defined according to the data that has been gathered in different days in concrete locations of the city centre of Barcelona. However, this data may be biased by the selected hours or moment of the year of the recording campaigns and the selected recording spots. To have a wider variety of events and, hence, be able to better parametrize the soundscape of the city, more recording campaigns could be carried out. It is suggested that, instead of long recording campaigns (of more than 1-hour each) as the ones conducted in this thesis, shorter recordings are captured in different locations of the city (such as streets with high density of traffic, parks, hospitals, schools, etc.). This way, it would be possible to validate if the recording campaigns carried out in this thesis (i.e., the BCNDatset and the Eixample dataset) do fully represent the soundscape of the city. Moreover, the results may be biased as well due to the change of the soundscape in different conditions (such as COVID-19 pandemic and its intrinsic consequences such as mobility restrictions, different seasons or different hours). For this reason, it would be convenient to take into account all these parameters for different recording campaigns and conduct a deep

study to check how they relate to the soundscape and explore if the taxonomy should be widened.

Classification algorithm enhancement: Currently, the developed algorithms take into account only which events are occurring in a determined moment and at a determined location. In this sense, the algorithm is just capable of determining what acoustic events are occurring, but not their equivalent level or if the noises are meeting the current regulations in the zone. In this sense, a potential future work would consist on adding a superior layer to the classification algorithm that would take into account: (1) the characteristics of the zone being monitored (e.g., residential area), (2) the equivalent noise level being perceived by the sensor, (3) the type of sound occurring at real-time and (4) the hour of the day at which the sounds are occurring. This way, this superior layer would be able to detect if the noise is exceeding the acoustic regulations of the zone and would facilitate the task of the technicians or experts that handle noise-related complaints in cities.

Add memory to the system: Another way of improving the classification system would consist on adding a memory layer to the system. That is, add a software layer that would take into account the acoustic events that occurred in past frames to predict or validate the events that are occurring in the present or in a near future. This memory layer would have to take into account the intrinsic nature of the acoustic events of the taxonomy, as not all the acoustic events are likely to be repeated if they occurred in the past. For example, if on the last 4-seconds a siren of an ambulance was present in the soundscape, it is very likely that the siren is still present in the next 4-seconds fragment. This is due to the nature of the siren event, that is typically long in time. However, this nature is not shared with other events such as car horns, that may happen independently and are usually shorter in time than siren sounds. A future work direction would consist on evaluating which events are more likely to persist in time and checking wether a memory layer in the system would allow to achieve better classification results.

Testing more data augmentation techniques: In this work, the main data augmentation technique that has been used is mix-up augmentation, which consists on combining two audio signals in one to have a wider variety of events (combination of simultaneous events in a single frame and class balancing). Concretely, data from three different datasets—two real-world dataset and one online dataset— have been combined, and a strategy has been followed to make all the labels match. Problems were encountered when using the online dataset, given that it was recorded at a lower sampling rate and it did not have the typical background noise that can be heard in cities, which made it less realistic.

As a future working line, it would be scientifically interesting to develop a system that generates realistic—yet synthetic— urban acoustic samples to enrich the training dataset. The audio generation could be achieved by means of a Variational AutoEncoder

o AutoCodificador Variacional (VAE) or a Generative Adversarial Network o Xarxa Generativa Antagònica (GAN).

Hardware selection testing: For this project, the hardware of the computing nodes has been chosen according to decisions taken when reading literature and specifications of commercial computing units and microphones, but no real-world testing has been carried out. A future line of work would be to try different computing units from other commercial brands (such as Banana Pi, Jaguar One or Hummingboard) in order to be able to objectively make a comparison with the already developed algorithms.

In this research line, it would be interesting too to evaluate the system with different microphone types, using either USB plug-and-play microphones from different brands, MEMS or high-precision microphones. The idea would be to physically characterize them and see the strengths and weaknesses of each model.

Sensors topology checking: The selected topology for this project involved having one sensor placed in a corner of a traffic intersection, allowing rings of 4 sensors on each crossroad. Having this four nodes close enough has enabled to study the effect of physical redundancy of sensors in the streets. However, other configurations could be studied in a future research work. For example, nodes could be positioned in the middle of each city block instead of in the corner, radically changing the topology. This potential location is illustrated in Figura 4.1 (green triangles). Another potential location would be to place the sensor in the middle of the crossroad (yellow squares). This way, only one sensor per cross-road would be required, reducing the price of the network. However, this would reduce the physical redundancy as well, so very few events (only the ones that have a louder volume) would be perceived or detected by different sensors. This would have a clear effect on the classification results, that should be analyzed. Any change in the topology would implicate a change in the configuration of the token rings, but in terms of software this would not be a major inconvenient.

Another future research work could be the study of the effect of height in the sensing nodes. Right now, all the studies have been carried out using a sensors height of 1.5 meters and a 45º inclination from the floor. Positioning the sensors, for instance, closer to the building facades or at the height of a traffic light would make them capture different acoustic information. The effect of changing the position of the nodes in the classifier system could be studied.

Deploying the system in different use-case scenarios: Finally, an interesting study would be to replicate the conducted experiments in cities different than Barcelona. Despite being chosen for its interesting characteristics (one of the noisiest cities in the world and with symmetric architecture), Barcelona is not the only city in which the proposed WASN could be deployed. The proposed system could be adapted to any modern city in the world, with the only difference that the topology would have to be adapted. The token rings could be maintained, even though the distance between

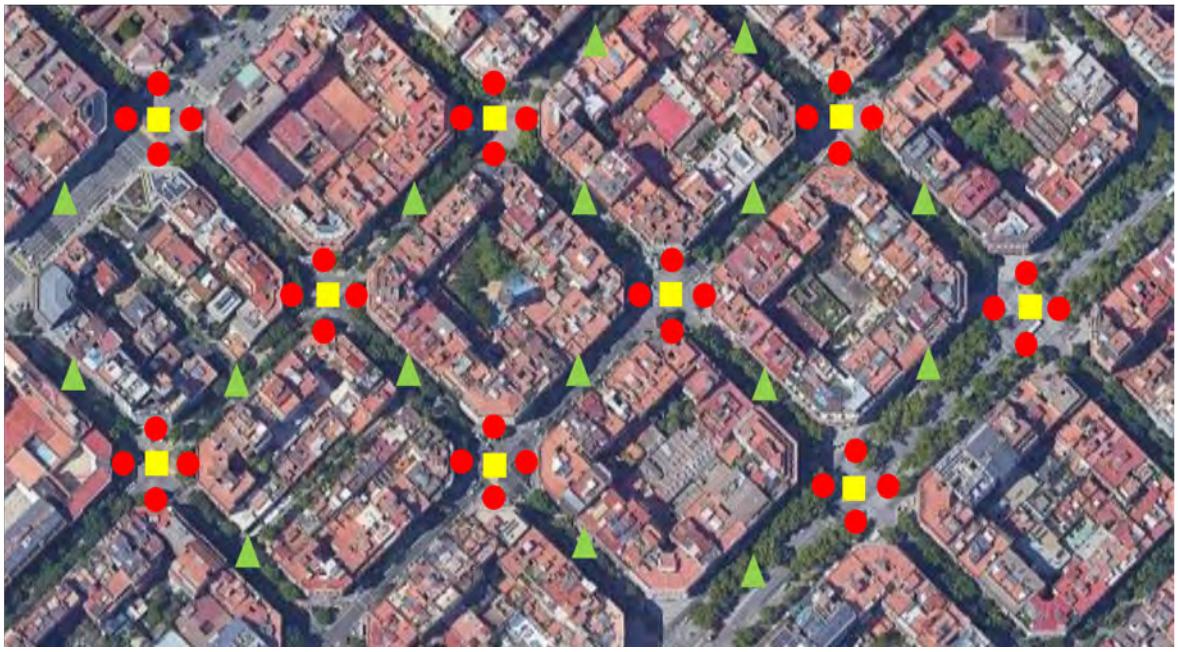


Figure 4.1: Potential positions for the sensors. Red dots indicate the current position of the sensors in the proposed topology, yellow squares and green triangles indicate potential locations that could be studied in a future work.

sensors would probably change. As a future line, it is suggested to try the system in other use-case scenarios with different characteristics (e.g., Girona, which is a smaller city, Madrid, Paris, etc.).

Moreover, the behavior of the system in time should be studied as well. Until now, all the experiments have been tested in shorts periods of time. It would be of scientific interest to deploy the system and leave it connected for a few months—or even years—to study the hardware weaknesses such as potential derives of the frequency response of the microphone or synchronization problems between different computing units, as well as the possibility of measuring the seasonal changes in the acoustic environment of the city.

Also, once the system was deployed in several areas and for long periods of time, it should be studied up to what extent the system could be used as a tool to improve the health of the population by changing the soundscape of the most acoustically polluted areas. It must be considered, though, that this last future work direction is very ambitious as it would imply to (1) set-up an agreement to the city councils of the city in which it was deployed and (2) develop a friendly user-interface so the workers of the noise control department could use it.

Now this is not the end.
It is not even the beginning of the end.
But it is, perhaps, the end of the beginning.
— Winston Churchill

Capítol 5

Articles complementaris al compendi

Article IV

Gerardo José Ginovart-Panisello, Ester Vidaña-Vila, Selene Caro-Via, Carme Martínez-Suquía, Marc Freixes, Rosa Ma Alsina-Pagès. 'Low-Cost WASN for Real-Time Soundmap Generation'. Presentat a: *8th International Symposium on Sensor Science* and published in *Engineering Proceedings*. Vol. 6, no. 1 (2021), pp. 57. DOI: 10.3390/I3S2021Dresden-10162.

Aquest primer article complementari té com a objectiu donar al lector més informació sobre la plataforma de *firmware* utilitzada per a la classificació. Tot i que la WASN descrita en l'article té com a objectiu generar mapes de so, els sensors són molt similars als utilitzats en aquesta tesi per classificar esdeveniments acústics. En aquest sentit, aquest article mostra tots els components necessaris per configurar una unitat de computació d'un sol node de detecció juntament amb el micròfon USB seleccionat i el procés de disseny i avaluació.

Article V

Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, Joan Navarro. 'Improving classification accuracy of acoustic real-world urban data using sensors physical redundancy'. Presentat a: *26th IEEE Symposium on Computers and Communications (ISCC)*. (2021), Athens, Greece. pp. 1-4. DOI: 10.1109/ISCC53001.2021.9631402

Aquest segon article complementari serveix com un pas intermedi entre el treball publicat en els dos últims articles del compendi. En aquest cas, s'analitza una hora d'enregistraments d'àudio del món real per avaluar si la redundància física és significativa quan s'utilitzen dades reals en llocs propers. En aquest article, només s'utilitzen 10 categories d'esdeveniments acústics, ometent els altres, i no s'aplica la classificació multietiqueta. El propòsit d'incloure el treball en aquesta dissertació és mostrar com transicionar d'un conjunt de dades net i controlat cap a la classificació de dades del món real.

Article VI

Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, Joan Navarro. 'Prototyping a low-cost Wireless Acoustic Sensor Network with physical redundancy to automatically classify acoustic events in urban environments'. Pòster presentat a: *UrbanSound Symposium 2021* and abstract published in *Engineering Proceedings*, Vol. 72, (2021), DOI: 10.3390/proceedings2021072004

Aquest tercer treball complementari il · lustra com sintetitzar les idees de l'article presentat anteriorment d'una manera resumida per a propòsits de difusió.

Article VII

Ester Vidaña-Vila, Dan Stowell, Joan Navarro, Rosa Ma Alsina-Pagès. 'Multilabel acoustic event classification for urban sound monitoring at a traffic intersection'. Pòster presentat a: *Deep Learning Barcelona Symposium 2021*.

Aquesta cuarta obra complementària mostra un altre pòster presentat en el transcurs del doctorat, i resumeix un treball que analitza com diferents configuracions experimentals que fan servir diferents tècniques d'augment de dades ajuden a millorar els resultats de la classificació.

Article VIII

Ester Vidaña-Vila, Joan Navarro, Rosa Ma Alsina-Pagès, Álvaro Ramírez. 'A Two-Stage Approach To Automatically Detect and Classify Woodpecker (Fam. *Picidae*) Sounds'. Publicat a: *Applied Acoustics*. Vol. 166, (2020), pp. 107312. DOI: 10.1016/j.apacoust.2020.107312.

Aquest cinquè article complementari exposa la investigació duta a terme en un camp diferent (bioacústica) utilitzant diferents tipus de característiques. Els ocells picots que habiten a la península Ibèrica han estat seleccionades com a espècies a classificar, ja que són d'interès per al seguiment dels entorns naturals.

Article IX

Júlia Blanch, Ester Vidaña-Vila, Rosa Ma Alsina-Pagès. 'Analysis of the Noise Impact of the Airport of Barcelona to the Llobregat Delta Natural Environment during the 2021 Lockdown period'. Presentat a: *8th Electronic Conference on Sensors and Applications*. (2021), DOI: doi:10.3390/ecsa-8-11267.

Aquest sisè treball complementari fusiona el camp del paper anterior (bioacústica) i l'escenari d'aquesta tesi (entorns urbans). Concretament, es dissenya un detector automàtic d'esdeveniments acústics per a dades captades en un parc natural prop d'una ciutat i l'aeroport, i per tant té en compte sons naturals i urbans.

Chapter 5

Complementary papers to the compendium

Paper IV

Gerardo José Ginovart-Panisello, Ester Vidaña-Vila, Selene Caro-Via, Carme Martínez-Suquía, Marc Freixes, Rosa Ma Alsina-Pagès. 'Low-Cost WASN for Real-Time Soundmap Generation'. Presented in: *8th International Symposium on Sensor Science* and published in *Engineering Proceedings*. Vol. 6, no. 1 (2021), pp. 57. DOI: 10.3390/I3S2021Dresden-10162.

This first complementary paper aims at giving the reader more information about the firmware platform used for classification. Even though the WASN described in the paper aims at generating sound maps, the sensing nodes are very similar to the ones used on this thesis to classify acoustic events. In this sense, this paper depicts all the necessary components to set-up a computing unit of a single sensing node together with the selected USB microphone and the design process and evaluation.

Paper V

Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, Joan Navarro. 'Improving classification accuracy of acoustic real-world urban data using sensors physical redundancy'. Presented in: *26th IEEE Symposium on Computers and Communications (ISCC)*. (2021), Athens, Greece. pp. 1-4. DOI: 10.1109/ISCC53001.2021.9631402

This second complementary paper serves as an intermediate step between the work published in the two last papers of the compendium. In this case, 1 hour of real-world audio recordings are analyzed to evaluate if physical redundancy is significant when using real-world data in close spots. In this paper, only 10 categories of acoustic events were used, omitting the others, and multi-label classification is not applied. The purpose of including the work in this dissertation is to show how to start moving from a controlled, clean dataset towards the classification of real-world data.

Paper VI

Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, Joan Navarro. 'Prototyping a low-cost Wireless Acoustic Sensor Network with physical redundancy to automatically classify acoustic events in urban environments'. Poster presented in: *UrbanSound Symposium 2021* and abstract published in *Engineering Proceedings*, Vol. 72, (2021), DOI: 10.3390/proceedings2021072004

This third complementary work illustrates how to synthesize the ideas of the paper presented above in a summarized way for dissemination purposes.

Paper VII

Ester Vidaña-Vila, Dan Stowell, Joan Navarro, Rosa Ma Alsina-Pagès. 'Multilabel acoustic event classification for urban sound monitoring at a traffic intersection'. Poster presented in: *Deep Learning Barcelona Symposium 2021.*

This forth complementary work shows another poster presented in the course of the PhD, and summarizes a work that analyzes how different experimental set-ups using different data augmentation techniques help improving classification results.

Paper VIII

Ester Vidaña-Vila, Joan Navarro, Rosa Ma Alsina-Pagès, Álvaro Ramírez. 'A Two-Stage Approach To Automatically Detect and Classify Woodpecker (Fam. *Picidae*) Sounds'. Published in: *Applied Acoustics.* Vol. 166, (2020), pp. 107312. DOI: 10.1016/j.apacoust.2020.107312.

This fifth complementary paper exposes the research carried out in a different field (i.e., bioacoustics) using different types of features. Woodpecker birds inhabiting the Iberian peninsula have been selected as the species to be classified, as they are of interest for the monitoring of natural environments.

Paper IX

Júlia Blanch, Ester Vidaña-Vila, Rosa Ma Alsina-Pagès. 'Analysis of the Noise Impact of the Airport of Barcelona to the Llobregat Delta Natural Environment during the 2021 Lockdown period'. Presented in: *8th Electronic Conference on Sensors and Applications.* (2021), DOI: doi:10.3390/ecsa-8-11267.

This sixth complementary work merges the field of the paper above (bioacoustics) and the topic of this thesis (urban environments). Concretely, an automatic acoustic event detector is used in a natural park near a city and the airport, and hence takes into consideration both natural and urban sounds.

Paper IV

Low-Cost WASN for Real-Time Soundmap Generation

Gerardo José Ginovart-Panisello, Ester Vidaña-Vila, Selene Caro-Via, Carme Martínez-Suquía, Marc Freixes, Rosa Ma Alsina-Pagès

Presented in *8th International Symposium on Sensor Science*, Published in *Engineering Proceedings*, May 2021, volume 6, issue 1, pp. 57. DOI: 10.3390/I3S2021Dresden-10162

IV

Abstract

Recent advances in technology have enabled the development of affordable low-cost acoustic monitoring systems, as a response of several fields of application that require a close acoustic analysis in real-time: road traffic noise in crowded cities, biodiversity conservation in natural parks, behavioural tracking in the elderly living alone and even surveillance in public places for safety reasons. This paper presents a low-cost wireless acoustic sensor network developed to gather acoustic data to build a 24/7 real-time soundmap. Each node of the network comprises an omnidirectional microphone and a computation unit, which processes acoustic information locally to obtain non-sensitive data (i.e., equivalent continuous loudness levels or acoustic event labels) that are sent to a cloud server. Moreover, it has also been studied the placement of the acoustic sensors in a real scenario, following acoustics criteria. The ultimate goal of the deployed system is to enable the following functions: *i)* to measure the L_{eq} in real-time in a predefined window, *ii)* to identify changing patterns in the previous measurements so that anomalous situations can be detected and *iii)* to prevent and attend potential irregular situations. The proposed network aims to encourage the use of real-time non-invasive devices to obtain behavioural and environmental information, in order to take decisions in real-time.

IV.1 Introduction

In recent years, the advances in technology have led WASNs (Wireless Acoustic Sensor Networks) emerge as a powerful tool to survey from the acoustic health of the population living on urban areas (Basten and Wessels 2014) to the biodiversity conservation in forests (Vidaña-Vila et al. 2020). In parallel, the development on smart-homes has allowed to include similar networks on indoor environments aiming to promote independence and well-being among the active elder population living on their own homes (Navarro et al. 2018). The advantage of WASNs compared to other monitoring systems (e.g., video surveillance systems

or networks composed of wearable devices) is that they are perceived as less intrusive by users (Sun et al. 2011), specially when data is processed locally on the node and, hence, private information of the user (i.e., raw audio data) is not shared to a central node or neighboring nodes.

For this reason, several research projects have developed networks composed of multiple sensing nodes with different features and capabilities. For example, in the context of the IDEA project (Botteldooren et al. 2011), Domínguez *et al.* (Domínguez et al. 2013) propose the usage of low-cost nodes (cost of around 50 €) to monitor outdoor environments that actively auto-check the frequency response of the microphone of each node by embedding a low-cost speaker that generates a periodical frequency sweep. This way, they are able to detect failures in the nodes. Another example of an outdoor acoustic sensor network is the one explained in (Bell and Galatioto 2013), that is centered on the framework of the MESSAGE project. In their work, Bell and Galatioto present the results obtained on a WASN of 50 nodes in which, apart from a noise detector module, each node incorporates traffic and chemical sensor modules. As computational unit, they use a microcontroller with low processing capabilities. Regarding indoor WASNs, the homeSound project (Alsina-Pagès et al. 2017) proposes a network architecture with several sensing nodes that send their information to a concentrator node composed of a GPU with parallel computing capabilities.

This work presents a proof of concept of a sensor and a generic WASN aimed to acoustically monitor indoor or outdoor environments to generate a 24/7 real-time soundmap. The paper is organized as follows: Section IV.2 details the requirements that must be satisfied when deploying the sensing nodes, Section IV.3 describes the design of the proposed sensor and the network in terms of hardware, Section IV.4 explains the evaluation carried out in the design in order to validate the feasibility of the proposal and, finally, Section IV.5 discusses the main conclusions of the work.

IV.2 Requirements

Regarding the sensors location, the following requirements must be satisfied according to the ISO 1996-2 (ISO 1996-2 : 2007 2007). For outdoor measurements, microphones must be located at a height of 4.0±0.5 meters from the floor in high building areas, and 1.2±0.1 meters in residential areas. The distance between the microphones and reflecting surfaces should be from 0.5 to 2 m. Regarding indoor measurements, microphones should be placed 0.5 meters apart from walls and 1 meter apart from significant sound-transmission elements. Distance between sensors should be greater than 0.7 m. Furthermore, before deployment, sensors should be calibrated to get reliable measurements in all nodes. All sensors should be tested with 94 dB level at 1 kHz at 1 m distance in a controlled environment such as an anechoic chamber, by means of a calibrator.

IV.3 Hardware Design

All units of the WASN are identical to simplify the scalability in number of nodes. Concretely, each node contains a Raspberry Pi 3B+ (Foundation n.d.) (RPi) as its computational unit. Since the system has been designed to be steadily active, the node may reach high temperatures. To avoid heating problems, a heat sink has been placed to cool it down. It is important to highlight that the heat sink should not include a fan, as it would generate noise, thereby affecting the measurements conducted by the microphone.

The selected computation unit (i.e., RPi) has four USB ports and a 40-pins GPIO header to connect different peripherals and a WIFI modem to transmit data. As a low-cost alternative of an acoustic sensor, a plug-and-play USB Microphone with an external ADC integrated in the serial bus has been chosen. The sensor has an omnidirectional acoustic pattern that allows to capture all possible sound sources from any direction at a maximum sampling rate of 48 KHz at 16 bits (GYVAZLA n.d.). This electret condenser microphone is USB powered. Thus, it increases the electrical power consumption of the unit. To ensure a correct full functionality, the node requires a 5 V 3 A power supply. Figure IV.1 shows the elements that compose each node of the network. In order to make the nodes suitable for a real-world deployment, all the elements are integrated in a small 3D printed rectangle box designed with SketchUp (Sketchup n.d.) with holes for the power wire, the microphone capsule and heat dissipation. The box integrates all the parts into a single node element minimizing the size to the maximum and protects the node.

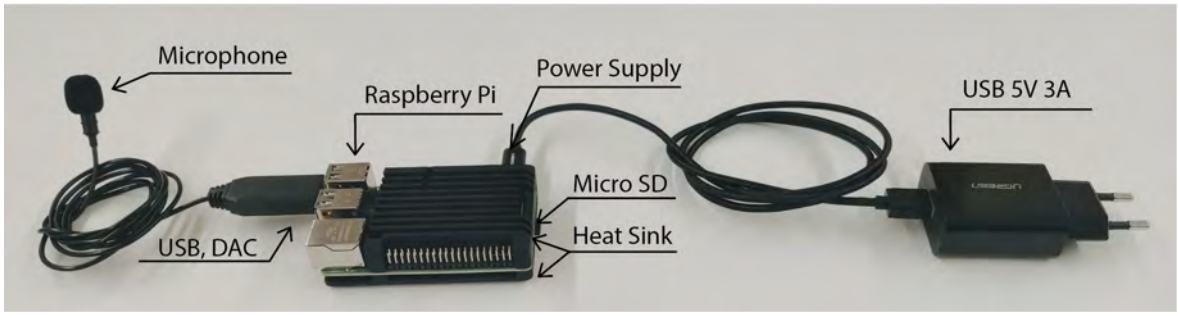


Figure IV.1: Hardware description for each node of the network.

IV.4 Design Process and Evaluation

The sensor design has to balance a low component price and a good performance to obtain an accurate and reliable WASN. The microphone features will limit the accuracy and sensitivity of the measurements. For this reason, different microphone models were compared in order to ensure that the microphone used in the setup is economical and has a frequency response as flat as possible. Concretely, authors compared the following models: *i)* Micro Electro Mechanical System (MEMS) were discarded as the frequency response of the microphone was not flat enough; *ii)* a high-precision measuring microphone (Behringer ECM8000) was discarded as well as it is too big for the purpose of the project (it doubles the size of the RPi) and requires and external ADC; *iii)* the KY-038 microphone was discarded too as its

Table IV.1: Main features and components of the nodes of the network.

Component	**Model**	**Main features**	**Price**
Microphone	LYM00002	Lavelier USB 16 bits/sample 48KHz	11€
Computational unit	Raspberry Pi Model 3B+	SDRAM 1 GB 64 bits CPU at 1.4 GHz WiFi connection	37€
Power supply	UGREEN CD122	18W/5V/3A	12€
USB wire	USB to microUSB		3€
Total			63€

output is analog, therefore requiring an external ADC; *iv)* a USB plug-and-play Microphone (LYM00002) has a flat frequency response, enough sensibility and incorporates an ADC. Hence, the USB microphone specifications together with its reduced price make this microphone ideal for the project.

Regarding to the selection of the CPU module, another comparison was done to ensure that the nodes are capable of locally processing data to avoid sending raw data streams to another node. Models such as Jaguar One or Banana Pi were discarded as their support community is not as big as the one offered by Raspberry. Those CPUs offering extra characteristics not needed for this project (e.g., Hummingboard, Cubieboard5) were discarded too. Finally, models lacking of a WiFi module (PcDuino4, ODROID-C2, Beaglebone Black) were discarded as well. Raspberry pi model 3B+ offers good computer capabilities at a reasonably low-cost and a wide support community.

To test the capabilities of the acoustic sensors, each node has been programmed to process an audio stream sampled at 22.05 kHz, with a bit depth of 16 bits and in 4 seconds windows. Specifically, the sensors calculate continuous acoustic descriptors such as the equivalent loudness level each 4 sec. As the audio streams are processed locally, sensitive information (i.e., raw audio data) is kept in the node and only non-private data is sent through the network. The nodes of this work run OS Raspian Lite, and conducted test have shown that the system uses the 100 % of CPU and 20 % of RAM when running the software test, which is continuously *i)* acquiring 4-second windows of raw audio data, *ii)* processing the audio streams to obtain acoustic descriptors such as the equivalent level, and *iii)* storing the data descriptors in the node's memory and sending them to a cloud server together with a time-stamp.

To synchronize the different nodes of the network, the Network Time Protocol (NTP) has been chosen. To validate the correct functioning of the synchronization, a test software was programmed to be executed automatically in the nodes after booting the system. In this test software, the node waits until a specific minute to start recording a *.wav file. Once the recording started, some acoustic impulses were generated at the same distance of two microphones and, later on, the two *.wav files were manually analysed. Results on the analysis validate that the delay between both files impulses was about 1 ms.

IV.5 Conclusions

A low-cost WASN has been designed for acoustic indoor and outdoor monitoring. Each node of the WASN includes a Raspberry Pi 3B+ which processes in real-time the audio captured by an USB microphone, to evaluate several acoustic features, which afterwards are sent to the cloud. The minimum requirements to draw a soundmap of the environment using the data sent by the nodes have been also analysed depending on the environment. Conducted tests ensure the synchronisation between the nodes, thus avoiding the need of a hub. The decentralised design together with the use of non-sensitive features, allow us to envisage the application of the proposed WASN in surveillance of active elderly in their own homes or in the street for noise monitoring solutions. Moreover, the proposal has taken into account scalability, so a more complex signal processing could be done in the nodes in the future. The authors set as future lines the detection of acoustic events of interest, which could be implemented in the nodes, and predefined alarms could be triggered accordingly.

Author's contributions

Conceptualization, RM.A-P.; software, GJ.G-P, S.C.V. and E.V-V; validation, M.F and RM.A-P; formal analysis, M.F; investigation, all authors; data curation, GJ.G-P and M.F; writing-original draft preparation, E.V-V, GJ.G-P, S.C.V and C.M-S; writing—review and editing, M.F, RM.A-P.; visualization, E.V-V, GJ.G-P and S.C.V; supervision, M.F and RM.A-P.; project administration, RM.A-P.; funding acquisition, RM.A-P.

Funding

Authors would like to thank La Salle for partially funding the research.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADC	Analog to Digital Converter
MEMS	Micro Electro Mechanical System
NTP	Network Time Protocol
WASN	Wireless Acoustic Sensor Network

References

Alsina-Pagès, Rosa Ma, Navarro, Joan, Alías, Francesc and Hervás, Marcos (2017). 'homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring'. In: *Sensors* vol. 17, no. 4, p. 854.

Basten, Tom and Wessels, Peter (2014). 'An overview of sensor networks for environmental noise monitoring'. In: *Proceedings of ICSV21*.

Bell, Margaret Carol and Galatioto, Fabio (2013). 'Novel wireless pervasive sensor network to improve the understanding of noise in street canyons'. In: *Applied Acoustics* vol. 74, no. 1, pp. 169–180.

Botteldooren, Dick, De Coensel, Bert, Oldoni, Damiano, Van Renterghem, Timothy and Dauwe, Samuel (2011). 'Sound monitoring networks new style'. In: *Acoustics 2011: Breaking New Ground: Annual Conference of the Australian Acoustical Society*. Australian Acoustical Society, pp. 1–5.

Domínguez, Federico, Reinoso, Felipe, Touhafi, Abdellah, Steenhaut, Kris et al. (2013). 'Active self-testing noise measurement sensors for large-scale environmental sensor networks'. In: *Sensors* vol. 13, no. 12, pp. 17241–17264.

Foundation, Raspberry Pi (n.d.). *Raspberry Pi 3 Model B+ (Datasheet)*. https://static.raspberrypi.org/files/product-briefs/Raspberry-Pi-Model-Bplus-Product-Brief.pdf (accessed on 10 Mar 2021).

GYVAZLA (n.d.). *USB Omnidirectional Microphone*. https://igyvazla.com/products/gyvazla-usb-microphone/ (accessed on 10 Mar 2021).

ISO 1996-2 : 2007 (2007). *Acoustics - Description, measurement and assessment of environmental noise - Determination of environmental noise levels*. International Organization for Standardization, Geneva, Switzerland.

Navarro, Joan, Vidaña-Vila, Ester, Alsina-Pagès, Rosa Ma and Hervás, Marcos (2018). 'Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios'. In: *Sensors* vol. 18, no. 8, p. 2492.

Sketchup (n.d.). *3D viewer on the web*. https://www.sketchup.com (accessed on 10 Mar 2021).

Sun, Zheng, Purohit, Aveek, Yang, Kathleen, Pattan, Neha, Siewiorek, Dan, Smailagic, Asim, Lane, Ian and Zhang, Pei (2011). 'Coughloc: Location-aware indoor acoustic sensing for non-intrusive cough detection'. In: *International Workshop on Emerging Mobile Sensing Technologies, Systems, and Applications*. Citeseer.

Vidaña-Vila, Ester, Navarro, Joan, Alsina-Pagès, Rosa Ma and Ramírez, Álvaro (2020). 'A two-stage approach to automatically detect and classify woodpecker (Fam. Picidae) sounds'. In: *Applied Acoustics* vol. 166, p. 107312.

Paper V

Improving classification accuracy of acoustic real-world urban data using sensors physical redundancy

Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, Joan Navarro

Published in *26th IEEE Symposium on Computers and Communications (ISCC)*, September 2021, Athens, Greece. DOI: https://doi.ieeecomputersociety.org/10.1109/ISCC53001.2021.9631402

Abstract

Latest advances in modern society together with the increase of the population living in urban areas have transformed these environments into noisy spaces. Current regulations limit the amount of noise-per-source that can impact the population. Hence, automatically identifying acoustic events in urban environments is of great interest for public administrations to preserve citizens' health. Therefore, alternatives that are typically composed of expensive sensing devices committed to individually survey a specific area have been researched. The purpose of this paper is to assess the performance of an alternative approach composed of a low-cost acoustic wireless sensor network that takes advantage of physical redundancy. Specifically, the evaluated system incorporates a deep neural network running in each sensor node and a distributed consensus protocol that implements a set of heuristics to benefit from the classification results of neighboring nodes surveying the same area (i.e., physical redundancy). To evaluate this system, real-world acoustic data were collected simultaneously from four different spots of the same crossroad in the centre of Barcelona and further processed by the system. Obtained results suggest that physical redundancy of sensors improves the classifier's confidence and increases the classification accuracy.

V.1 Introduction

Unwanted sounds that disturb individuals and their communication are commonly referred to as noise (Bello et al. 2019). Prolonged exposure to high levels of noise is harmful for human beings (Office 2017). Noise affects several everyday life activities (Test et al. 2011), may be the cause of chronic injuries on the sympathetic nervous system (Su-bei 2007), and could contribute to originate psychological disorders such as severe annoyance (Guski et al. 2017).

In this regard, the World Health Organization recommends the maximum noise levels (typically using L_{Aeq} metrics) under different conditions to ensure acoustic health in populated environments (e.g., teaching facilities, residential areas) (Hurtley 2009). Similarly, the European Commission has established a set of regulations (Flindell and Walker 2004) to limit the amount of noise that the population is exposed to. These regulations limit the maximum L_{Aeq} noise levels for specific noise sources such as road traffic noise, railway noise or airport noise among others (Office 2017).

The rising interest on protecting citizens from harmful noises has motivated the research on monitoring the acoustic activity in urban spaces (Bello et al. 2019; Sevillano et al. 2016). Traditionally, this is done by deploying a set of interconnected acoustic sensors, coined as Wireless Acoustic Sensor Network (WASN), in a given area to automatically listen to the events—and their equivalent L_{Aeq} levels—that happen throughout the day (Vidaña-Vila et al. 2020b) and, thus, conduct noise surveillance tasks. One of the main design challenges in this type of distributed systems is to keep scalability while maintaining cost-effectiveness (Pham and Cousin 2013). That is, being able to acoustically cover a large-scale area with a budget-constrained equipment. Typically, this limits the scope of the measurements and areas to be monitored, which results in very few acoustic sensors per squared kilometer (Mydlarz et al. 2019).

In this regard, in a previous work (Vidaña-Vila et al. 2020b), authors proposed an alternative WASN inspired by the concept of physical redundancy (Piper et al. 2017) that was committed to (over)populate a urban space with several low-cost acoustic sensor devices composed of inexpensive hardware (i.e., Raspberry Pi Model 2B (RPi) and USB microphone OUT-AMLO-0872).

This system used a preliminary distributed intelligence layer running on top of each acoustic sensor node together with a deep neural network. This layer was committed to analyze in real-time the acoustic samples from all the neighboring sensors that are physically close (i.e., less than 100 m) and reach a consensus on which was the most probable event that happened over a fixed set of 10 possible classes. In (Vidaña-Vila et al. 2020b), this approach was evaluated using with the UrbanSound 8k dataset (Salamon et al. 2014) under laboratory conditions mimicking (Bergadà and Alsina-Pagès 2019) a specific corner of the Eixample of the city of Barcelona. However, no real-world evaluation of this system was conducted so far. That is, assessing the system performance using real-operation data (with class imbalance, sounds belonging to event classes out of the training data set, multiple events happening concurrently, acoustic events masked by background noise, etc.).

Therefore, the work presented on this paper aims to assess the performance of this system when exposed under realistic real-world circumstances and detail which *minor* enhancements have been made to it in order to operate appropriately. Specifically, the purpose of this paper is manifold: (1) detail the specific heuristic rules carried on the distributed intelligence layer to improve the system performance, (2) to further evaluate the approach presented in (Vidaña-Vila et al. 2020b) under real-operation conditions, (3) assess up to what extent acoustic physical redundancy contributes to improve the classification accuracy, and (4)

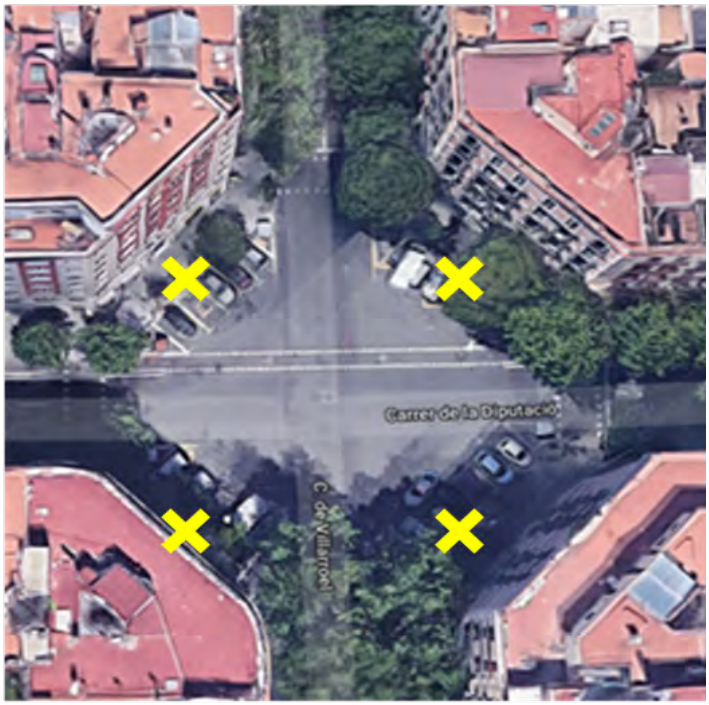


Figure V.1: Aerial view from Google Maps of the crossroad where the recording was conducted.

describe the faced difficulties on migrating from laboratory conditions toward real-operation.

V.2 Data Collection Scheme

In order to evaluate the performance of the algorithm (Vidaña-Vila et al. 2020b) in a real-operation scenario, authors organized a recording campaign that took place on 17 Nov. 2020 at 13:00h to collect a real-world dataset containing four simultaneous recordings located in pre-determined measuring points in a crossroad of Eixample in Barcelona. For this purpose, authors used four Zoom H5 recorders in the intersection between Villarroel street and Diputació street on the city centre of Barcelona. Each recorder was deployed on one of the four corners of the intersection, as it can be seen in Fig. V.1. During the campaign, recorders were placed over tripods, accomplishing the conditions of a minimum distance of 4 m from the nearest wall, 1.5 m from the floor and with an inclination of 45^o from the floor. The recorders synchronization was as follows:

1. All the recorders were set up at a sampling frequency of 44,100 Hz and with the same gain value (minimum gain, because the city centre of Barcelona presents high values of noise) and mounted them over the tripods at the same height (1.5 m).

2. After that, they were placed very close in space in one of the corners, the rec button was activated and authors generated three impulses (hands clapping). These impulses were used to later synchronize the audio recordings, as not all the rec buttons were pressed simultaneously.

3. Next, each recorder was deployed on a different corner of the crossroad and, and then, three more impulses (hand claps) that would be locally heard by the recorder were generated. This second set of impulses indicated that from that moment all the information being recorded was no longer interfered by the set-up process. Hence, the simultaneous valid data started with the last clapping of the last recorder that was positioned (see Fig. V.1).

4. To finish the recording campaign, authors generated three impulses to indicate that the recorder was going to be moved (and hence, the data collection ended) and joined again all the recorders in a single corner to generate three final impulses, to ensure the synchronisation also at the end of the test.

After synchronising the four recording sources, the dataset was ready to be classified using the algorithm detailed in (Vidaña-Vila et al. 2020b).

V.3 Description of the Classification Algorithm

Each one of the four acoustic nodes is committed to run a distributed intelligence algorithm to increase its individual event classification performance (Vidaña-Vila et al. 2020b). The same piece of software is running in all nodes. The algorithm consists of two layers: the first layer, coined as local classifier, implements a SqueezeNet (Iandola et al. 2016) deep neural network that classifies in real-time the acoustic events occurring. The output of this layer (i.e., a vector containing the probability of the event belonging to each one of the possible classes) is supplied to the second layer. The second layer, coined as consolidation, features a distributed consensus protocol that (1) shares the local classification results with the neighboring nodes, and (2) takes into account the classification results from the neighbouring sensors—including the local predictions—to give a final classification label. The precise specifications regarding the communication between sensors and their network topology are wider detailed in (Vidaña-Vila et al. 2020b). Further details on each layer and how they have been upgraded to adapt themselves to real-operation conditions are provided in the following paragraphs.

Regarding the deep neural network, the SqueezeNet network takes as input the spectrograms of 4-seconds windows of audio data. In our previous work (Vidaña-Vila et al. 2020b), the network was trained using the UrbanSound 8K dataset (Salamon et al. 2014), which contains acoustic data from 10 types of different sounds from New York City: *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren,* and *street music*. To obtain reliable results from the classifier system in the real-operation environment, the training dataset has been enriched using real-world urban acoustic data from other sources (see Section V.4.1). This enables us to train the system using a larger number of samples focused in the noises produced by traffic and people in the centre of Barcelona.

Regarding the distributed consensus protocol, it consists on a heuristic set of rules to be applied on each node (i.e., acoustic sensor). These rules take into account the probability of occurrence of the different events (i.e., the Softmax output of the neural network) on each of the sensors. The heuristic rules that have been specifically defined for this work are:

Rule 1: If the output of the local neural network classifies an event with a confidence (i.e., probability value provided by the last Softmax layer of the deep network) smaller than 80 %, the protocol assigns a provisional label of *unknown* to that 4-seconds fragment. Note that this label may change according to the probability values of the labels assigned by neighboring nodes. This rule aims to identify those events that the deep network may have assigned a wrong label.

Rule 2: If the previous condition is not met (i.e., the classification confidence is greater than 80 %) and the local neural network detects an event whose equivalent level L_{eq} is typically low (i.e., *dog* or *people*), then the outputs of the neighbour sensors are ignored and that fragment is labelled with the classification result of the local neural network. This rule aims to empower the local classifications on those events that, probably, would not be heard by neighboring nodes.

Rule 3: If the output of the local neural network classifies an event whose equivalent level L_{eq} is typically high (such as *traffic, horn* or *siren*) or the event was provisionally tagged as *unknown* by Rule 1, then the assigned label is calculated depending on the probability values among the four different neighbouring nodes. Specifically, the final assigned label will be the one that presents the highest probability value—among all the nodes—in a category that typically has high L_{eq}.

For instance, if the local neural network (1) assigned a label with a probability smaller than 80%—hence the event would be provisionally classified as *unknown* by Rule 1—and (2) the highest probability of the classification results from the neighbouring nodes belonged to a category with high L_{eq}, then the local node would update the *unknown* label to the label assigned by its neighbours. On the contrary, if the class was *unknown* but the highest probability belonged to an event with low L_{eq}, then the final label would still remain *unknown*.

V.4 Experimental Evaluation

To validate the effectiveness of taking advantage of physical redundancy of sensors, we propose to use the classification algorithm (i.e., deep neural network and distributed consensus protocol) described above to classify acoustic events from a real-operation urban environment. In this regard, we have used the aforesaid data collected simultaneously in 4 close spots.

V.4.1 Experiment set up

In (Vidaña-Vila et al. 2020b), authors used exclusively the UrbanSound 8k (Salamon et al. 2014) dataset to evaluate the system under laboratory conditions. However, as the purpose of the dataset is to capture the *sounds* of New York, it lacks from the predominant class in the city centre of Barcelona (road traffic noise). To address this issue and enrich the training dataset, data from 3 datasets were used:

1. UrbanSound 8k (Salamon et al. 2014): This was the base dataset containing samples of acoustic events different from traffic noise. Authors decided to remove the classes *air conditioner* and *street music* as using these two classes considerably degraded the

Table V.1: System performance using physical redundancy.

	Non-*unknown* events (Experiment 1)	Classification accuracy (Experiment 2)
Before consensus	585 events	66.84 %
After consensus	846 events	91.16 %
Improvement	**+261 events**	**+24.32 %**

performance of the local classifier in the real-world scenario which is something that will be analyzed in the future to improve the generalization of the algorithm and the system.

2. BCNDataset (Vidaña-Vila et al. 2020a): Even though this dataset contains information from 14 categories, only the audio fragments labelled as *road traffic noise, brakes, horns, sirens,* and *people* were used to train the model.

3. Andorra Dataset (Alsina-Pagès et al. 2019): It was used to increase the number of *horn* and *siren* samples. No traffic noise was added from this dataset in order to avoid class imbalance.

Data from these three datasets was used to train the local classifier of each node. Hence the system was trained to classify the following categories: *traffic*, *dog*, *people*, *gun*, *jackhammer*, *drilling*, *horn* and *siren*. Note that there is also a class called *unknown* that is assigned to those events that the neural network is not able to classify (see Section V.3).

Finally, to test the classifier, one hour of manually annotated audio data from the recording campaign detailed in Section V.2 was used. Note that none of these real-world data (containing events not previously seen by the classifier such as bells, birds or wind, which will obviously degrade the system performance) was used in training. More precisely, when manually labelling the collected data, authors detected that only 769 audio fragments (out of 900 fragments) of 4 seconds contained information belonging to any of the classes of the training set. These fragments were used to build the *known* dataset.

V.4.2 Experimental Evaluation

The experimental evaluation is twofold. On the one hand, it first assesses up to what extent the distributed consensus protocol boosts the confidence of individual nodes. On the other hand, it later measures the advantages of using physical redundancy in terms of classification accuracy.

Experiment 1. A local analysis on the behaviour of the deep neural network (without taking advantage of physical redundancy of sensors) showed that only the 585 out of the 900 fragments of the dataset were given a label different from *unknown*, which means that in a 35 % of the samples, none of the classes showed a confidence greater than 0.8 (see Heuristic Rule 1 in Section V.3). However, as a result of applying the distributed consensus protocol (i,e., considering the labels and class probabilities of the four simultaneous recordings), the number

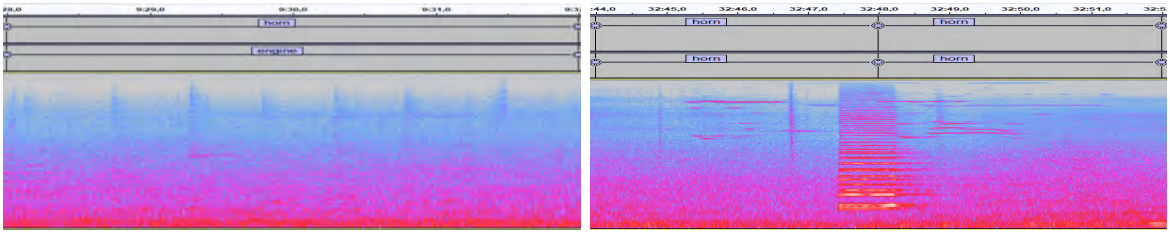


Figure V.2: Spectrogram of a 4-second sample manually labelled as *horn* that the system has been unable to classify (left). Spectrogram of two consecutive 4-seconds samples each manually labelled as *horn* that the system has classified correctly (right).

of non-*unknown* detected events rocketed to **846**. Therefore, using physical redundancy enables individual nodes to benefit from the decisions suggested by their neighbours.

Experiment 2. To test the capability of the system on identifying the events belonging to any of the categories of the training set, the 769 audio fragments of the *known* dataset were selected. For these known events, the accuracy before applying the distributed consensus protocol was 66.84 %, and raised up to **91.16 %** after it. Hence, it can be seen that using audio data from different close measuring points contributes to increase the overall classification accuracy of the proposed system. A summary of these results is shown in Table V.1.

It must be considered that most of the test data belongs to the *traffic* category, which is the predominant noise in the chosen place of the city of Barcelona. The classifier system is unable to recognise some of the events.

To find out the logic behind this behaviour, authors carefully listened to the raw audio data and observed their spectrograms. For instance, it was observed that those events that were miscategorized as *traffic* (i.e., *people, horn, siren*) actually contained traffic noise masking the labelled event. Fig. V.2 shows two *horn* events: one that was misclassified (left) and one that was properly classified (right). The wrongly labelled event is hard to identify by sight even for humans, as its salience is comparable to the rest of the acoustic events happening simultaneously. Actually, the horn event happens on the first second of the window, and when listening to it, the perception is that the noise source is far, probably in another street. We believe that this event would be successfully detected in another node when extending this system to a grid of sensors.

V.5 Discussion and Conclusions

This work presents a preliminary performance evaluation of a low-cost WASN with physical redundancy when classifying urban events from the crossroad of two crowded streets in the city centre of Barcelona. Three different corpora have been used to train the deep neural network running at each node, a set of classification heuristics to exploit physical redundancy has been defined, and four simultaneous recordings have been used to test the system in a real-operation environment.

Obtained results suggest that the distributed consensus protocol—aimed to take into consideration the classification results from neighboring nodes—increases the overall

classification confidence by reducing the number of non-*unknown* detected events. Also, physical redundancy may contribute to increase the classification accuracy for all the acoustic classes that exist in the training set. It is worth noting that, given the intrinsic real-world nature of the test data set, the obtained results may suffer the consequences of class imbalance. That is, not all the events had the same probability to occur. For instance, gun shots are highly unlikely in Barcelona. Indeed, when looking at the absolute accuracy results, it must be taken into account that most of the test data belongs to *traffic* category, which clearly unbalances the test corpora and makes the accuracy results to (over)shine. Rather, this work aims to consider the benefits of using the heuristics implemented on the distributed intelligence layer in order to improve the classification confidence (i.e., identifying known events).

After conducting an in-depth manual analysis, the system fails to properly classify some of the events because the same audio fragment contains multiple types of noise (e.g. *traffic*, which is mainly present in most of the samples).

The future work aims to extend the training dataset to avoid missing several types of events present in the city centre of Barcelona. Also, the system should be tested with acoustic samples from other hours of the day to glimpse all the variations in the soundscape of the city centre of Barcelona. To address the issue of classifying acoustic samples with several concurrent events, a multiclass classifier shall be considered.

Acknowledgements

Authors thank Gerard Ginovart for his help on the recording campaign and Secretaria d'Universitats i Recerca of the Department d'Empresa i Coneixement of the Generalitat de Catalunya for grants 2017-SGR-966 and 2017-SGR-977.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

References

Alsina-Pagès, Rosa Ma, Garcia Almazán, Robert, Vilella, Marc and Pons, Marc (2019). 'Noise events monitoring for urban and mobility planning in Andorra La Vella and Escaldes-engordany'. In: *Environments* vol. 6, no. 2, p. 24.

Su-bei, MENG (2007). 'Harm to human health from low frequency noise in city residential area [J]'. In: *China Medical Herald* vol. 35.

Bello, Juan P, Silva, Claudio, Nov, Oded, Dubois, R Luke, Arora, Anish, Salamon, Justin, Mydlarz, Charles and Doraiswamy, Harish (2019). 'Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution'. In: *Communications of the ACM* vol. 62, no. 2, pp. 68–77.

Bergadà, Pau and Alsina-Pagès, Rosa Ma (2019). 'An Approach to Frequency Selectivity in an Urban Environment by Means of Multi-Path Acoustic Channel Analysis'. In: *Sensors* vol. 19, no. 12, p. 2793.

Flindell, IH and Walker, JG (2004). 'Environmental noise management'. In: *Advanced Applications in Acoustics, Noise and Vibration*, p. 183.

Guski, Rainer, Schreckenberg, Dirk and Schuemer, Rudolf (2017). 'WHO environmental noise guidelines for the European region: A systematic review on environmental noise and annoyance'. In: *International journal of environmental research and public health* vol. 14, no. 12, p. 1539.

Hurtley, Charlotte (2009). *Night noise guidelines for Europe*. WHO Regional Office Europe.

Iandola, Forrest N., Moskewicz, Matthew W., Ashraf, Khalid, Han, Song, Dally, William J. and Keutzer, Kurt (2016). 'SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size'. In: *CoRR* vol. abs/1602.07360. arXiv: 1602.07360.

Mydlarz, Charlie, Sharma, Mohit, Lockerman, Yitzchak, Steers, Ben, Silva, Claudio and Bello, Juan Pablo (2019). 'The life of a New York City noise sensor network'. In: *Sensors* vol. 19, no. 6, p. 1415.

Office, Parliamentary Counsel's (2017). 'Protection of the Environment Operations (Noise Control) Regulation 2017'. In: *Legal Service Bull.* vol. 1, p. 44.

Pham, Congduc and Cousin, Philippe (2013). 'Streaming the sound of smart cities: Experimentations on the smartsantander test-bed'. In: *2013 IEEE international conference on green computing and communications and IEEE internet of things and IEEE cyber, physical and social computing*. IEEE, pp. 611–618.

Piper, Ben, Barham, Richard, Sheridan, Steven and Sotirakopoulos, Kostas (2017). 'Exploring the "big acoustic data" generated by an acoustic sensor network deployed at a crossrail construction site'. In: *Proceedings of the 24th International Congress on Sound and Vibration (ICSV), London, UK*, pp. 23–27.

Salamon, J., Jacoby, C. and Bello, J. P. (Nov. 2014). 'A Dataset and Taxonomy for Urban Sound Research'. In: *22nd ACM International Conference on Multimedia (ACM-MM'14)*. Orlando, FL, USA, pp. 1041–1044.

Sevillano, Xavier et al. (2016). 'DYNAMAP–Development of low cost sensors networks for real time noise mapping'. In: *Noise mapping* vol. 1, no. open-issue.

Test, Tsafnat, Canfi, Ayala, Eyal, Arnona, Shoam-Vardi, Ilana and Sheiner, Einat K (2011). 'The influence of hearing impairment on sleep quality among workers exposed to harmful noise'. In: *Sleep* vol. 34, no. 1, pp. 25–30.

Vidaña-Vila, Ester, Duboc, Leticia, Alsina-Pagès, Rosa Ma, Polls, Francesc and Vargas, Harold (2020a). 'BCNDataset: Description and Analysis of an Annotated Night Urban Leisure Sound Dataset'. In: *Sustainability* vol. 12, no. 19, p. 8140.

Vidaña-Vila, Ester, Navarro, Joan, Borda-Fortuny, Cristina, Stowell, Dan and Alsina-Pagès, Rosa Ma (2020b). 'Low-cost distributed acoustic sensor network for real-time urban sound monitoring'. In: *Electronics* vol. 9, no. 12, p. 2119.

Paper VI

Prototyping a low-cost Wireless Acoustic Sensor Network with physical redundancy to automatically classify acoustic events in urban environments

Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, Joan Navarro

Presented in *Poster presented at UrbanSound Symposium 2021.* Abstract published in *Proceedings*, May 2021, volume 72, DOI: 10.3390/proceedings2021072004

VI

Prototyping a Low-cost Wireless Acoustic Sensor Network with Physical Redundancy to Automatically Classify Acoustic Events in Urban Environments

{Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, Joan Navarro} La Salle - Universitat Ramon Llull
{ester.vidana, rosamaria.alsina, joan.navarro}@salle.url.edu

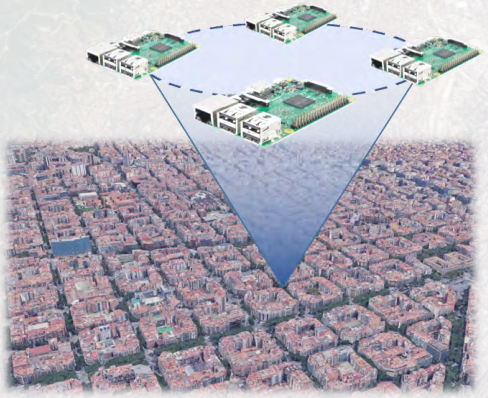
laSalle UNIVERSITAT RAMON LLULL

Summary

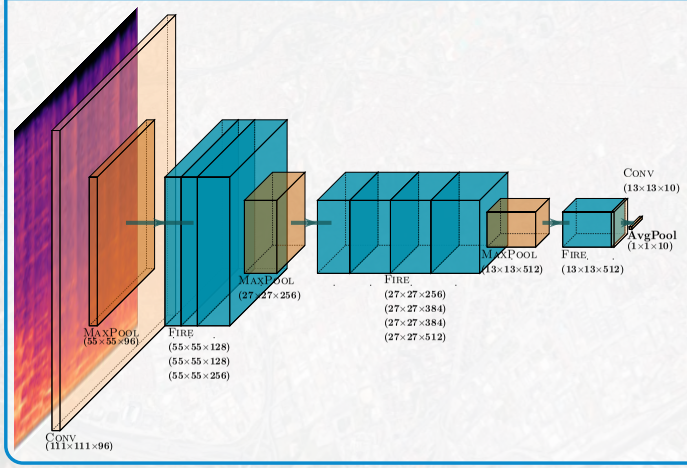
Each different **noise source** in an urban environment has a different impact on **citizens' well-being**. Typically, the **platforms** used to **monitor** the noise level on cities are composed of **expensive sensors** placed in specific locations, which **limits their coverage area**. This work pursues an alternative approach and proposes to:

1. Use **low-cost devices for sensing and computing** (i.e., Raspberry Pi Model 2B and USB omnidirectional microphones).
2. **Deploy** them in a topology that takes advantage of **physical redundancy** (i.e., multiple nodes can monitor the same acoustic event simultaneously).

To conduct urban sound monitoring, a **distributed classification algorithm** runs in each node. The algorithm consists of a (1) **deep neural network** and (2) **distributed consensus protocol** that merges the classification results of neighbouring nodes and outputs a final label to the events occurring at real-time.



Classification: Deep Neural Network & Distributed Consensus Protocol



The event classification is as follows:

1. Each node (Raspberry Pi) computes the **spectrogram of a 4-seconds** acoustic window.
2. The spectrogram is fed to a **deep neural network** with a SqueezeNet topology **pre-trained** with the UrbanSound 8K dataset and traffic noise samples from Barcelona.
3. The deep net computes a 10-components (one per category) vector with the occurrence **probabilities of each event**.
4. A distributed consensus protocol **shares** the event with maximum local probability with the **neighboring nodes**.
5. Once all the nodes have received the events heard by their closest neighbors respectively, the events (and their probabilities) identified from the neighbors are **compared** against the local classification and a **final label** is assigned to the 4-seconds window.

Recording Campaign

To **test** the system prototype with **real-operation data**, a recording campaign was conducted. **Four** Zoom H5 recorders were deployed in a crossroad of the city center of Barcelona to **simultaneously collect acoustic data with physical redundancy**.
Those data were used to verify the performance improvement—in terms of classification accuracy—of the proposed system.

Conclusions & Long-Term Goal

Initially, the system was **tested** using the **UrbanSound 8k dataset** under laboratory conditions and, in this contribution, it has been **validated with data from real-operation simultaneous recordings** obtained from the recording campaign on the city centre of Barcelona. With these **new data** collected simultaneously from four physically close spots, the proposed distributed consensus protocol **improves** the acoustic classification accuracy **by a +24%** compared to the classification done by a single device. Experiment results validate the feasibility of the proposal and justify the **usage of physical redundancy to improve the classification accuracy of acoustic events**.
The long-term objective of the work is to physically **implement the proposed approach** in the Eixample district on the **city centre of Barcelona** to be able to identify different noise sources, which would allow to **objectively survey the acoustic comfort** of the citizens in the street.

Paper VII

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

Ester Vidaña-Vila, Dan Stowell, Joan Navarro, Rosa Ma Alsina-Pagès

Poster presented at Deep Learning Barcelona Symposium 2021.

VII

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
[2] Department of Cognitive Sciences & Artificial Intelligence, Tilburg University, Netherlands
[3] GRITS - Grup de Recerca en Internet Technologies and Storage, La Salle – Universitat Ramon Llull.

Objectives

Conceive a system able to detect and classify, in real-time, a predefined set of urban acoustic events that may occur simultaneously by means of:

- Training a multi-label deep-learning-based algorithm.
- Testing the system with real-world collected data.
- Running the system over a Wireless Acoustic Sensors Network.

Context

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. 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. 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. Therefore, identifying the sources of those potentially harmful sounds has emerged as a hot research topic nowadays.
Moreover, Barcelona is one of the noisiest cities Europe. Actually, Barcelona has been categorized as the seventh noisiest city in the world. For this reason, it has been selected as the use-case scenario for this project.

Data gathering

To be able to evaluate the system with real-world data, two recording campaigns were carried out:

- Autumn campaign: was recorded on the 17 November 2020 from 12:00 to 14:30 when there were COVID-19 mobility restrictions.
- Spring campaign: was recorded on the 31 May 2021 from 15:30 to 18:00 (Mobility restrictions were softened).

Data were manually labelled in frames of 4-seconds of duration taking into account events occurring simultaneously (i.e., polyphonic labelling). Once it was labelled, it was divided into Train/Validation/Test splits. Concretely, the division was done into contiguous regions of 5—71 minutes length.

Feature extraction

As features, spectrograms were used. More concretely, each 4-second duration audio fragment was transformed into a spectrogram to train a Deep Neural Network (MobileNet).

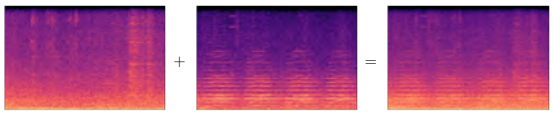


Figure 1: Example of mix-up data augmentation using two random 4-second fragments containing several acoustic events.

Experimental evaluation

Four experiments were carried out using data augmentation techniques with different datasets: BCNDataset, UrbanSound 8K dataset.
Experiment 0: No data augmentation.
Experiment 1: Data augmentation using data from BCNDataset to balance the classes.
Experiment 2: Data augmentation using BCNDataset + UrbanSound 8k dataset.
Experiment 3: Same as experiment 2 but using more instances of augmented data.
In all the experiments, the data augmentation technique used was audio mixing (mix-up augmentation).
All the spectrograms of the dataset were normalized.

Labelled data

Table 1: Number of events manually annotated on the dataset.

Label	1st Campaign	2nd Campaign	TOTAL
rtn	2177	2118	4295
peop	300	612	912
brak	489	424	913
bird	357	960	1317
motorc	769	565	1334
eng	203	913	1116
cdoor	133	161	294
impls	445	170	615
cmplx	85	73	158
troll	162	152	314
wind	8	23	31
horn	43	33	76
sire	18	57	75
musi	8	30	38
bike	51	24	75
hdoor	25	60	85
bell	24	27	51
glass	17	32	49
beep	31	0	31
dog	3	25	28
drill	0	14	14

Results

Table 2: Macro and micro average F-1 scores.

Experiment	F1- Macro Average
Experiment 0	12%
Experiment 1	39%
Experiment 2	36%
Experiment 3	33%

Experiment	F1- Micro Average
Experiment 0	46%
Experiment 1	70%
Experiment 2	75%
Experiment 3	67%

Whereas in F1- Macro Average all the classes have the same importance, in F1- Micro Average the most populated classes have more importance.
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.

Conclusions

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 polyphonic real-world data.
The system has a good performance when classifying events with more than 100 instances on the Validation and Test set. However, it behaves poorly when classifying those classes with few instances except for the *bell* event.

Acknowledgements

We would like to thank Gerard Ginovart for his valuable assistance on the recording campaign in both seasons.

Contact Information

- Email: ester.vidana@salle.url.edu
- Web: https://www.salleurl.edu/es/signal-processing

Paper VIII

A Two-Stage Approach To Automatically Detect and Classify Woodpecker (Fam. *Picidae*) Sounds

Ester Vidaña-Vila, Joan Navarro, Rosa Ma Alsina-Pagès, Álvaro Ramírez

Published in *Applied Acoustics*, 2020, volume 166, pp. 107312. DOI: 10.1016/j.apacoust.2020.107312.

Abstract

Inventorying and monitoring which bird species inhabit a specific area give rich and reliable information regarding its conservation status and other meaningful biological parameters. Typically, this surveying process is carried out manually by ornithologists and birdwatchers who spend long periods of time in the areas of interest trying to identify which species occur. Such methodology is based on the experts' own knowledge, experience, visualization and hearing skills, which results in an expensive, subjective and error prone process. The purpose of this paper is to present a computing friendly system able to automatically detect and classify woodpecker acoustic signals from a real-world environment. More specifically, the proposed architecture features a two-stage Learning Classifier System that uses (1) Mel Frequency Cepstral Coefficients and Zero Crossing Rate to detect bird sounds over environmental noise, and (2) Linear Predictive Cepstral Coefficients, Perceptual Linear Predictive Coefficients and Mel Frequency Cepstral Coefficients to identify the bird species and sound type (i.e., vocal sounds such as advertising calls, excitement calls, call notes and drumming events) associated to that bird sound. Conducted experiments over a data set of the known woodpeckers species belonging to the *Picidae* family that live in the Iberian peninsula have resulted in an overall accuracy of 94,02%, which endorses the feasibility of this proposal and encourage practitioners to work toward this direction.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

L'interval de pàgines 214-236 s'ha eliminat d'aquesta versió per motius de copyright.

Paper IX

Analysis of the Noise Impact of the Airport of Barcelona to the Llobregat Delta Natural Environment during the 2021 Lockdown period

Júlia Blanch, Ester Vidaña-Vila, Rosa Ma Alsina-Pagès

Published in *8th Electronic Conference on Sensors and Applications*, November 2021. DOI: doi:10.3390/ecsa-8-11267.

Abstract

The noise caused by airports and its impact on human health, together with train, road traffic, leisure and wind nose has been widely analyzed, even in the reports published in 2019 by the WHO. Noise effect has also been studied in the literature on other species, such as birds and amphibians. In this work, we focus on a natural environment of special singularity due to its location: the natural space of the Delta del Llobregat, next to the city of Barcelona. Placed in an area close to the Port of Barcelona, and right on the way out of the planes taking off at Barcelona airport. In this paper, we present a first analysis of the typology of the sounds found in the natural environment of the Delta del Llobregat after conducting a simultaneous recording campaign at three separate spots of biological interest, determined by the park's curators. We identify the interfering sounds, as well as the amount of wildlife sounds in relation to the noises caused by the airport activity. The recordings and posterior analysis were made on March 5, 2021, when airport activity was still greatly diminished by the mobility restrictions. Also, we apply machine learning techniques to classify the acoustic events produced by both airport activity and wildlife aiming to build an automatic system that would allow to gather labelled data in future works.

IX.1 Introduction

The effect of aircraft noise on humans, among other noise polluters, has been widely studied and analyzed over the last few decades (Hurtley 2009). Fewer studies have also analyzed the impact of those sounds on wildlife (e.g. birds or amphibians), and despite the well-known consequences that noise can have on animals such as reproductive or long-term survival

problems, there are still natural parks over-exposed to sounds produced by humans (Radle 2007). In this work, we aim to collect audio files and analyze the soundscape of the *Delta del Llobregat* natural park, which is a Protected Area (PA) located next to the city center of Barcelona. Concretely, the selected location is surrounded by the Port of Barcelona and the *Josep Tarradelles Barcelona-El Prat* airport in Spain. Using acoustic data gathered at three spots of biological interest inside the natural park, we aim to train a machine learning model able to classify real-world acoustic events that would allow researchers to easily obtain more data to find patterns in the behaviour of wildlife in the selected areas.

Similar works have been conducted in other natural parks located close to noisy areas such as airports. For example, in (Alquezar and Macedo 2019), the overlap between natural areas and Brazilian airports is studied. Moreover, the legislation of different countries regarding the location of airports near protected areas are analyzed, and measures to mitigate the impact of aircrafts to wildlife are proposed. Another example can be found on (Radle 2007). In that work, A.L. Radle focuses on the impact of noise to wildlife on different ecosystems (e.g. terrestrial wildlife, marine wildlife or noise in national parks). Similarly, in (Iglesias-Merchan et al. 2015), C. Iglesias-Merchan *et al.* evaluate the impact of aircraft noise in a protected area in the Central Mountains of Spain. Finally, on (Alquezar et al. 2020), mist-nets and sound automatic recording units are used to classify bird species near natural areas close to different Brazilian airports. They evaluate several biodiversity indexes and identify airport avoider bird species and airport adapter bird species. Results show that, in quieter locations, the abundance of different bird species is, indeed, richer.

The work presented in this paper exposes the results of a manually labelled recording campaign carried out in the Delta del Llobregat protected area. Concretely, three simultaneous recordings of 2 hours of duration have resulted in acoustic events from 14 different categories: some of them produced by humans and some others produced by the environment wildlife. Then, the classification results of three different machine learning algorithms trained and tested over the collected dataset are compared. The reason to apply machine learning techniques over the recorded data comes from the idea that automatically classifying the acoustic events present on the soundscape of the selected location would allow to automatically have more data that could be used to analyze over time the impact of the airport sound over the bird species inhabiting the protected area.

The remainder of this paper is organized as follows: first, Section IX.2 explains the methodology carried out to gather data in three different spots. Section IX.3 details the analysis conducted to the designed dataset after labelling it. Then, Section IX.4 reports the classification algorithms trained with our data and compares their results. Finally, Section IX.5 closes the paper and proposes some future work.

IX.2 Airport Recording Campaign

Once the recording points were decided, and having requested prior permission from the consortium for protection and management of the natural spaces of the Delta del Llobregat,

we planned a recording campaign for 5th of March 2021. At that date, due to the COVID-19 pandemic, take-offs and landings of flights were happening more or less with a frequency of a flight every 15 or 20 minutes.

The recording equipment required was: *i)* tripod, *ii)* Zoom H5 Recorder, *iii)* pen and writing support and *iv)* data collection sheet (see Figure IX.1). The three recorders were synchronized with 3 hand claps. Later on, the three recorders were separated and placed at their final designed locations, and after finishing the recording setup, all the technicians started the annotations in the data collection sheets. At the end of the recordings, another synchronization was conducted, to be able to adjust the data stamp if the three clocks were not precisely synchronous. The three recordings lasted for 2 hours, starting at 16:20 in the afternoon. The distance between the three chosen locations was around 500m.

Figure IX.1: Locations of the three recordings in Delta del Llobregat.

IX.3 Data analysis

After the recording campaign, an exhaustive analysis was conducted over the data. Firstly, a manual labelling process was carried out using Audacity (open-source software for audio and recordings treatment that can be downloaded for free at https://www.audacityteam.org/). The volume of the acoustic events detected is the one represented in Table IX.1.

Regarding the feature extraction process, the following parameters were obtained for each of the acoustic events: (1) Mel Frequency Cepstral Coefficients (MFCC), which represent the short-term power spectrum of a sound (Mermelstein 1976), (2) the Spectral Centroid, used in digital signal processing to characterise a spectrum, (3) the Spectral Roll-Off, and (4) the Zero Crossing Rate.

As shown on Figure IX.2, not all the categories have the same duration. Therefore, the average length of all the categories was used to split the events into windows of the same duration (0.94 seconds). Considering those divisions, the dataset was created in a way that all the audio slices belonging to the same acoustic event (e.g. an aircraft passing by), were placed only on the training set or the testing set. Finally, the 80% of the audio fragments were used for training and the remaining 20% were used for testing.

Category	description	Number of events
1 - airp	Airplanes taking off or landing	33
2 - alarm	Alarms from the airport or surrounding states	81
3 - animals	Sounds produced by animals	188
4 - bicy	Bicycles	6
5 - bird	Single bird vocalizations	5726
6 - birds	Multiple bird vocalizations	1493
7 - complex	Unidentified sounds	79
8 - duck	Duck vocalizations	1437
9 - flutter	Ducks moving their wings	2
10 - nature	Leaves from trees moved by the wind	19
11 - peop	People talking	91
12 - rtn	Road traffic noise	32
13 - water	Water sound	13
14 - wind	Wind sound	2

Table IX.1: Number of events for each of the categories of the labelled dataset.

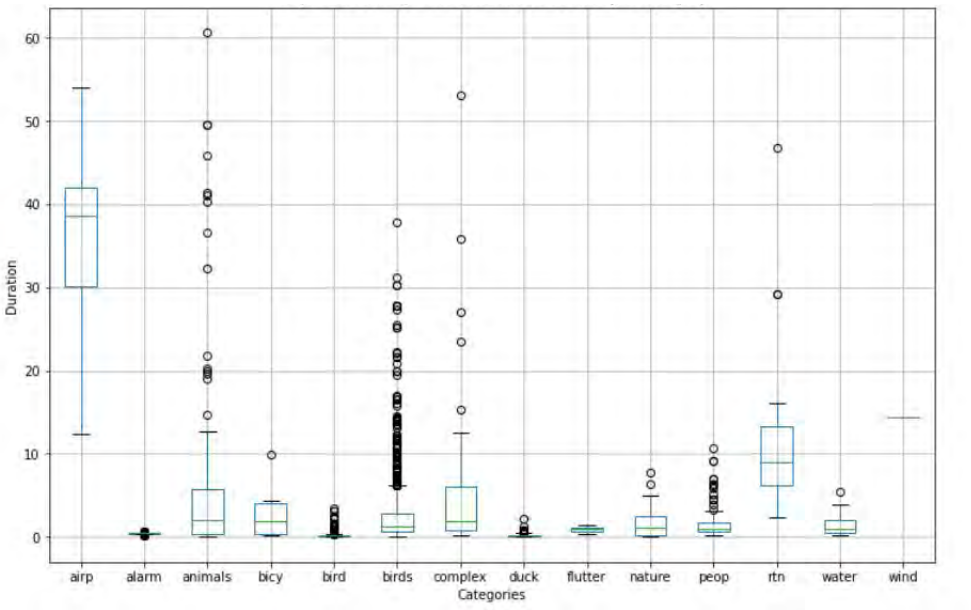


Figure IX.2: Boxplot of average duration time of events per category.

IX.4 Classification algorithm

Several machine learning algorithms have been tested to automate the acoustical detection of events. The accuracy given for each model is evidenced on Table IX.2.

IX.4.1 k-Nearest Neighbor

k-NN has given efficient results for acoustic event detection in other fields (Hoyos-Barceló et al. 2017; Liu et al. 2010). A grid search was performed to check what number of neighbors results in the best accuracy value. Finally, the best result (accuracy value of 53,5%) was obtained when using a value of k = 6.

Usually, the sounds produced by airplanes (*airp* category) are confused with *complex*

Algorithms	Accuracy
k-NN	53,5%
Decision Tree	51,3%
Random Forest + Bagging	68,8%
SVM (kernel: polynomial)	38,6%
SVM (kernel: sigmoid)	15,0%
SVM (kernel: RBF)	83,2%

Table IX.2: Accuracy value for the tested algorithms.

sounds. As the *complex* category contains acoustic information that we could not identify in the labelling process, it is possible that some of the windows from that category contain fragments of airplane sounds. Also, some *alarm* events are confused with the *rtn* category. Since transit sound is continuous in background on almost all of the recordings, some events catalogued as *alarm* could contain also *rtn* background noise. Finally, the algorithm tends to confuse the categories *bird* and *birds*, which means that it is unable to differentiate the number of birds present on a concrete window.

IX.4.2 Decision Tree

The model created with a decision tree is designed with a maximum profundity (largest way from the root node to the leaf node) of 6, since it is the one that results in a higher accuracy (of 51,3%). Again, maximum profundity was chosen after conducting a grid search.

In this case, the decision tree model shows that categories *airp* and *peop* have clear patterns, and hence there is no confusion identified on these events. Alternatively, all the categories related to animals (*animals , bird, birds, duck*) are often confused. Also, the fragments belonging to categories with the poorest samples (*alarm, bicy, complex, flutter, nature, transit, water or wind*) are the ones that result in the worst classification results.

IX.4.3 Random Forest

Random Forest has already been used in other research projects of acoustic events detection and classification (Phan et al. 2014). To design our concrete model, we have conducted a grid search varying the maximum depths parameter. We found that the best performance of the model was achieved for max_depth = 48, with an accuracy of 68,6%. Then, Bagging, Boosting and Voting methods were applied to try to increase the accuracy of the model, and after applying Bagging the accuracy raised to 68,8%.

The accuracy obtained in this algorithm is the best one so far, but it is also important to study the weaknesses of the model by means of analyzing the events that it confuses the most. With this classifier, the algorithm confuses sporadic events of all the categories. However, some patterns can be identified again. The algorithm tends to confuse the *rtn* and *airp* categories, and the *bird* with *birds*, which proves that it is not able to identify the number of birds vocalizing simultaneously. Something remarkable that has not happened on other algorithms is that some *birds* events are confused with the *peop* category, showing that it

confuses bird vocalizations with human voice. The reason behind this confusion may be that there are some similarities between bird vocalizations and human sounds, as stated in some studies (Doupe and Kuhl 1999).

IX.4.4 Support Vector Machine

One of the most widely used methods for the classification of sound events is the Support Vector Machine (SVM). In this work, the Radial Basis Function (RBF) kernel (Vavrek et al. 2010) has proven to be the one that obtains the best classification results out of 4 (linear, sigmoid, polynomial and RBF). The linear kernel never converged and therefore, there are no results to present. When using the sigmoid kernel, the classifier was able to identify correctly only the following categories: *animals, bird* and *duck*, and obtained poor results for the other ones. With the polynomial kernel, the classifier tended to classify events from other categories as *bird*. Finally, RBF kernel obtained the best results among all the classifiers presented in this work. The obtained confusion matrix of the algorithm when using the RBF kernel can be seen on Figure IX.3. This kernel results in an accuracy of 83,2%. On the confusion matrix, it can be seen how the system is able to classify the categories: *airp, animals, bird, birds*, and *duck*, respectively.

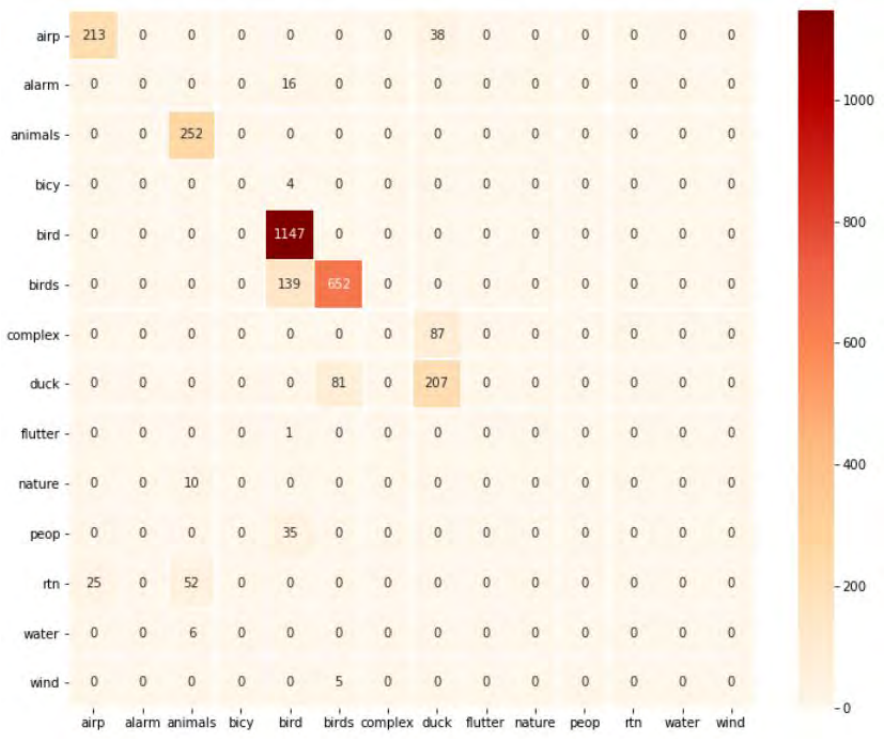


Figure IX.3: Confusion matrix of the SVM algorithm.

IX.5 Conclusions

After analyzing the machine learning results, it has been detected that, in general, all of the implemented algorithms have confusion patterns over different classes. The main reason

for it might be the lack of data from some of the categories in the training set. The most common confusion happens between the *bird* and *birds* categories, which may be caused by the splitting of different windows of the same acoustic event in different fragments, and due to the similarity of the spectrum of both signals. As the window is usually shorter than the duration of the *birds* event, it may have happened that some of the windows of the labelled event contained only information of a single vocalization. This fact was not considered when dataset was created.

To improve the current results, in future work, a wider recording campaign should be done. This would probably allow the algorithms to create more accurate patterns for detection, hence resulting in a more efficient model with better classification results.

Author's contributions

Conceptualization, RM.A-P and E.V-V; software, J.B; validation, all authors; formal analysis, J.B; investigation, all authors; data curation, J.B and E.V; writing-original draft preparation, review and editing, all authors; visualization, J.B ; supervision, E.V and RM.A-P; project administration and funding acquisition, RM.A-P.

Funding

We would like to thank Secretaria d'Universitats i Recerca of the Department d'Empresa i Coneixement of the Generalitat de Catalunya for the grant 2017-SGR-966. The research that has concluded in these results has been carried out thanks to funds from the Secretariat of Research and Universities of the Generalitat de Catalunya and the Ramon Llull University, thanks to the project *Sons al Balcó*, code 2021-URL-Proj-053.

Acknowledgements

We would like to thank El Delta del Llobregat Natural Park for their valuable assistance on selecting the most convenient locations during the recording campaign.

Conflict of interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KNN K-Nearest Neighbor
LD Linear dichroism
MFCC Mel Frequency Cepstral Coefficients
PA Protected Area
RBF Radial Basis Function
SVM Support Vector Machine
WHO World Health Organisation

References

Alquezar, Renata D and Macedo, Regina H (2019). 'Airport noise and wildlife conservation: What are we missing?' In: *Perspectives in Ecology and Conservation* vol. 17, no. 4, pp. 163–171.

Alquezar, Renata D, Tolesano-Pascoli, Graziela, Gil, Diego and Macedo, Regina H (2020). 'Avian biotic homogenization driven by airport-affected environments.' In: *Urban Ecosystems* vol. 23, no. 3.

Doupe, Allison J and Kuhl, Patricia K (1999). 'Birdsong and human speech: common themes and mechanisms'. In: *Annual review of neuroscience* vol. 22, no. 1, pp. 567–631.

Hoyos-Barceló, Carlos, Monge-Álvarez, Jesús, Shakir, Muhammad Zeeshan, Alcaraz-Calero, Jose-María and Casaseca-de-La-Higuera, Pablo (2017). 'Efficient k-NN implementation for real-time detection of cough events in smartphones'. In: *IEEE journal of biomedical and health informatics* vol. 22, no. 5, pp. 1662–1671.

Hurtley, Charlotte (2009). *Night noise guidelines for Europe*. WHO Regional Office Europe.

Iglesias-Merchan, Carlos, Diaz-Balteiro, Luis and Soliño, Mario (2015). 'Transportation planning and quiet natural areas preservation: Aircraft overflights noise assessment in a National Park'. In: *Transportation Research Part D: Transport and Environment* vol. 41, pp. 1–12.

Liu, Chien-Liang, Lee, Chia-Hoang and Lin, Ping-Min (2010). 'A fall detection system using k-nearest neighbor classifier'. In: *Expert systems with applications* vol. 37, no. 10, pp. 7174–7181.

Mermelstein, Paul (1976). 'Distance measures for speech recognition, psychological and instrumental'. In: *Pattern recognition and artificial intelligence* vol. 116, pp. 374–388.

Phan, Huy, Maaß, Marco, Mazur, Radoslaw and Mertins, Alfred (2014). 'Random regression forests for acoustic event detection and classification'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 23, no. 1, pp. 20–31.

Radle, Autumn Lyn (2007). 'The effect of noise on wildlife: a literature review'. In: *World Forum for Acoustic Ecology Online Reader*, pp. 1–16.

Vavrek, Jozef, Pleva, Matúš and Juhar, Jozef (2010). 'Acoustic events detection with support vector machines'. In: *Electrical Engineering and Informatics, Proceeding of the Faculty of Electrical Engineering and Informatics of the Technical University of Košice*, pp. 796–801.