

# List of Publications

## ARTICLE 1



**Title:** Energy-Efficient Acoustic Violence Detector for Smart Cities

**Authors:** Marta Bautista-Durán, Joaquín García-Gómez, Roberto Gil-Pita, Inma Mohino-Herranz, Manuel Rosa-Zurera

**Journal:** International Journal of Computational Intelligence Systems (IJCIS)

- Year: 2017
- Volume: 10. Issue: 1
- Pages: 332-349
- ISSN (online): 1875-6883. ISSN (print): 1875-6891
- License: Open Access under the CC BY-NC license
- D.O.I.: <https://doi.org/10.2991/ijcis.10.1.89>

**Ranking:**

- JCR (2017): 2
- Quartile Rank - Computer Science, Artificial Intelligence: 54/132 (Q2).
- Quartile Rank - Computer Science, Interdisciplinary Applications: 49/105 (Q2).

**Contribution to the article:** The doctoral student has contributed to the whole development of this publication, including the conceptualization of the problem, methodology, creation of the dataset, software programming, analysis of the results, and writing.

## ARTICLE 2



**Title:** Cost-constrained Drone Presence Detection through Smart Sound Processing

**Authors:** Joaquín García-Gómez, Marta Bautista-Durán, Roberto Gil-Pita, Inma Mohino-Herranz, Miguel Aguilar-Ortega, César Clares-Crespo

**Book:** Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM)

- Year: 2019
- Volume: 1
- Pages: 766-772
- ISBN: 978-989-758-351-3
- D.O.I.: 10.5220/0007556007660772

**Contribution to the article:** The doctoral student has contributed to the whole development of this publication, including the conceptualization of the problem, methodology, creation of the dataset, software programming, analysis of the results, and writing.

## ARTICLE 3



**Title:** Analysis of the performance of Evolved Frequency Log-Energy Coefficients in Hearing Aids for different Cost Constraints and Scenarios

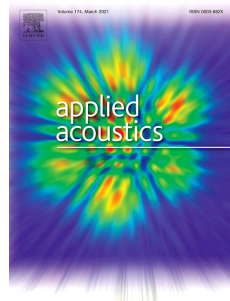
**Authors:** Joaquín García-Gómez, Inma Mohino-Herranz, César Clares-Crespo, Alfredo Fernández Toloba, Roberto Gil-Pita

**Library:** AES E-Library. 145th AES Convention

- Year: 2018
- Paper number: 10111
- ISBN: 978-1-942220-25-1
- D.O.I.: <https://doi.org/10.17743/aesconv.2018.978-1-942220-25-1>

**Contribution to the article:** The doctoral student has contributed to the whole development of this publication, including the conceptualization of the problem, methodology, software programming, analysis of the results, and writing.

## ARTICLE 4



**Title:** Linear detector and neural networks in cascade for voice activity detection in hearing aids

**Authors:** Joaquín García-Gómez, Roberto Gil-Pita, Miguel Aguilar-Ortega, Manuel Utrilla-Manso, Manuel Rosa-Zurera, Inma Mohino-Herranz

**Journal:** Applied Acoustics

- Year: 2021
- Volume: 175
- Article number: 107832
- License: Open Access under the CC BY-NC-ND license
- D.O.I.: <https://doi.org/10.1016/j.apacoust.2020.107832>

### **Ranking:**

- JCR (2019): 2.44
- Quartile Rank - Computer Science, Artificial Intelligence: 9/32 (Q2).
- Quartile Rank - Acoustics: 49/105 (Q2).

**Contribution to the article:** The doctoral student has contributed to the whole development of this publication, including the conceptualization of the problem, methodology, software programming, analysis of the results, and writing.

## ARTICLE 5



**Title:** Smart Sound Processing for Defect Sizing in Pipelines Using EMAT Actuator Based Multi-Frequency Lamb Waves

**Authors:** Joaquín García-Gómez, Roberto Gil-Pita, Manuel Rosa-Zurera, Antonio Romero-Camacho, Jesús Antonio Jiménez-Garrido, Víctor García-Benavides

**Journal:** Sensors. Special issue “State-of-the-Art Sensors Technology in Spain 2017”

- Year: 2018
- Volume: 18. Number: 3
- Article number: 802
- Pages: 1298-1305
- ISBN (pdf): 978-3-03842-960-9. ISBN (pbk): 978-3-03842-959-3
- License: Open Access under the CC BY-NC-ND license
- D.O.I.: <https://doi.org/10.3390/s18030802>

### Ranking:

- JCR (2018): 3.076
- Quartile Rank - Instruments & Instrumentation: 13/61 (Q1)
- Quartile Rank - Physics, Applied: 42/148 (Q2)
- Quartile Rank - Engineering, Electrical & Electronic: 87/266 (Q2)

**Contribution to the article:** The doctoral student has contributed to the whole development of this publication, including the conceptualization of the problem, methodology, taking of measurements, software programming, analysis of the results, and writing.

## DOCTORAL STUDENT DECLARATION

The doctoral student Joaquín García-Gómez states that the publications which support this thesis have not been used previously by other researchers as an endorsement of other compendium doctoral theses.

Alcalá de Henares, 5 de marzo de 2021.

Fdo. D. Joaquín García Gómez

# Abstract

Smart cities are places that try to implement new technologies and ideas in a sustainable and intelligent way to obtain benefits in a wide range of areas, focusing on creating social improvements, economic growth and new opportunities. In this thesis, four applications that can contribute to the improvement of the quality of life of people and must be present in this kind of spaces are researched: violent situation detection, drone presence detection, voice activity detection in hearing aids and pipeline defect assessment. All of them can help in solving issues related to public security, welfare, social inclusion and natural resources management, among others.

To develop these applications, different types of data can be obtained from the cities, including audio, video, radar or radio-frequency signals. Acoustic signals are a rich source of study due to the large amount of information they provide about the environments that surround us, and for that reason they have been considered in this thesis. Furthermore, the advantages of microphones compared to other devices like video cameras are numerous, such as their smaller size, consumption and price, their tolerance to adverse environmental conditions, or their capability to provide an omnidirectional sensing.

In this thesis, machine learning techniques are developed to detect different sound events in those signals. A typical pattern recognition scheme is presented in all the systems, including feature extraction, feature selection and detection stages. These processes are restricted in terms of computational cost, since the number of operations carried out in a microprocessor is directly related to the consumption of the device, and we want the systems to work autonomously to the extent possible. For this reason, and as massive datasets are not generally available in these issues, more complex techniques such as deep learning have been avoided.

Promising results are obtained along the thesis, and we can conclude that it is possible to apply computationally constrained sound event detection techniques to the four applications mentioned above, reaching a balance between consumption and performance. Furthermore, additional optimization techniques based on cascade-detectors seem to be useful when dealing with very restrictive devices such as hearing aids.

# Resumen

Las ciudades inteligentes son lugares en que se tratan de implementar nuevas tecnologías e ideas de manera sostenible e ingeniosa, con el objetivo de conseguir mejoras en una gran variedad de ámbitos, destacando especialmente la consecución de mejoras sociales, crecimiento económico y nuevas oportunidades. En esta tesis se han investigado cuatro aplicaciones que pueden ayudar a mejorar la calidad de vida de las personas y que deberían estar presentes en este tipo de lugares: la detección de situaciones violentas, la detección de presencia de drones, la detección de actividad vocal en audífonos y el análisis de defectos en tuberías. Todas ellas pueden contribuir a resolver problemas relacionados con la seguridad pública, el bienestar, la inclusión social y la gestión de recursos naturales, entre otros.

Existen gran variedad de datos presentes en las ciudades que pueden ayudar a desarrollar estas aplicaciones, como señales de audio, vídeo, radar o radiofrecuencia. Las señales acústicas son una valiosa fuente de estudio, ya que proporcionan una gran cantidad de información acerca de los entornos que nos rodean, y por esta razón han sido consideradas en esta tesis. Además, son muchas las ventajas de los micrófonos en comparación con otro tipo de dispositivos como las videocámaras: tienen un tamaño, consumo y precio menores, poseen una tolerancia mayor a condiciones ambientales adversas, y permiten grabar de forma omnidireccional.

En esta tesis se han desarrollado técnicas de aprendizaje automático para detectar eventos sonoros en las señales. En todas las aplicaciones se ha implementado un sistema de reconocimiento de patrones que incluye las fases de extracción de características, selección de las mismas y detección. Estos procedimientos se han restringido en cuanto a coste computacional, ya que el número de operaciones que lleva a cabo un microprocesador se encuentra directamente relacionado con el consumo del dispositivo, y se desea desarrollar sistemas que trabajen de forma autónoma en la medida de lo posible. Por esta razón, y dado que generalmente no existen bases de datos extensas en estos campos, se ha evitado el uso de técnicas más complejas como el aprendizaje profundo.

A lo largo de la tesis se han obtenido resultados satisfactorios, y se puede por tanto afirmar que es posible aplicar técnicas de detección de eventos acústicos restringidas en términos computacionales a las aplicaciones mencionadas anteriormente, alcanzando un equilibrio entre consumo y rendimiento. Además, la aplicación de técnicas de optimización adicionales basadas en detectores en cascada ha demostrado ser útil en un dispositivo final restrictivo, como es el caso de un audífono.



*A mis padres, Isabel y Juan, y a mi hermano Juan Antonio.  
Gracias por estar siempre.*

*“La felicidad puede hallarse hasta en los más oscuros momentos,  
si somos capaces de usar bien la luz”.*

Albus Dumbledore

## Agradecimiento

*Se cumplen cuatro años desde que decidí matricularme y ampliar mi formación universitaria sacando adelante un doctorado. Cuatro años que literalmente han tenido de todo, incluyendo una pandemia mundial que ha puesto patas arriba el mundo tal y como lo conocíamos. Ahora este periodo llega a su fin con la entrega y defensa de mi tesis doctoral.*

*Echando la vista atrás a esta etapa en la Universidad de Alcalá, quiero dar las gracias a Roberto Gil, por haberme iluminado tanto y tan bien con sus conocimientos y destrezas a lo largo de estos años. Del mismo modo, quiero agradecer a Manuel Rosa su sabiduría y preocupación por cómo iban las cosas, y a Manuel Utrilla, por su apoyo e interés. Por supuesto, agradezco a Stephan Chalup, mi tutor durante la estancia en Australia, y a Ali Bakhshi, mi compañero de laboratorio allí, su acogida y soporte para que me sintiera parte del grupo al otro lado del mundo.*

*Estos años en el grupo de investigación no hubieran sido lo mismo sin los compañeros y ya amigos con los que he compartido horas en el fondo, pero también muchas risas y alguna que otra quedada fuera. Gracias especialmente a Inma, que ha sido un gran apoyo y con la que he tenido mucha complicidad desde el principio, a Marta, que fue la persona con la que empecé este proyecto, a Cosme, que era (y es) un auténtico crack, y a Freddy, que ha demostrado ser un gran compañero de fondo y amigo. Gracias también a Miguel y a Fausto, por esos cafés tan necesarios donde intentábamos arreglar el mundo, y a todos los demás compañeros que han estado conmigo: Héctor, César, Alfredo...*

*En lo personal, gracias en primer lugar a mi familia, por estar siempre ahí: a mi madre, a mi padre y a mi hermano, que son los tres pilares de mi vida, y espero que no dejen de serlo nunca. Gracias por vuestra paciencia, por vuestros consejos, por saber perfectamente cuándo algo va mal, y por entenderme tan bien desde que recuerdo. A mis amigas y amigos de Azuqueca, de la uni y de Australia, gracias por estar, por sacarme una y mil sonrisas y por compartir juntos los mejores planes (que estoy seguro de que pronto volverán). Gracias también a los que estuvieron pero se fueron, porque seguro que dejaron algo de lo que aprender.*

*Por último, agradecer al lector por dedicarle tiempo a mi tesis, espero que sea tiempo bien invertido y despierte algún interés o alguna nueva idea en la cabeza.*

*La presente Tesis Doctoral ha sido financiada por el Ministerio de Ciencia, Innovación y Universidades bajo el proyecto RTI2018-098085-B-C42, el Ministerio de Economía y Competitividad bajo el proyecto TEC2015-67387-C4-4-R de los fondos FEDER, la Comunidad de Madrid bajo la Ayuda de Excelencia del Profesorado con referencia EPU-INV/2020/003, y la Universidad de Alcalá bajo los proyectos CCG2015/EXP-056, CCG2016/EXP-033 y CCGP2017-EXP/060, y por medio del programa de contratos predoctorales FPI.*

# Contents

<b>Contents</b>	<b>III</b>
<b>List of Figures</b>	<b>V</b>
<b>List of Tables</b>	<b>VI</b>
<b>I Introduction, state-of-the-art and methodology</b>	<b>1</b>
<b>1 Introduction and scope</b>	<b>2</b>
1.1 Introduction . . . . .	2
1.2 State-of-the-art . . . . .	5
1.2.1 State-of-the-art in VSD . . . . .	6
1.2.2 State-of-the-art in DPD . . . . .	8
1.2.3 State-of-the-art in VADHA . . . . .	8
1.2.4 State-of-the-art in PDA . . . . .	9
1.3 Problem formulation and scope of the thesis . . . . .	11
1.4 Materials and methods . . . . .	12
1.4.1 Overview of the experiments in VSD . . . . .	22
1.4.2 Overview of the experiments in DPD . . . . .	22
1.4.3 Overview of the experiments in VADHA . . . . .	23
1.4.4 Overview of the experiments in PDA . . . . .	25
1.5 Structure of the thesis . . . . .	26
<b>Bibliography</b>	<b>28</b>
<b>II Publications</b>	<b>37</b>
<b>2 Article 1: Energy-Efficient Acoustic Violence Detector for Smart Cities</b>	<b>38</b>
<b>3 Article 2: Cost-constrained Drone Presence Detection through Smart Sound Processing</b>	<b>47</b>
<b>4 Article 3: Analysis of the performance of Evolved Frequency Log-Energy Coefficients in Hearing Aids for different Cost Constraints and Scenarios</b>	<b>55</b>
<b>5 Article 4: Linear detector and neural networks in cascade for voice activity detection in hearing aids</b>	<b>64</b>

<b>6 Article 5: Smart Sound Processing for Defect Sizing in Pipelines Using EMAT Actuator Based Multi-Frequency Lamb Waves</b>	<b>77</b>
<b>III Conclusions and future lines</b>	<b>97</b>
<b>7 Conclusions</b>	<b>98</b>
7.1 Summary of the conclusions . . . . .	98
7.1.1 Conclusions for VSD . . . . .	98
7.1.2 Conclusions for DPD . . . . .	99
7.1.3 Conclusions for VADHA . . . . .	99
7.1.4 Conclusions for PDA . . . . .	100
7.2 Future lines . . . . .	101
7.3 List of publications . . . . .	102

# List of Figures

1.1	Illustration of the four applications in a hypothetical smart city. Source: Own elaboration from <i>freepik</i> vectors. . . . .	4
1.2	Scheme of supervised machine learning method applied along this thesis. . . . .	12
1.3	Feature extraction process. Source: Own elaboration from <i>freepik</i> vectors. . . . .	14

# List of Tables

1.1	Main parameters of the system implemented for VSD application. . . . .	23
1.2	Main parameters of the system implemented for DPD application. . . . .	24
1.3	Main parameters of the system implemented for VADHA application. . . . .	25
1.4	Main parameters of the system implemented for PDA application. . . . .	26

## Part I

# Introduction, state-of-the-art and methodology

# Chapter 1

## Introduction and scope

### 1.1 Introduction

The idea of a smart city appeared for the first time in (Heng and Low, 1993), when Singapore presented itself as an “intelligent city”. From that moment, different organizations and authors have raised several definitions and understandings of this concept. In (Giffinger et al., 2007), the authors presented a smart city as a place that performs its activity by intelligently combining the industry, education, citizen participation and technical infrastructure fields to serve its citizens. Later, IBM company defined a smart city as an instrumented, interconnected and intelligent city (Harrison et al., 2010): instrumented city because it must collect and integrate real data in real time from sensors, applications, personal devices, etc; interconnected city because all these data must be integrated into a computing platform to provide a set of services; and intelligent city because complex elements are required to meet this objective, including analytical calculations, modelling, optimization and visualization of services. Another famous definition was made by the authors in (Hancke et al., 2013), who considered a smart city as a city that operates in a sustainable and intelligent way, so that all the infrastructure and citizen services are integrated, and smart devices are used for monitoring and control.

To this day, and due to the fact that the smart city concept is relatively recent, it is not completely clear which technologies and ideas must be considered in their development, since this concept covers a large number of fields and technologies. One thing is clear: smart cities must implement the latest technologies to obtain benefits in a wide range of areas, focusing on creating social improvements, economic growth and new opportunities (Chamoso et al., 2018). And what is probably more important is that the final beneficiary in most of the perceptions of a smart city is the citizen: the main focus is “people first and foremost” (Boulos et al., 2015). This will be beneficial for the vast majority of people, since forecasts indicate that by 2050 around two-thirds of the world’s population will live in urban areas (6 billion people of the world’s total 9.7 billion people) (State of Green, 2020, United Nations, 2018).

The technologies that are part of a smart city can and must be applied to a wide range of aspects of the daily life of an urban area. For clarity, it is necessary to classify the different services that are usually available in a city. Authors in (Neirotti et al., 2014) provided a very clear and cohesive classification of them depending on the domain they belong to:

- Natural resources and energy domain: smart grids, lightning, renewable energies, waste management, water management, food and agriculture.
- Transport and mobility: city logistics, mobility information, mobility of people and services exposing district information models.



- Smart building: facilities management, construction services and housing quality.
- Daily life: entertainment, hospitality, pollution control, public security, health, welfare, social inclusion, culture and management of public spaces.
- Government: e-governance, e-democracy and transparency.
- Economy and society: innovation and entrepreneurship, cultural heritage management, digital education and human capital management.

In this thesis, four different applications that can contribute to improving the quality of life of people and must be present in this kind of spaces have been researched. In the following lines, they will be briefly described and classified into one or more domains from the previous classification.

- Violent Situation Detection (VSD). The public security issue, which is included within the daily life domain, tries to protect citizens and their belongings based on the active involvement of public organizations, the police, and even citizens themselves, including the collection and monitoring of information for crime prevention (Khan et al., 2014). A system capable of detecting when a violent situation is taking place in a public space (streets) or private space (home) would be extremely useful in a smart city. The idea is to detect the first signs of violence, such as a heated argument or the start of a physical fight between two or more people, avoiding the fatal end that these situations can imply, from serious injuries to the death of a citizen. In addition, this application can help to the improvement of the transport and mobility domain, particularly the mobility of people, since crime situations could be mitigated in public transport, avoiding disorders in the bus, the train or the underground, and ensuring the security and tranquility of all the passengers.
- Drone Presence Detection (DPD). City logistics issue, which is included within the transport and mobility domain in smart cities, tries to improve the logistics by efficiently integrating business needs with traffic, geographical and environmental conditions (Nowicka, 2014). Related to this topic, it seems evident that unmanned aerial vehicles, also known as drones, will be present in these futuristic cities. They will play some important roles, such as delivering goods and merchandise, serving as mobile hot spots for broadband wireless access, and maintaining surveillance and security (Vattapparamban et al., 2016). However, they could also be used by malicious entities or people to produce physical or cyber attacks, or to threaten the society with an invasion of privacy of the citizens and public administrations. Because of that, a system capable of detecting drones and checking if their flights are permitted or not will be necessary to ensure the public security issue previously mentioned, allowing the protection of citizens and their belongings.
- Voice Activity Detection in Hearing Aids (VADHA). The daily life domain includes the welfare and social inclusion group, which tries to improve the quality of life by stimulating social learning and participation, with particular attention to certain groups of citizens, such as the elderly and people with disabilities (Hussain et al., 2015). In this sense, people who suffer from hearing losses with varying degrees of severity must have access to intelligent devices capable of overcoming this deficiency. Hearing aids are devices with limited battery life, so it is essential that they include algorithms capable of detecting whether a conversation to which the user belongs is taking place. In this way, the device can be activated when this event occurs, and it can hibernate and save power otherwise. Although

**Figure 1.1:** Illustration of the four applications in a hypothetical smart city. Source: Own elaboration from *freepik* vectors.



this application has been included inside welfare and social inclusion groups, it provides benefits that go beyond them, since it allows people with this disability to access mobility information, entertainment, culture, etc.

- Pipeline Defect Assessment (PDA). The natural resources and energy domain includes the renewable energy group, which will try to exploit natural resources that are regenerative or inexhaustible, such as heat, water or air (González-Briones et al., 2018, Viitanen and Kingston, 2014). In 2018 renewables made up 26% of global electricity generation, and it is expected to reach 45% by 2040 (Murdock et al., 2019). Thus, it appears to be distant when this kind of energy production will be fully implemented around the world. Until then, gas and oil must be carried through pipelines buried underground and underwater along the countries and cities, which requires a certain maintenance. Furthermore, similar pipelines also distribute water, one of the most critical resources together with electricity. Nowadays, these distribution systems are non-intelligent, and sometimes it is tricky to diagnose the system early enough to detect a leak in one of the distribution pipes, especially if it is not readily visible (Hancke et al., 2013). In this kind of cities, it would be useful to implement advanced sensing capable of monitoring the condition of those pipelines, in order to detect whether corrosion and defects are appearing on them. In this way, the maintenance teams could repair the materials before a gas, oil or water leak happens, reducing costs and avoiding power and water cuts, which can affect to the citizens and companies.

In Figure 1.1, an illustration of a smart city is presented to aid the reader in identifying the

four applications. It can be observed a fight that has started between two people, a drone which is surrounding public administration buildings, two elderly people who are using hearing aids to communicate with each other, and an underwater pipeline that needs to be repaired. The VSD system has alerted the policeman to the fight, the DPD has warned the authorities about the presence of an unauthorized drone, the VADHA provides the two elderly people with a better listening experience just when a conversation is taking place, and the PDA system has allowed the maintenance company to know where the defect is located along the pipeline.

To carry out the tasks mentioned above, it is necessary to have data that allow us to detect whether these events are taken place or not. The type of data that can be used varies across the applications: audio, video, radar, radio frequency or temperature signals, among others. This thesis has researched how acoustics signals can contribute to those applications in two different ways: the use of sounds (in VSD, DPD and VADHA) and ultrasounds (in PDA). The following paragraphs review the state-of-the-art related to event detection with audio signals, both in a general way and specifically in each application.

## 1.2 State-of-the-art

In recent years, machine learning and signal processing advances have contributed to the development of new techniques for Sound Event Detection (SED). This field has an excellent potential for many applications, as well as many research challenges. The reason why sounds provide such a rich source of study is that this kind of signals carry a great amount of information about the environments that surround us, including both individual physical events and sound scenes as a whole (Virtanen et al., 2018). This becomes clear when we imagine ourselves with our eyes closed standing in the street of an urban area. We would probably listen to other pedestrians walking on the sidewalk, the sound emitted by the vehicles in a traffic jam, or the kids playing in a nearby park. But what if a discussion or a fight starts between two people? And if a drone starts flying around us? Would we be capable of recognizing when a person starts talking to us? The answer to all these questions must be an unequivocal ‘yes’. We should be able to identify these situations without needing to open our eyes and use the visual information that they provide to us. Consequently, an artificial intelligence system that processes the same information should be able to extract the same conclusions as ourselves, with the difference that they should use the signals recorded by microphones, or other devices, instead of the ones received by the ears.

Ultimately sound is an important source of information about the events that happen in urban life. The growing development of acoustic sensor networks means that urban sound monitoring is becoming an increasingly appealing alternative, or a complement, to video cameras and other forms of environmental sensing. This is due to the significant benefits that microphones provide compared with video cameras. Firstly, microphones are generally less expensive and smaller than video cameras, so they can be more easily placed anywhere. Secondly, environmental conditions (daily changes in light, fog, pollution or rain) affect negatively to the visibility, making the quality of the video camera signal worse, while the quality of the microphone signal remains intact in this sense. Thirdly, sound is less susceptible to occlusion, since it can travel through obstacles even if diffraction effects cause scattering in the signal. Unfortunately, the same does not happen when something or someone appears on the scene between a video camera and the situation of interest. Fourthly, microphones are capable of providing omni-directional sensing, while video cameras can hardly record a panoramic view bigger than  $180^\circ$ . Finally, one of the most important advantages of microphones is their small consumption compared with video cameras. It means that if these devices are programmed to work autonomously because they can not be directly connected to a power source, or they are powered by a solar cell, or the city just needs to be less

polluting, the battery of the microphones will last much longer.

Regarding this last issue, it must be considered that the elements of a smart city must be effective and innovative enough to avoid being harmful to the environment (Chamoso et al., 2018). The way we can make these devices less polluting is directly related to their consumption, apart from feeding them with renewable sources of energy instead of carbon-based ones, which lies outside the scope of this thesis. For this reason, the algorithms presented in this study are constrained in terms of computational cost, and the computational cost parameter goes hand in hand with the final consumption of the device. This will make it easier for the resulting devices to extend their battery life, that is, to increase their autonomy. We will see further on how the computational cost has been controlled in the different applications, depending on the available resources of each of them.

Furthermore, approximately half of the world’s population has a smartphone with a microphone (3.5 billion people according to December 2020 statistics (O’Dea, 2020), out of 7.8 worldwide citizens). This fact makes easier the deployment of these applications in a hypothetical future where microphone signals could be used for smart cities purposes. In this assumption, microphones from the users could continue providing information even when the citizens carried them in their pockets, which would not be the case with video cameras. However, it is too early to state whether it will be possible, since issues related to the right to privacy and the right to anonymity should be discussed previously.

Once the type of signals used in this thesis has been defined, it is important to define the term ‘sound event’ before continuing. A sound event is a specific sound produced by a distinct physical sound source that typically has a well-defined, brief, duration in time (e.g., a car passing by, a bird singing, or a doorbell). This concept must be differentiated from the ‘sound scene’ one, which refers to the entirety of sound formed by sounds from various sources (e.g., the sound scene of a street, which contains cars passing by, footsteps, people talking, etc.). A sound event can be classified or detected, which consists in locating in time the occurrences of a specific type of sound by finding all the temporal positions when the sound is active. In this thesis, SED will be the subject matter, resulting in two different classes in each dataset: occurrence of the event (‘1’ or ‘positive’) or absence of event (‘0’ or ‘negative’) (Virtanen et al., 2018).

As stated at the beginning, SED requires using several techniques related to machine learning. The term ‘machine learning’ was first coined in the ’50s, and it referred to the transfer of intelligent activities made by human to machine (Guyon et al., 2008). Since that moment, the research efforts have focused on finding and extracting relationships in data. This methodology can be applied in every task defined by a series of examples or cases rather than by predefined rules. Another definition for machine learning is the subset of artificial intelligence that builds a mathematical model based on sample data (“training set”) to make predictions or decisions (our case) in a new set of data (“test set”) (Zhang, 2020).

The implementation of SED techniques differs quite considerably between the four applications due to the peculiarities of each of them. Because of that, the review of the state-of-the-art will be presented separately, starting with the literature of VSD in Section 1.2.1, continuing with the state-of-the-art of DPD in Section 1.2.2, addressing later the literature of VADHA in Section 1.2.3, and finishing with the state-of-the-art of PDA in Section 1.2.4.

### 1.2.1 State-of-the-art in VSD

Many people suffer from violence issues every day in society, and statistics show this number has maintained or almost increased recently. The first that must be determined when addressing this problem is the definition of the term ‘violence’, since each person could understand it in a different

way. The World Health Organization defines violence as “the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, which either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation” (Krug et al., 2002). Other related works define violence in very diverse ways, principally referring to the physical act itself. Some examples are: “physical violence or accident resulting in human injury or pain” (Penet et al., 2012), “a series of human actions accompanying with bleeding” (Chen et al., 2011), or “any situation or action that may cause physical or mental harm to one or more persons” (Giannakopoulos et al., 2006). However, anything produced in an injurious or damaging way might be described as violent, even if there is no implicit physical effect. Depending on the nature of violence, it can be classified into physical, psychological or even sexual violence. Nowadays, women and children are the largest and most vulnerable group of victims. Referring to the first one, recent researches show that 35% of women around the world have suffered physical or sexual violence during their lives (World Health Organization et al., 2019), and 43% of women from the European Union declared suffering psychological violence at least once (FRA-European Union Agency for Fundamental Rights, 2014). On the other hand, lots of children are abused every day in the schools. In 2016, more than 9% of students from Spain suffered from bullying, and approximately 21% were usually insulted (Trueba Sánchez et al., 2016). These statistics demonstrate how these situations continue being relevant at present, and this problem must be treated to prevent the tragic end that sometimes takes place.

Studies in the literature have tried to find a solution to this problem previously, some of them using both audio and video signals. They have applied MFCC-based audio and advanced motion features (Chen et al., 2011), or SVMs for action scene detection and bloody frame detection (Acar et al., 2016), and the results obtained with the combination of those sources seem to be efficient. However, as stated at the beginning of this thesis, the main disadvantages of video are its high computational cost and intrusiveness, and its poor coverages. Furthermore, audio and video have been tested in the literature separately and in combination (Dias, 2016). Their conclusions show that the system works properly using just audio source. When video information is added, the performance improves slightly, but computational cost increases in a big way. In the state-of-the-art, other proposals where audio is used to detect violence by itself can be found (Giannakopoulos et al., 2006), since violent situations are commonly accompanied by signs like arguments, shouts or an increase in the volume of the conversation. However, they tested a small dataset (only 20 minutes of audio).

Related to the data used, some problems were found in the datasets available in the literature. In (Jain and Vishwakarma, 2020), a review of the most popular datasets for violence detection is made, including 15 different video datasets from 2004 to 2017, and in (Ramzan et al., 2019) a review of the state-of-the-art violence detection techniques is carried out. The main conclusion from those studies is that artificial vision techniques have been highly extended, so the datasets available usually implement video-features more than audio-features, and in most of them the images are recorded with surveillance cameras, which sometimes do not include audio information. Other datasets are composed of videos from films or games, which have been recorded in optimal conditions with pretended violence (Nievas et al., 2011, Perperis et al., 2011, Schedi et al., 2015, Sjöberg et al., 2014). Another problem is that in some datasets, the definition of violence is directly related to the appearance of gunshots or blood in a sequence, which differs significantly from the objective pursued in this thesis.

### 1.2.2 State-of-the-art in DPD

Unmanned aerial vehicles, also known as drones, are extensively used in nowadays societies due to the advantages they provide, and they are expected to play a major role in connected smart cities in the future, including tasks like package delivery, traffic monitoring, policing, drone taxi, ambulance services, firefighting, rescue operation, etc. (Khan et al., 2018). However, people sometimes misuse them, trying to invade the privacy of others or bypass the security systems (Altawy and Youssef, 2016). A recent example is the drone that crashed on the White House in January 2015, causing a brief lockdown (Miller, 2015). Officials said that the device was a two-foot wide remote-controlled quadcopter that is sold in stores. More recently, a drone bypassed all security efforts in Puerta del Sol (Madrid) last New Year's Eve, when the square was almost empty due to Covid-19 restrictions (Keane, 2021). This device carried a black ribbon along with a Spanish flag and appeared during the TV broadcasting.

For all these reasons, it is essential to develop a system capable of detecting the presence of drones in particular environments where they can be used for malicious purposes, such as households, public buildings, or restricted-access areas. In the state-of-the-art there are many studies that deal with this issue (Ganti and Kim, 2016). In general, this problem can be approached using different data sources, like radar information, radio frequency, video, or even audio signals. All of them have some drawbacks. Some manuscripts use radar signals for aircraft detection (Drozdowicz et al., 2016), but the small size of the drones complicates the task. Related to radio frequency based methods, they are useful for the problem at hand, since radio frequency is the communication mode used between drones and the remote controller (Nguyen et al., 2016). However, the use of Wi-Fi range (2.4-5 GHz) in no-license channels causes the appearance of high interferences. Talking about temperature-based detection (Farlik et al., 2019), it is an efficient solution if the drone uses a propulsion engine, which usually appears in fixed-wing drones. However, most current drones are made of plastic and their electric engines do not radiate much heat. Video disadvantages were explained previously in this thesis, but it appears an additional difficulty when distinguishing between drones and birds, even after including bird flight patterns that drones do not follow (Ganti and Kim, 2016).

Some proposals have based their study on audio information, mixed or not with video. Authors in (Case et al., 2008) propose using an array of microphones and an infrared camera to get the information. They try to trace the path followed by the drone through beamforming techniques. Others use only one microphone (King and Faruque, 2016), but they focus on detecting a particular model of drone, so the results could not be generalizable. In other manuscript (Ganti and Kim, 2016), the authors analyze video information to detect the difference between frames, and in this way they track the drone movement, while they use audio information for detecting the vehicle with a threshold in frequency. The problem is that it is not very effective when background noise is high. In addition, audio appears to be more reliable for detecting drones according to some studies (Liu et al., 2017).

For several reasons such as privacy, it is difficult to find public drone audio datasets in the literature. Recently a dataset was uploaded (Al-Emadi et al., 2019), but it only included 662 s of drone sound from the Parrot Bebop and Parrot Mambo models. The authors doubled the duration of the dataset by adding background noise to them, and other unknown sounds were included too.

### 1.2.3 State-of-the-art in VADHA

Hearing loss is a common issue, especially when people become older and start to suffer from hearing impairment (Meister et al., 2015). It is a problem that affects over 5% of the world

population, having the largest impact on people over 65 years old (one-third of them has a loss greater than 40 dB) (World Health Organization, 2020). This issue has several health implications, including social isolation, depression, altered physical function, reduced activity participation, falls, greater cognitive decline, and higher risk of dementia (Amieva et al., 2018). To provide these people a better quality of life, hearing aid devices are the best option, as their use has a positive impact on long-term cognition (Amieva et al., 2015). However, these devices must present several restrictions due to the consumption and real-time processing requirements. Firstly, the computational capability of the device and the number of assembled components must be restricted to satisfy the battery life requirements. Secondly, the processing algorithms must present a low-delay, which in numerical terms cannot exceed 20 ms (Stone and Moore, 2002). All the mentioned characteristics must be taken into account during the design process of these devices (Gil-Pita et al., 2017). For example, the low-delay requirement limits the length of the time frame in the time-frequency analysis, therefore limiting the frequency resolution.

Several algorithms are presented in hearing aids, including feedback cancellation, environment classification or speech enhancement (Gong and Xia, 2015, Lee et al., 2017). However, all of them depend on the VAD algorithm, which allows differentiating between conversations and noise. The field of VAD is plenty of research in the literature: some authors used a decision-directed parameter estimation along with Hidden Markov Model (HMM) (Sohn et al., 1999); others measured the Long-Term Spectral Estimation (LTSE) between speech and noise, and compared the envelope to the average noise spectrum (Ramirez et al., 2004); in (Wisdom et al., 2015), the second-order non-circularity of speech and noise complex subbands is used; authors in (Mukherjee et al., 2018) used features based on Line Spectral Frequency (LSF) along with extreme learning classifiers; a smartphone app was developed in (Sehgal and Kehtarnavaz, 2018) for real-time VAD with Convolutional Neural Networks (CNNs); a Deep Neural Networks (DNN) based model was implemented in (Kim and Hahn, 2018); just to mention a few. Recently, a study compared a large number of previous proposals and showed that, despite the usefulness of a long temporal context and a look-ahead for VAD, they require much more CPU consumption than the available for off-the-shelf hearing aids (Graf et al., 2015)

VADHA issue has already been studied in the literature too. In (Gil-Pita et al., 2015), the authors proposed a computationally efficient system for sound environment classification and VAD, which provided a proper classification of some audios into speech, music and noise. They considered the computational limitations previously stated, and because of that, they proposed a novel set of designed features inspired in the MFCCs, denominated Evolved Frequency Log-Energy Coefficients (EFLECs). These coefficients achieve a performance equivalent to the MFCC one, but reducing the computational complexity. This is achieved by using uniform filters instead of triangular ones, removing the Discrete Cosine Transform (DCT) block, and applying evolutive algorithms to select the limits of the frequency bands where the filters are distributed, instead of using the standard Mel scale.

#### 1.2.4 State-of-the-art in PDA

It is estimated that the length of the oil and gas pipelines worldwide is higher than 3.5 million kilometers (CIA, 2019). During their useful life, some failures can take place, including corrosion, weld defects and third-party damage. The first of them is the most common, being the first cause of failures according to studies from Europe (European Gas Pipeline Incident Data Group, 2015), United States (Wang et al., 2017), Canada (Alberta Government, 2017) and United Kingdom (United Kingdom Onshore Pipeline Operators' Association, 2018). When corrosion takes place over a long period of time, it usually results in the appearance of leakages, whose consequences

are devastating: heavy economic losses appear (Lu et al., 2020d), and also the environment and personal safety can be threatened (Lu et al., 2020b,c). As an illustration, an oil pipeline leaked last November 2013 in the Chinese city of Qingdao, giving rise to some explosions that left at least 62 people died and 136 injured, and causing economic losses of more than 751 million Chinese yuan, apart from spreading oil to the sea (Meng, 2013). Subsequently, an investigation found 8000 safety problems and corrosion issues along the nation's pipeline network.

To reduce the probability of these events happening, it is important to monitor the state of the pipelines before the leakage occurs through the use of corrosion detection methods. In (Lu et al., 2020a), the authors provide a deep classification of the different technologies, differentiating between hardware-based and software-based methods. All the methods have different characteristics, as well as advantages and disadvantages. One of them is the ultrasonic guided wave method, an acoustic technique widely used in nondestructive testing (NDT) based on the generation of waves through a sensor. This sensor generates a wave that is propagated through the boundary of the pipeline, reflecting back and forth at the interface and resulting in complex waveform conversion and mutual interference. The sensors can be selected to stimulate single or various modes of the guided wave. The disadvantages of this technique are the following: the selected detection frequency must be obtained previously; professional staff are required to interpret the data; the echo signal of the wave can be affected by the outer layer of the pipeline, the inhomogeneity of weld and the severity of the defect; and there can be high requirements for the sensors. On the other hand, this technique offers important benefits: it is especially suitable for covering long distances, which can save detection time and costs, and reduce the labor intensity; it allows the company to detect defects of the whole section of the pipeline (HAO and SHI, 2008); in addition, it provides low attenuation and high testing efficiency in the experiments (Liu et al., 2020).

The literature approaches generate ultrasonic guided waves through three main methods: piezoelectric transducer (PZT), magnetostrictive patch transducer (MPT) and electromagnetic acoustic transducer (EMAT) (Green Jr, 2004, Huan et al., 2019). The biggest issue when using the first two techniques is that there must exist a robust sonic contact and coupling with the pipeline, so the inspection is not efficient in some applications. For its part, EMAT can test the pipelines in a non-contact and coupling-free way, and there is no need to previously accommodate surfaces with oxide, dirt or coatings (Hao et al., 2011, Xie et al., 2017). Other advantages provided by EMAT are its high speed of inspection and the capability to test structures submitted to high temperatures and fast movement (Pei et al., 2016, Petcher et al., 2014, Urayama et al., 2010). In addition, this technology is capable of exciting multiple types of waves: Lamb, shear, longitudinal and Rayleigh. These EMAT-generated guided-waves are usually generated through some magnetics and testing coils, in a way that the magnetics create a static magnetic field, and the testing coils induce and receive an eddy current or an alternating magnetic field (He et al., 2017). Furthermore, when the system is implemented with meander-line-coils, the waves are generated in a directional way, which allows differentiating between circumferential and axial scans.

The pipeline inspection using EMAT has been previously studied in the literature. In (Clough et al., 2017), the authors provide a screening technique that generates Shear Horizontal (SH) waves to axially examine pipelines through circumferential scans. They explain the behavior of different wave modes when they interact with defects, in both experimental measurements and artificially created ones, but they do not provide error measurements, so it is difficult to know how well the defect sizing is carried out. In (Nakamura et al., 2017), the authors use other type of guided waves (torsional) and test several aluminum pipes, concluding that amplitude and phase information have enough detection sensitivity, being this last one more powerful for



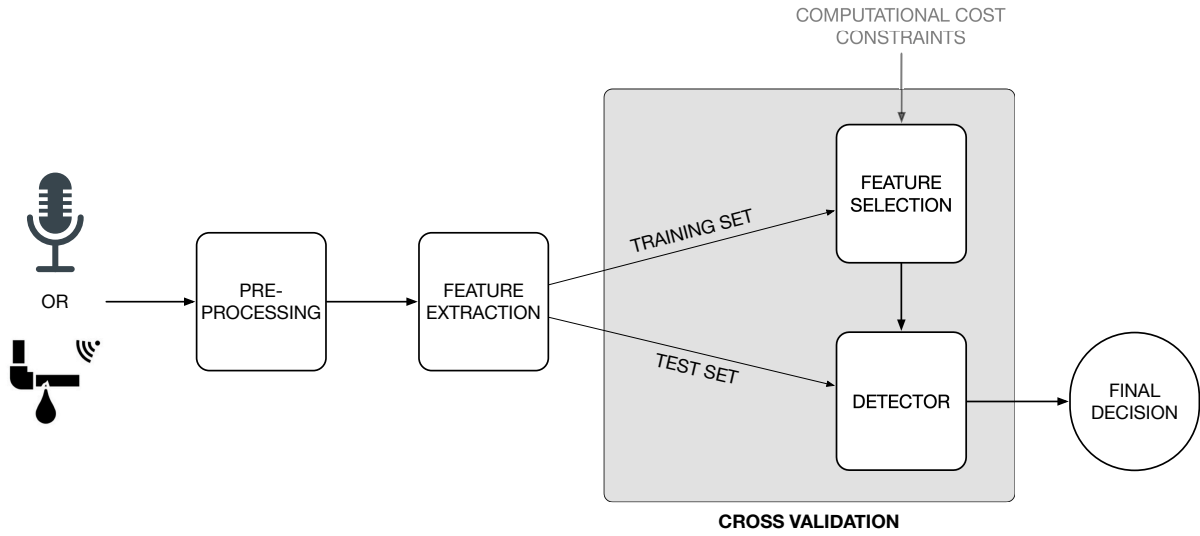
quantitative purposes. Even they classify the usefulness of amplitude and phase information into good, unsatisfactory or poor, for different types of defects, performance measurements related to defect sizing are not presented to the author again.

Machine learning methods have been previously applied to this issue in the literature. Authors in (Sun et al., 2020) use deep neural networks for defect inspection. However, they evaluate their algorithms in a plate instead of pipelines, and though the results using deep neural networks are very promising, these techniques require large datasets with a wide variety of defects and are very costly in computational terms. It may represent a problem in some cases when applying portable handheld inspection, since a large proportion of the battery life may be consumed during the signal generation and taking of measurements, so the processing block must be energy-efficient to provide enough autonomy. Something similar is proposed in (Lu et al., 2018), where deep neural networks are applied to other type of signals for pipeline sizing. In other proposal (Layouni et al., 2017), the authors apply feature extraction (maximum magnitude, peak-to-peak distance, mean average, standard deviation, integral of the normalized signal), pattern-adapted wavelets and artificial neural networks for defect detection and sizing. Similarly, authors in (Mohamed et al., 2015a,b) apply feature extraction combined with Artificial Neural Networks (ANNs), or with Support Vector Machines (SVMs), but they provide insufficient information about the dataset used, without detailing the number and characteristics of the defects used in their experiments. In the last three cited studies the results are promising, but the authors use Magnetic Flux Leakage (MFL) inspection for acquiring the signals, a method which has proved to have several limitations (Safizadeh and Azizzadeh, 2012, Shi et al., 2015): it is hardly applied in practice, as a big qualitative analysis of the signal is needed because the working conditions can not match the laboratory conditions; it is limited to the material surface and near surface, but the detection of axial narrow and long defects is restricted; the probe is susceptible to the pipe wall and its anti-interference ability is low, in a way that false data will be collected when impurities appear; among others.

### 1.3 Problem formulation and scope of the thesis

The research questions that have been identified from the review of the literature and will be answered through-out this thesis are the following:

- RQ1: Are the datasets available in the state-of-the-art suitable for testing the four applications? If not, is it possible to create new acoustic datasets?
- RQ2: Is it feasible to solve these problems using standard machine learning techniques without applying more deeply learning ones, which generally involve a higher computational cost?
- RQ3: Is it possible to carry out a quantitative analysis of the computational cost required by those systems to increase their autonomy, trying to reach a compromise between the performance of the algorithms and the computational cost associated with them, without highly degrading the proper functioning of the system?
- RQ4: Can advanced optimization methods based on cascade-detectors reduce even more the computational cost of the system in VADHA issue while keeping the same performance? Similarly, is it possible to use these cascade-detectors to improve the performance of the system without increasing the resulting computational cost?

**Figure 1.2:** Scheme of supervised machine learning method applied along this thesis.

- RQ5: Are the typical pattern recognition methods suitable for an application in which ultrasounds are used as signals (PDA)? In this issue, is it possible to heuristically find useful features to solve the problem?

## 1.4 Materials and methods

From the raised scope of the thesis, several research studies have been carried out. In this section, a summary of the different investigations carried out is presented. They will be explained in detail in the articles attached in Part II: Publications.

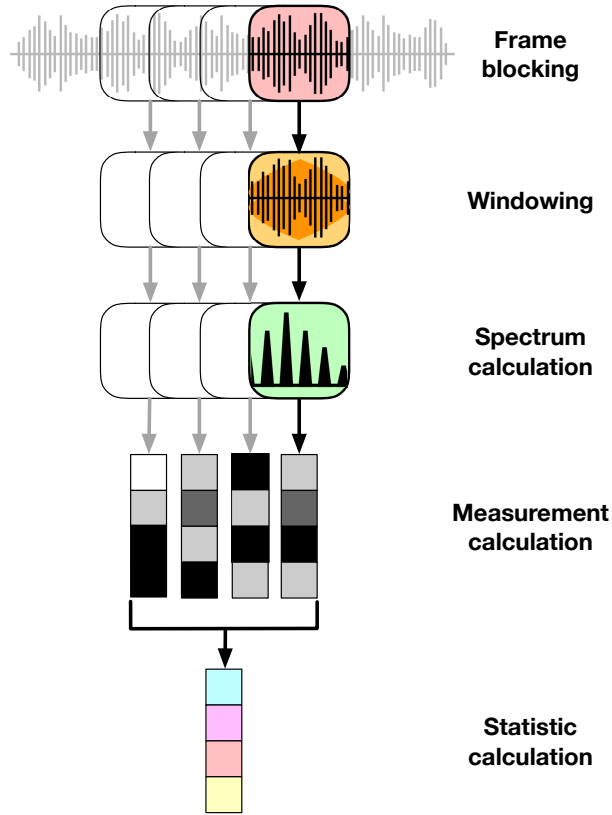
Machine learning is an extensive field that can be approached in different ways. One of them is supervised learning, which will be used along this thesis as it is the mainstream and typically the most efficient and generic method for developing this kind of systems (Virtanen et al., 2018). It consists in having a training set that includes labeled data and learning a general rule that maps inputs to outputs. By way of analogy, this is like a teacher or supervisor that gives a student a problem (finding the relationship between inputs and outputs) and its solutions (labeled output data), and later he asks that student to learn how to solve new problems (unseen data). In Figure 1.2 the scheme of supervised learning that has been followed along the research is shown.

The system takes an input captured by a microphone (in the case of sounds) or a sensor (in the case of ultrasounds). In this thesis, we have worked with datasets composed of signals captured with one of the previous devices. Once the signals are available in our system, they go through different modules or stages: pre-processing, feature extraction, feature selection, and detection, including cross-validation techniques, until the final decision is taken. Each of them is detailed below.

- Pre-Processing. In some audio signals it is necessary to implement this module before feature extraction is applied, in order to enhance certain characteristics of the signal for maximizing the performance of the detection. For example, when the audio data is collected from various sources, there are usually variations in the sampling frequency and the amount of captured audio channels (Virtanen et al., 2018). In this thesis, this occurs in VSD, DPD

and VADHA. These variations have been addressed by converting the audio signal into a uniform format, both re-sampling it into a fixed sampling frequency and down-mixing it into a fixed number of channels. On the one hand, sampling frequency has been set to 22 kHz in VSD, 16 kHz in VADHA, and 8 kHz in DPD. The reason for using different sampling frequencies is due to the different quality of the audios that makes up the different datasets, since the sampling frequency must always be at the lowest of the frequencies of all the audios. In addition, having different values of this parameter allows us to check if SED can be performed in different recording conditions. On the other hand, the number of channels was fixed to one, as most of the audios do not provide multi-channel recordings. For their part, this stage of pre-processing allows us to discard signals in PDA when the sensor does not run properly along the pipeline. Other parameters are standardised in this application along all the measurements.

- Feature extraction. Once the signals are standardised to the desired values of sampling frequency and number of channels, it is time to extract useful information from them. Feature extraction is a widely extended stage (Virtanen et al., 2018). The objective of extracting acoustic features is to represent the audio in a compact and non-redundant way. The idea is that these features vary slightly among the examples which are part of the same event, and present distant values between examples that are part of different events (in our case, the presence or absence of an event) (Gold et al., 2011). In addition, the amount of memory and computational power required by these features will always be lower than using the raw data. As shown in Figure 1.3, several procedures can be identified within this stage: frame blocking, windowing, spectrum calculation, measurement calculation, and statistic calculation.
  - Frame blocking. In this part of the stage, the audio signal is sliced into fixed-length frames, shifted with a timestep. This is due to the fact that audio signals are generally non-stationary and their parameters change rapidly over time, so it is better to analyze periodically short-time segments or frames, where the signal is quasi-stationary. The length of the frames is different in each application: 512 frames in VSD, which sampled at 22 kHz translates into 23.22 ms; 512 frames in DPD, which sampled at 8 kHz is equivalent to 64 ms; and 128 samples in VADHA, which sampled at 16 kHz translates into 8 ms. Typical frames are between 20 and 60 ms, but in the VADHA application the total delay introduced by the processing time cannot exceed 20 ms (Stone and Moore, 2002).
  - Windowing. When spectral features are extracted (VSD, DPD and VADHA applications), it is important to smooth the frames with a windowing function. It will avoid the appearance of distortions in the spectrum due to abrupt changes at the frame boundaries. In this thesis, when frequency-based features have been calculated, a Hann window (also known as Hanning or raised cosine window) has been applied (Harris, 1978). It has been used because it is the most popular one, resulting in an outstanding overall performance (it works properly in around 95% of cases according to (National Instruments, 2019)). It provides fair frequency resolution and reduced spectral leakage, thus increasing the dynamic range of analysis, as leakage can swamp signal components of close frequencies and much smaller magnitudes.
  - Spectrum calculation. Features can be calculated directly in the time-domain, but sometimes it is interesting to work in the frequency-domain. In this thesis we have worked in one or both domains, depending on the application: time-domain and

**Figure 1.3:** Feature extraction process. Source: Own elaboration from *freepik* vectors.

frequency-domain in VSD and DPD, frequency-domain in VADHA, and time-domain in PDA. The spectrum calculation involves computing the Discrete Fourier Transform (DFT), which represents the time-domain signal as a superposition of sinusoidal base functions, each of them with different value of magnitude and phase (Oppenheim and Schaffer, 1989).

- Measurement calculation. After the previous steps, acoustic measurements are computed through different equations and formulas. In this sense, the amount of measurements available in the literature for SED is substantial. However, one of the most frequent and that deserves to be explained at this point since it will appear in almost all the applications are the Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980). They are designed to emulate the human auditory perception, which focuses only on magnitudes of frequency components. The perception of these magnitudes by the human ear is highly non-linear, and, in addition, the perception of frequencies is also non-linear. On this basis, the extraction of these measurements uses non-linear representation for magnitudes (power spectrum and logarithm) and non-linear frequency scaling (Mel-frequency scale). This scale is implemented by using a set of filters that integrate the spectrum at non-linearly spaced frequency ranges, with narrow band-pass filters at low frequencies and larger bandwidth at higher frequencies (Virtanen et al., 2018). In summary, these measurements provide a compact and smooth representation of the spectral envelope, so that most of the energy is concentrated in the first coefficients (Mohino-Herranz, 2017). The computing process of the MFCCs is: once the DFT is computed, a bank of triangular filters spaced ac-

ording to the Mel frequency scale (Stevens et al., 1937) is applied, and after that the logarithm of the Discrete Cosine Transform (DCT) is computed.

- Statistic calculation. Once the different measurements have been calculated, some parameters or statistics are applied to a segment of samples consisting of several fixed-length frames previously mentioned. The number of frames that compose the segment depends on how frequently we want the decision to be taken: the decision is made every 5 s in the VSD, every 1 s in the DPD, and every 16 ms in VADHA. The parameters applied to these segments are the mean and the standard deviation, which are the most commonly used in the literature, although others have also been calculated depending on the different applications. Once these statistics are applied to the measurements, feature extraction has finished, and it is possible to move to the next step.

Before continuing, it must be noted that in the application of PDA the feature extraction is made directly in the signal that arrives from the sensor, since the features applied on them are different from the ones usually implemented for acoustic purposes. The analysis of each frame of the signal disappears as other measurements that will be detailed later are considered, and the number of samples of the signal in each point of the pipeline is limited to the length of the circumference of the pipeline. In a way, the traditional time-analysis and frequency-analysis used in the rest of applications are not valid when using ultrasounds, but the general methodology can be applied almost identically.

- Feature selection. There may be grounds to select a subset of features from all the features computed in the previous stage. In this thesis, a subset of relevant features has been selected due to the more extensive set initially calculated and to control the computational cost of the system. There are several reasons for using feature selection:
  - The simplification of the complexity of a model makes them easier to be interpreted by future researchers or users (James et al., 2013).
  - It enables training times to be shorter, that is, the machine learning algorithm can be trained faster.
  - It allows reducing generalization problems by minimizing overfitting (or reduction of variance) (Bermingham et al., 2015). This phenomenon appears when the model cannot identify the relevant information in the training data and, instead of that, it specializes in those data without extracting a general rule from them. Overfitting happens when the trained models violate the principle of parsimony (Hawkins, 2004) (also known as Occam’s Razor principle (Walsh, 1979)), which states that models and procedures must contain all that is necessary for the modeling but nothing more. Overfitting models include more terms or use more complicated approaches than are required. As a result, when new data from other datasets are tested, the algorithm performance fails, remaining unable to generalize.
  - The accuracy of the model improves if the right subset is chosen.
  - The curse of dimensionality problem is avoided. This phenomenon (Bellman et al., 1957) relates to the fact that when the dimensionality increases, the volume of the space increases so fast that the available data appear sparse. This sparsity becomes a problem for any statistical method, as generally the amount of data needed grows exponentially with the dimensionality.

In the literature, there exist different alternatives regarding feature selection procedures. The simplest way to carry out this task is to test each possible subset of features finding the one which gets the best performance, that is, the lowest error rate. This is an exhaustive search of the space that is computationally unthinkable, even more if our goal is to develop an energy-efficient system. The most common and feasible methods used in the literature are classified into wrapper, filter and embedded methods (Guyon and Elisseeff, 2003).

- Wrapper methods are a simple and powerful way of carrying out feature selection (Kohavi and John, 1997). They use a predictive model to score subsets of features, giving us an idea of its usefulness. Each subset of features is used to train a model, whose performance is tested on a hold-out set. The major disadvantage of this method is its computational intensity, due to the use of the classifier in each subset of features. On the other hand, they usually provide the best subset of features for a particular model, problem or dataset.
- Filter methods select subsets of features by scoring them using another proxy measure, different from the typical error rate (e.g., the mutual information, the pointwise natural information, the inner-class or intra-class distance, etc.). This makes the method faster, but still capturing the usefulness of the feature set. The main difference with the wrapper methods is that filters usually need less computational resources to be executed, but the chosen feature subset is not trained to a specific type of predictive model (Zhang et al., 2013). Consequently, the performance provided by this type of methods is usually lower than the one offered by wrapper ones.
- Embedded methods choose the best subset of features in the process of training. Their computational complexity is between the previous two methods.

In this thesis, the number of computed features is computationally affordable (around 150 features in the worst cases, VSD and DPD). Because of that, the methodology which provides the best subset of features in terms of performance has been applied, that is, wrapper methods. Within these methods, one of the most extended is the randomized one, which uses search strategies such as simulated annealing (Meiri and Zahavi, 2006), hill climbing (Long et al., 2011) or genetic algorithms (Shah and Kusiak, 2004). The latter have been successfully applied in the past to the issue of feature selection (Babatunde et al., 2014, Tan et al., 2008), and it has been used in this thesis. These algorithms, proposed in (Holland et al., 1992), are based on the principle of survival of the fittest, in a way that some modifications are applied to the chromosomes of the individuals of one population to find for the one which fulfills better with the requirements of our system. The individuals would be the different subsets of features, being the chromosomes each of the available features previously computed (if the chromosome is ‘1’, the feature is included in the subset; if the chromosome is ‘0’, the feature is discarded). The following process is applied iteratively:

1. First, several individuals (subsets of features) are randomly generated.
2. It is checked if one of them is identical to another, and in this case, one of them is changed (through including or discarding one of the features).
3. Constraints are applied to each individual of the population. In this thesis, computational cost has been restricted, so it is checked if the sum of costs of each of the features that compose the individual exceeds the set value. In this case, one or more features are disabled in the individual (set to ‘0’) until it fulfills the cost requirement.

4. The different individuals are ranked according to the performance evaluation parameter obtained after applying a simple detector over the design set.
5. The best individuals are selected as “survivors”, while the rest are regenerated through crossovers between the survivors.
6. All the individuals, excluding the best one, are mutated with a very low probability.

The stopping point of the algorithm is controlled through the number of iterations. In addition, the full process is repeated several times to avoid local minima and facilitate generalization in the results.

- **Detector.** Once the subset of features is determined, it is time to assign each observation to one of the classes (in a binary way in the case of VSD, DPD and VADHA, and as a predictor in the case of PDA). The overall aim in this step is to build a discriminant function capable of assigning each of the observations to one of the classes. This function should meet two requirements: it must ensure generalization, in a way that the algorithm must work properly if another dataset is tested, without “learning the data” during the learning process; and it must optimize a set parameter, which will be different in each application: error rate in the case of DPD and VADHA, probability of detection for a set probability of false alarm in the case of VSD, and Root-Mean Square Error (RMSE) in the case of PDA. To fulfill the previous requirements, the detector must follow the principle of Ockham’s Razor (Jefferys and Berger, 1992). This principle states that having two theories on equal terms and with the same consequences, the simplest theory has more probability of being the correct one. From a preliminary study of the data from this thesis, three detectors in total have been applied along the different applications, depending on the requirements of each of them. They are the Least-Squares Linear Discriminant (LSLD), the Least-Squares Quadratic Discriminant (LSQD), and the Multilayer Perceptron (MLP). Others like k-Nearest Neighbors (KNN) or Support Vector Machines (SVM) were tested as well, but because of the worse results obtained, they have not been included in the articles.

- **Least-Squares Linear Discriminant (LSLD).** This detector is based on the work made by (Van Trees, 1968). Its main advantages compared to other detectors are: it is not necessary to know the probability density function, which can be difficult to characterize; the values of the coefficients can be directly obtained from the design data without using optimization algorithms; and the classification rule is straightforward, so the overfitting problems are minimized, avoiding loss of generalization. As a disadvantage, this method provides decision boundaries that can be very simple for some issues. However, this applies above all to multi-class problems, while in this thesis 2-class (binary) problems are studied in most applications. In the following lines, the steps that allow us to implement the LSLD in a binary application will be detailed. In a linear classifier, the decision rule ( $\mathbf{g}$ ) is a function of a linear combination of the components of the observation. It is expressed in the following equation:

$$\mathbf{g} = f(y) = f\left(\sum_{n=1}^C w_n x_n + b\right), \quad (1.1)$$

being  $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_C]^T$  the input vector;  $\mathbf{w} = [w_1 \ w_2 \ \dots \ w_C]^T$  the weights vector;  $C$  the number of features; and  $b$  a constant value called “bias”.

In binary issues, the decision boundary is usually a hyperplane. This occurs whenever the number of features is equal to or higher than three. The decision is made by applying a threshold to the linear combination. It can be expressed according to the following expression.

$$y = \sum_{n=1}^C w_n x_n + b = \begin{cases} -1, & \text{if } y < 0 \\ +1, & \text{if } y \geq 0 \end{cases}, \quad (1.2)$$

being +1 y -1 each of the classes.

The most extended linear classifiers are based on the Linear Discriminant Analysis (LDA) proposed by Fischer (Xanthopoulos et al., 2013), whose objective is to combine the features to make the discrimination between classes as effective as possible. The optimum application of the LDA tries to minimize the distance between the same class patterns and maximize the distance between the different class patterns at the same time. There are numerous studies about LDA and its variants. In this thesis, the LSLD has been applied, a detector that tries to minimize the Mean Square Error (MSE). Now the mathematical formulation upon which LSLD is based will be detailed. Firstly, the desired output ( $\mathbf{t}$ ) and the coefficients of the linear combination ( $\mathbf{v}$ ) are defined:

$$\mathbf{t} = (t_1 \ t_2 \ \dots \ t_I) = (1 \ 1 \ \dots \ -1) \quad (1.3)$$

$$\mathbf{v} = (w_1 \ \dots \ w_C \ b), \quad (1.4)$$

being  $I$  the number of instants in which the decision has to be taken.

Now we define the design patterns through the matrix  $\mathbf{Q}$ , which contains the input features for the detection:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{P} \\ \text{ones}(1, I) \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1I} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2I} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{C1} & x_{C2} & x_{C3} & \dots & x_{CI} \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix} \quad (1.5)$$

So, the output of the linear detector for the design data is a vector obtained as a linear combination of the inputs according to equation (1.6)

$$\mathbf{y} = \mathbf{v} \cdot \mathbf{Q} = [y_1 \ y_2 \ \dots \ y_I] \quad (1.6)$$

The error is the difference between the desired output and the obtained output:

$$\mathbf{e} = \mathbf{y} - \mathbf{t} = \mathbf{v} \cdot \mathbf{Q} - \mathbf{t} \quad (1.7)$$

The error is a vector with size  $1 \times I$ , and the MSE is computed according to equation (1.8).

$$MSE = \frac{1}{I} \sum_{n=1}^I e_n^2 \quad (1.8)$$



The minimization of the MSE is obtained by deriving with respect to  $\mathbf{v}$  and equaling to zero.

$$\mathbf{e} \cdot \mathbf{Q}^T = \text{zeros}(1, C + 1) \quad (1.9)$$

$$\mathbf{v} \cdot \mathbf{Q} \cdot \mathbf{Q}^T = \mathbf{t} \cdot \mathbf{Q}^T \quad (1.10)$$

The resulting equation is (1.11), which is known as the equation of Wiener-Hopf (Van Trees, 1968).

$$\mathbf{v} = \mathbf{t} \cdot \mathbf{Q}^T \cdot (\mathbf{Q} \cdot \mathbf{Q}^T)^{-1} \quad (1.11)$$

With this expression, it is possible to determine the values of the coefficients in order to minimize the MSE for a subset of features.

- Least-Squares Quadratic Discriminant (LSQD). This detector provides satisfactory results with a speedy learning process (Gil-Pita et al., 2012). The intelligence of the LSLD is increased by adding quadratic terms to the linear combinations, which directly implies an improvement of the performance and the complexity, and inevitably the probability of appearance of generalization problems is increased. These new terms are shown in equation (1.12).

$$y = w_0 + \sum_{n=1}^C w_n x_n + \sum_{n=1}^C \sum_{m=1}^C v_{mn} x_m x_n, \quad (1.12)$$

where  $w_n$  and  $v_{mn}$  are the linear and quadratic weights, respectively. In the experiments, a simplified version of this detector will be applied, using only the diagonal terms, that is, those in which  $v_{mn} = 0, \forall m \neq n$ . Subject to such consideration, equation (1.12) is simplified into equation (1.13).

$$y = w_0 + \sum_{n=1}^C w_n x_n + \sum_{n=1}^C v_{nn} x_n^2 \quad (1.13)$$

An extended pattern matrix  $\mathbf{Q}$  can be defined containing the input features and their quadratic values. It is shown in equation (1.14).

$$\mathbf{Q} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{C1} & x_{C2} & x_{C3} & \dots & x_{CI} \\ x_{11}^2 & x_{12}^2 & x_{13}^2 & \dots & x_{1I}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{C1}^2 & x_{C2}^2 & x_{C3}^2 & \dots & x_{CI}^2 \end{bmatrix} \quad (1.14)$$

In addition, the weights can be rearranged in a vector  $\mathbf{v}$  according to equation (1.15).

$$\mathbf{v} = [ w_{10} \ w_{11} \ \dots \ w_{1C} \ v_{111} \ v_{122} \ \dots \ v_{1CC} ] \quad (1.15)$$

This way, the output can be obtained as  $\mathbf{y} = \mathbf{v} \cdot \mathbf{Q}$ , and the coefficients that minimize the MSE can be obtained through the Wiener-Hopf equations, as in the LSLD. This is shown in equation (1.16).

$$\mathbf{v} = \mathbf{t} \cdot \mathbf{Q}^T \cdot (\mathbf{Q} \cdot \mathbf{Q}^T)^{-1} \quad (1.16)$$

The main difference with the previous detector is that the boundaries are quadratic functions, so the system becomes more complex. As stated above, the greater intelligence of the classifier, the best results can be obtained in the experiments, but more generalization problems can appear.

- Artificial Neural Networks (ANNs): Multilayer Perceptron (MLP). ANNs are learning machines consisting of several simple processing elements called neurons, interconnected between them (Principe et al., 2000). There exist many types of neurons, but one of the most common is the one proposed in (McCulloch and Pitts, 1943). After that, the perceptron was proposed (Rosenblat, 1958) as a neuron with adjustable weights and a step-activation function type, being the first model capable of learning through supervised training. The perceptron splits the observation space into two regions through a hyperplane, being its linear basis function the pondered sum of the inputs, similarly to the equation (1.1) of the previously explained LSLD. In this model, the neurons were only activated when the stimulation was total, that is, when the result of the activation function  $I(y)$  was positive. Later, it was discovered that neurons emit electrical activity impulses with a variable frequency and present some activity at rest, so nonlinear activation functions started to be used. The most implemented one nowadays is the sigmoid, a mathematical function with a characteristic “S”-shaped curve (Han and Moraga, 1995). Within this function, the best known is the Log-Sigmoid, which is obtained using a logistic function, and the Tan-Sigmoid, which uses the hyperbolic tangent.

The MLPs have one or more layers of neurons sequentially arranged, in a way that the outputs of the neurons of a layer are the inputs of the neurons of the next layer. It is a direct propagation network, so the outputs of the network are calculated as functions of the inputs and the weights. The output of the neurons of the first layer  $\mathbf{X}$  is a matrix of size  $N_{neu} \times N$ , where  $N_{neu}$  is the number of neurons of the first hidden layer. This matrix can be obtained using equation (1.17).

$$\mathbf{X} = I(\mathbf{VQ}), \quad (1.17)$$

where  $\mathbf{V}$  is a matrix of size  $N_{neu} \times (C + 1)$  which contains the weights of the first layer. The universal approximation theorem states that any classification can be implemented using a single-layer perceptron (only one hidden layer) with enough neurons (Kurková, 1992). Because of that, in this thesis only two layers will be implemented: the hidden layer and the output layer, which combines the subspaces generated by each of the neurons of the first layer. In this way, the decision boundary will be formed by the nonlinear combinations of the hyperplanes, and with a sufficient number of hyperplanes any kind of decision boundary can be implemented.

The activation function implemented is the Tan-Sigmoid in the hidden layer and the linear in the output layer. Related to the number of neurons selected, it is a parameter that will be tested along the applications which use MLPs (VADHA and PDA), as its value is difficult to determine upon being highly dependent on the specific issue (Sheela and Deepa, 2013). It must be noted that this parameter only affects to the hidden layer, as the number of neurons in the output layer will be equal to one in the case of binary detection (the scope of this thesis).

The output of the MLP  $\hat{\mathbf{y}}$  can be obtained using equation (1.18).

$$\hat{\mathbf{y}} = \hat{\mathbf{w}} \cdot \begin{pmatrix} ones(1, I) \\ \tanh(\mathbf{V}\mathbf{Q}) \end{pmatrix} = \hat{\mathbf{w}} \cdot \hat{\mathbf{Q}}, \quad (1.18)$$

where  $\hat{\mathbf{w}}$  is a vector of size  $1 \times (N_{neu} + 1)$ , which contains the weights of the second layer. As in the case of LSLD and LSQD, the decision is computed by applying a threshold to the output  $\hat{\mathbf{y}}$ .

The weights of the neurons that make up the network are determined through training algorithms. They are usually based on partial derivatives, such as the Gradient method, the Newton method or the Gauss-Newton method. In this thesis, the Levenberg-Marquardt algorithm has been implemented (Levenberg, 1944, Marquardt, 1963). It is a quasi-Newton iterative method in which the objective function to minimize must be a sum of quadratic terms (like the mean squared error function), and which provides a fast optimization of the parameters using a set of training data. In addition, it is one of the most robust methods from the literature, as in many cases it finds a solution even if it is initialized very far from the final result.

The use of MLPs has several disadvantages. Firstly, the convergence is not guaranteed, since many local minima are usually presented. To avoid this problem, the training process has been repeated several times (10), so that the weights are initialized in a different way each time. It drastically reduces the probability of stopping the algorithm in local minima, or at least obtaining the same local minimum in all the repetitions. Secondly, this algorithm usually implies a loss of generalization, so the performance falls when new data from a new dataset are tested. To address this issue, a part of the data is used for validation, so the learning process is stopped in time. Thus, the design data are divided into two subsets: training data (80%) and validation data (20%). Test data are not used neither for training nor validation purposes. The training process finishes when the performance of the validation set does not improve after a certain number of iterations.

- Cross-Validation (CV). The feature selection and detection processes previously detailed are part of a cross-validation process, which is a validation method that tries to accurately estimate the performance of the general system. The motivation for using it is that the final model usually overfits the training data, and the performance is too optimistic. With this validation technique, it is possible to know how well the model would generalize in a new dataset. A dataset is usually divided into two subsets: the training set (also called the design set), which is known to the detector as it uses it to learn, and the test set, which is composed of unknown data to test the system. In this thesis, a non-exhaustive cross-validation method has been applied, in particular the *k-fold cross-validation* (Hastie et al., 2009). This method splits the dataset into  $k$  different folds and runs  $k$  times the whole design process (feature selection, parameter optimization and design of the detector), so in each experiment 1 fold is used as the test set and the remaining  $k - 1$  folds are used as the design set. The final evaluation metric is the average of the obtained metrics. This way, all the samples are used for both training and test.

Once a general perspective of the material and methods used along the thesis has been presented, the particularities of each of the four applications will be detailed in the following lines. We will explain which dataset has been used for training the algorithms or how it has been

created in case there are not suitable datasets in the literature, what are the parameters set in the experiments, etc.

### 1.4.1 Overview of the experiments in VSD

In the VSD issue, our goal is to detect the signs of violence that appear previously to those extreme events, but it is difficult to find a real-life acoustic dataset for VSD. For that reason, a dataset composed of real violent situations from *YouTube* website has been created, considering as violent the presence of heated arguments, people shouting and screaming among them, or even physical fights (García-Gómez et al., 2016).

The audio content has been extracted from the video signal, and all the audios were set to 22,050 Hz of sampling frequency, which was the minimal frequency of the original audios. Most of them were recorded with mobile phone cameras and similar, so they don't have high quality. This is an interesting characteristic because these recordings are more like the real world situations, where it usually appears background noise, compared to a fictional scenario, such as films. Once downloaded, the file segments were labeled. It consisted in listening to the audios and tagging where a scene was considered violent or not. This task was carried out by two people, comparing the results to make the labeled data more objective than if just one person is involved.

Regarding the implementation of a VSD system in smart cities, the system proposed in this thesis has taken account of the possibility of working in an autonomous way, such as being powered through solar cells. Because of that, some restrictions related to the computational cost the features need to be computed have been applied. Specifically, the number of Floating Operations Per Second (FLOPS) required by each of the features have been evaluated. The software developed in (Qian, 2015) allowed us to detect the different kind of operations of each line of code (arithmetic operations, elementary functions, variance calculation, determinant calculation, etc.). After that, different limits related to the number of FLOPS (from 1 to 15 MFLOPS) have been imposed during the algorithm learning process, particularly in the feature selection process, as they are directly related to the power consumption. To justify these values, a study was carried out (Fernández-Toloba et al., 2018) using a low-power processor (ARM Cortex-M4) provided with MEMS microphones and powered by a small solar cell of  $1 \text{ dm}^2$ , which spends  $1 \text{ W/dm}^2$ . Assuming a minimum average of 2.5 hours of sun per day (a typical value in several winter in regions such as Spain), the average total power will be 100 mW. If 57.4 mW are consumed by the microphones to record the audio samples, and other portion of energy is consumed when transmitting data, around 10 mW could be used to execute the detection algorithm. An ARM Cortex-M4 typically consumes approximately 0.2 mW/MHz, and a conservative relation of 3 FLOP per Hz can be assumed as a multiply requires 3 cycles according to the technical reference manual (ARM Developer, 2020). With these values, around 16.6 MFLOPS could be executed at most with the power available in the processor.

In Table 1.1 the parameters used in the experiments carried out in VSD are explained, including the used dataset, the computed features, the applied detectors, the applied computational cost constraints, and the parameter which evaluates the performance of the algorithms.

It must be noted that, apart from LSLD and LSQD, MLPs were also applied in the experiments, but the results have not been included because the performance drastically fell, probably because of overfitting problems.

### 1.4.2 Overview of the experiments in DPD

As previously stated, it is not easy to find acoustic datasets in the DPD issue. Because of that, we created a new dataset (García-Gómez et al., 2017). As in the VSD issue, sounds from different

**Table 1.1:** Main parameters of the system implemented for VSD application.

Parameter	Value	
Dataset	New dataset	Yes
	Total duration	27,802 s
	Violence duration	3,051 s
	Ratio of violence	10.97 %
	Number of audios	109
	Min audio length	15 s
	Max audio length	4,966 s
Features	Frequency-domain	Mel-Frequency Cepstral Coefficients (MFCCs) Delta MFCCs ( $\Delta$ MFCCs) Spectral Rolloff (SR) Spectral Centroid (SC) Spectral Flux (SF)
	Time-domain	Pitch Harmonic Noise Rate (HNR) Ratio of Unvoiced Frames (RUF) Short Time Energy (STE) Energy Entropy (EE) Zero Crossing Rate (ZCR)
Detectors	LSLD, LSQD	
Cost Constraints	1, 3, 5, 10, 15 MFLOPS	
Evaluation	Probability of Detection for low Probability of False Alarm	

models of drones (DJI Phantom, UDI 817, Parrot AR, Cheerson CX10, Eachine Racer 250, etc.) in motion were collected and labeled from *YouTube* and *FreeSound* websites. In addition, we make the dataset more challenging after adding sounds from other sources that can appear in the scene at the same time as a drone. Specifically, sounds from planes, helicopters, shavers, building work, diggers, motorbikes, mowers, F1 cars, cut-off wheels, fire sirens and drag racers were included. In some cases, the sound produced by them results quite similar to the one made by drones.

The implementation of a DPD system has been thought similarly to the VSD system, that is, in an autonomous way. As the same features have been computed, a similar study related to computational cost constraints has been implemented. However, the limits associated with the number of FLOPS have been set to be more restrictive (the maximum limit is set to 4 MFLOPS, while in VSD it was set to 15 MFLOPS) in order to be consistent with the sampling frequency used (8 kHz against 22 kHz).

In Table 1.2, similarly to Table 1.1, the parameters used in the experiments carried out in DPD are explained, including the used dataset, the computed features, the applied detectors, the applied computational cost constraints, and the parameter which evaluates the performance of the algorithms.

### 1.4.3 Overview of the experiments in VADHA

In this thesis, we have applied the EFLECs to VADHA, including computational cost restrictions related to the IPS (Instructions Per Second) (García-Gómez et al., 2018, Gil-Pita et al., 2017). After that, we have tried to apply additional optimizations based on cascade-detectors, implementing a simple detector in a first stage (LSLD) followed by a second stage based on MLPs, which must be used when the decision of the system is not clear. The objective was to reduce the

**Table 1.2:** Main parameters of the system implemented for DPD application.

Parameter	Value	
Dataset	New dataset	Yes
	Total duration	3,671 s
	Drone duration	1913 s
	Ratio of violence	50.08 %
	Number of audios	36
	Min audio length	6 s
	Max audio length	316 s
Features	Frequency-domain	Mel-Frequency Cepstral Coefficients (MFCCs) Delta MFCCs ( $\Delta$ MFCCs) Spectral Rolloff (SR) Spectral Centroid (SC) Spectral Flux (SF)
	Time-domain	Pitch Harmonic Noise Rate (HNR) Ratio of Unvoiced Frames (RUF) Short Time Energy (STE) Energy Entropy (EE) Zero Crossing Rate (ZCR)
Detectors	LSLD, LSQD	
Cost Constraints	0.5, 1.0, 1.5, ..., 4.0 MFLOPS	
Evaluation	Error Rate	

computational cost of the system while maintaining the performance of a more complex VAD system, or even to keep the same computational cost while improving the performance of the system.

The field of VAD has been widely researched in the literature, and the fact that the final implementation is set to hearing aid devices does not involve any additional requirement to the typical datasets of the field. Because of that, there was no need to create a new dataset, and the QUT-NOISE TIMIT one has been used (Dean et al., 2010). It contains a large number of conversations between people, including 10 different background noises from real and common places (café, home, street, car and reverberant places).

The computational constraints applied to VADHA have been much more restrictive than in the previous applications (between 10 and 200 KIPS instead of millions of them), as the power available in these devices is much lower and other important algorithms must also be implemented in them. To determine this range of values, the power consumption of a hearing aid in which only the compression algorithm was implemented and the consumption of a similar device where the VAD algorithm was also implemented were compared (García-Gómez et al., 2021). Using a real device with a DSP working at 1.92 MHz, an average power consumption of 0.87 mW was estimated in the case of only implementing the main compression algorithm, which required 1100 KIPS, and an average power consumption of 0.90 mW was calculated in the case of also implementing the proposed VAD algorithm running with 200 KIPS. Thus, it is shown that the consumption of the latter is negligible, saving power that can be used by other algorithms.

In Table 1.3, the parameters used in the experiments carried out in VADHA are explained, including the used dataset, the computed features, the applied detectors, the applied computational cost constraints, and the parameter which evaluates the performance of the algorithms.

<sup>1</sup>It must be noted that a subset of the dataset QUT-NOISE TIMIT was considered in the experiments due to the long duration of the original set, which contains around 600 hours of audios.

**Table 1.3:** Main parameters of the system implemented for VADHA application.

Parameter	Value		
Dataset	New dataset	No. QUT-NOISE TIMIT	
	Total duration <sup>1</sup>	21,600 s	
	Voice duration	≈10,800 s	
	Ratio of voice	≈50 %	
	Number of audios	240	
	Min audio length	60 s	
	Max audio length	120 s	
	SNR	Medium (0, 5 dB) and low (-5, -10 dB)	
Features	Frequency-domain	Evolved Frequency	Log-Energy Coefficients (EFLECs)
Detectors	LSLD in cascade with MLPs (1, 2, 3, 4, 5, 10, 15, 20 neurons)		
Cost Constraints	10, 20, 30, ..., 100, 120, 140, ..., 200 KIPS		
Evaluation	Error Rate		

#### 1.4.4 Overview of the experiments in PDA

In this thesis, we collaborated with the company Innerspec Technologies Europe S.L., involved in providing advance Non-Destructive Testing (NDT) solutions, for addressing the problem of PDA. The first step was the generation of Lamb guided-waves at different frequencies using EMAT technology, whose advantages were detailed previously. The Innerspec Powerbox H and the MRUT PMX scanner were the hardware devices employed. They allow to axially scan pipelines with a single or double sensor, and thus measure attenuation and velocity changes in the signal due to the presence of corrosion, cracks or other defects around the circumference of the pipe. The generation of the waves involves the appearance of different modes of the signals, affecting the dispersion effect in a different way to each of them according to the frequency excited.

Companies that work in the defect sizing field compete between them for commercializing their products into the market. Because of that, they do not usually release the datasets used for testing the algorithms. The data used for this purpose have been shared between the mentioned company and the researchers of our research group. They have some real pipelines in their facilities, which have been inspected by their devices to provide us the dataset. In addition, an experimental dataset has been developed through the Finite Element Method (FEM) included in the Partial Differential Equations Toolbox of Matlab. This is due to the limited number of real pipelines and defects, and to try to study the relation between the parameters of the signals (amplitude, time of arrival related to group velocity, phase velocity, etc.) and the shape and dimensions of the defects.

In this application, the features computed during the feature extraction process are substantially different from the acoustic features calculated in the previous applications, since in this application standard acoustic signals are not processed. In this case, the signal changes in every new scan, that is, when the sensor moves to a new position in the pipeline, but there is not a typical time-domain like in the rest of the applications. A heuristic study was made to evaluate whether some features contain useful information to address the problem, and later six of them were computed and submitted to a feature selection process. Due to the limited number of features, computational cost constraints have not been applied in the system.

In Table 1.4, the parameters used in the experiments of this application are shown. The details of the two generated datasets are presented, as well as the computed features, the applied

detectors, and the parameter which evaluates the performance of the algorithms.

**Table 1.4:** Main parameters of the system implemented for PDA application.

Parameter	Value		
Datasets	Simulated dataset	New dataset Number of defects Depth of the defects Excited Frequencies	Yes 418 0.5, 1, 1.5, ..., 9 mm 158, 250, 350, 450, 548 kHz
	Real dataset	New dataset Number of defects Depth of the defects Excited frequencies	Yes 3 1.85, 4.63, 6.18 mm 158, 548 KHz
Features	Time-domain	Maximum Amplitude (dB) Phase delay ( $\mu$ s) Average energy (dB) Group delay ( $\mu$ s) Maximum amplitude of the echo (dB) Average energy of the echoes ( $\mu$ s)	
Detectors	MLPs with 1, 2, 3, 4, 5 neurons		
Evaluation	RMSE (mm)		

## 1.5 Structure of the thesis

This thesis by compendium of articles has been divided into three parts, which in turn, are made up of several chapters.

- Part I contains the current chapter (Chapter I), in which the scope of this thesis is exposed, including the importance of four different applications in smart cities (VSD, DPD, VADHA and PDA), the reason for using acoustic signals for solving the issues, the different machine learning systems implemented on them, and a description of the experiments carried out.
- Part II contains Chapters 2, 3, 4, 5 and 6, which are the publications that have given rise to the thesis.
  - Chapter 2 includes the publication “Energy-Efficient Acoustic Violence detector for Smart Cities”, published in the International Journal of Computational Intelligence Systems. In this article the VSD application is developed.
  - Chapter 3 is the publication “Cost-constrained Drone Presence Detection through Smart Sound Processing”, which is included within Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019). In this article, the DPD application is proposed.
  - In Chapters 4 and 5, the conference paper “Analysis of the performance of Evolved Frequency Log-Energy Coefficients in Hearing Aids for different Cost Constraints and Scenarios”, included within 145th AES Convention E-Library, and the publication “Linear detector and neural networks in cascade for voice activity detection in hearing aids”, published in Applied Acoustics journal, are attached. The VADHA application is explained in both chapters. The latter is an expansion of the former, including new methodologies and improvements.



- Chapter 6 focuses on the publication “Smart Sound Processing for Defect Sizing in Pipelines Using EMAT Actuator Based Multi-Frequency Lamb Waves”, published in Sensors journal. The application PDA is included there.
- Part III contains the last chapter (Chapter 7), which collects the conclusions from the obtained results, in a general way and specifically for each application, as well as the future research lines.

# Bibliography

- Acar, E., Hopfgartner, F., and Albayrak, S. (2016). Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies. *Neurocomputing*, 208:225–237.
- Al-Emadi, S., Al-Ali, A., Mohammad, A., and Al-Ali, A. (2019). Audio Based Drone Detection and Identification using Deep Learning. In *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, pages 459–464.
- Alberta Government (2017). Energy Annual Report 2016-2017.
- Altawy, R. and Youssef, A. M. (2016). Security, privacy, and safety aspects of civilian drones: A survey. *ACM Transactions on Cyber-Physical Systems*, 1(2):1–25.
- Amieva, H., Ouvrard, C., Giulioli, C., Meillon, C., Rullier, L., and Dartigues, J.-F. (2015). Self-reported Hearing Loss, Hearing Aids, and Cognitive Decline in Elderly Adults: A 25-Year Study.
- Amieva, H., Ouvrard, C., Meillon, C., Rullier, L., and Dartigues, J.-F. (2018). Death, depression, disability, and dementia associated with self-reported hearing problems: A 25-year study. *The Journals of Gerontology: Series A*, 73(10):1383–1389.
- ARM Developer (2020). Arm Cortex-M4 Processor Technical Reference Manual Revision r0p1. FPU instruction set table.
- Babatunde, O. H., Armstrong, L., Leng, J., and Diepeveen, D. (2014). A genetic algorithm-based feature selection.
- Bellman, R., Corporation, R., and Collection, K. M. R. (1957). *Dynamic Programming*. Rand Corporation research study. Princeton University Press.
- Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, 5(1):1–12.
- Boulos, M. N. K., Tsouros, A. D., and Holopainen, A. (2015). Social, innovative and smart cities are happy and resilient?: insights from the WHO EURO 2014 International Healthy Cities Conference.
- Case, E. E., Zelnio, A. M., and Rigling, B. D. (2008). Low-cost acoustic array for small UAV detection and tracking. In *2008 IEEE National Aerospace and Electronics Conference*, pages 110–113. IEEE.

- Chamoso, P., González-Briones, A., Rodríguez, S., and Corchado, J. M. (2018). Tendencies of technologies and platforms in smart cities: A state-of-the-art review. *Wireless Communications and Mobile Computing*.
- Chen, L.-H., Hsu, H.-W., Wang, L.-Y., and Su, C.-W. (2011). Violence Detection in Movies. In *2011 Eighth International Conference Computer Graphics, Imaging and Visualization*, pages 119–124. IEEE.
- CIA (2019). Field Listing: Pipelines. The World Factbook.
- Clough, M., Fleming, M., and Dixon, S. (2017). Circumferential guided wave EMAT system for pipeline screening using Shear Horizontal ultrasound. *NDT & E International*, 86:20–27.
- Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.
- Dean, D., Sridharan, S., Vogt, R., and Mason, M. (2010). The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pages 3110–3113. International Speech Communication Association.
- Dias, L. J. G. (2016). *Detecting violent excerpts in movies using audio*. PhD thesis.
- Drozdowicz, J., Wielgo, M., Samczynski, P., Kulpa, K., Krzonkalla, J., Mordzonek, M., Bryl, M., and Jakielaszek, Z. (2016). 35 GHz FMCW drone detection system. In *2016 17th International Radar Symposium (IRS)*, pages 1–4. IEEE.
- European Gas Pipeline Incident Data Group (2015). 9th report of the european gas pipeline incident data group (period 1970 – 2013).
- Farlik, J., Kratky, M., Casar, J., and Stary, V. (2019). Multispectral detection of commercial unmanned aerial vehicles. *Sensors*, 19(7):1517.
- Fernández-Toloba, A., Sánchez-Hevia, H. A., Espino-Sanjosé, R., Clares-Crespo, C., García-Gómez, J., and Gil-Pita, R. (2018). Solar powered autonomous node for wireless acoustic sensor networks based on arm cortex m4. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- FRA-European Union Agency for Fundamental Rights (2014). *Violence against women: An EU-wide survey. Main results report*. FRA, European Union Agency for Fundamental Rights.
- Ganti, S. R. and Kim, Y. (2016). Implementation of detection and tracking mechanism for small UAS. In *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1254–1260. IEEE.
- García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., Mohino-Herranz, I., and Rosa-Zurera, M. (2016). Violence detection in real environments for smart cities. In *Ubiquitous Computing and Ambient Intelligence*, pages 482–494. Springer.
- García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., and Rosa-Zurera, M. (2017). Feature Selection for Real-Time Acoustic Drone Detection Using Genetic Algorithms. In *Audio Engineering Society Convention 142*. Audio Engineering Society.

- García-Gómez, J., Gil-Pita, R., Aguilar-Ortega, M., Utrilla-Manso, M., Rosa-Zurera, M., and Mohino-Herranz, I. (2021). Linear detector and neural networks in cascade for voice activity detection in hearing aids. *Applied Acoustics*, 175:107832.
- García-Gómez, J., Mohino-Herranz, I., Clares-Crespo, C., Fernández-Toloba, A., and Gil-Pita, R. (2018). Analysis of the performance of Evolved Frequency Log-Energy Coefficients in Hearing Aids for different Cost Constraints and Scenarios. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., and Theodoridis, S. (2006). Violence content classification using audio features. In *Hellenic Conference on Artificial Intelligence*, pages 502–507. Springer.
- Giffinger, R., Fertner, C., Kramar, H., Meijers, E., et al. (2007). City-ranking of European medium-sized cities. *Cent. Reg. Sci. Vienna UT*, pages 1–12.
- Gil-Pita, R., Alvarez-Perez, L., and Mohino, I. (2012). Evolutionary diagonal quadratic discriminant for speech separation in binaural hearing aids. *Advances in Computer science*, 20(5).
- Gil-Pita, R., Ayllón, D., Ranilla, J., Llerena-Aguilar, C., and Díaz, I. (2015). A computationally efficient sound environment classifier for hearing aids. *IEEE Transactions on Biomedical Engineering*, 62(10):2358–2368.
- Gil-Pita, R., García-Gomez, J., Bautista-Durán, M., Combarro, E., and Cocana-Fernandez, A. (2017). Evolved frequency log-energy coefficients for voice activity detection in hearing aids. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.
- Gold, B., Morgan, N., and Ellis, D. (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley.
- Gong, Z. and Xia, Y. (2015). Two speech enhancement-based hearing aid systems and comparative study. In *2015 5th International Conference on Information Science and Technology (ICIST)*, pages 530–534. IEEE.
- González-Briones, A., Prieto, J., De La Prieta, F., Herrera-Viedma, E., and Corchado, J. M. (2018). Energy optimization using a case-based reasoning strategy. *Sensors*, 18(3):865.
- Graf, S., Herbig, T., Buck, M., and Schmidt, G. (2015). Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(1):1–15.
- Green Jr, R. E. (2004). Non-contact ultrasonic techniques. *Ultrasonics*, 42(1-9):9–16.
- Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.
- Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2008). *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg.
- Han, J. and Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International Workshop on Artificial Neural Networks*, pages 195–201. Springer.
- Hancke, G. P., Hancke Jr, G. P., et al. (2013). The Role of Advanced Sensing in Smart Cities. *Sensors*, 13(1):393–425.

- Hao, K., Huang, S., Zhao, W., Wang, S., and Dong, J. (2011). Analytical modelling and calculation of pulsed magnetic field and input impedance for EMATs with planar spiral coils. *NDT & E International*, 44(3):274–280.
- HAO, M.-w. and SHI, X.-w. (2008). Oil and Gas Pipeline Leak Detection Technologies [J]. *Pipeline Technique and Equipment*, 5.
- Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83.
- Harrison, C., Eckman, B., Hamilton, R., Hartswick, P., Kalagnanam, J., Paraszczak, J., and Williams, P. (2010). Foundations for smarter cities. *IBM Journal of research and development*, 54(4):1–16.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.
- He, J., Dixon, S., Hill, S., and Xu, K. (2017). A new electromagnetic acoustic transducer design for generating and receiving S0 Lamb waves in ferromagnetic steel plate. *Sensors*, 17(5):1023.
- Heng, T. M. and Low, L. (1993). The intelligent city: Singapore achieving the next lap: Practitioners forum. *Technology Analysis & Strategic Management*, 5(2):187–202.
- Holland, J. H. et al. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press.
- Huan, Q., Chen, M., and Li, F. (2019). A practical omni-directional SH wave transducer for structural health monitoring based on two thickness-poled piezoelectric half-rings. *Ultrasonics*, 94:342–349.
- Hussain, A., Wenbi, R., da Silva, A. L., Nadher, M., and Mudhish, M. (2015). Health and emergency-care platform for the elderly and disabled people in the Smart City. *Journal of Systems and Software*, 110:253–263.
- Jain, A. and Vishwakarma, D. K. (2020). State-of-the-arts Violence Detection using ConvNets. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 0813–0817. IEEE.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.
- Jefferys, W. H. and Berger, J. O. (1992). Ockham’s razor and Bayesian analysis. *American Scientist*, 80(1):64–72.
- Keane, S. (2021). ‘macabre’ New Year’s Eve Drone Evaded Security In Spain’s Madrid. *Euro Weekly News*.
- Khan, M. A., Alvi, B. A., Safi, A., and Khan, I. U. (2018). Drones for good in smart cities: A review. In *Proc. Int. Conf. Elect., Electron., Comput., Commun., Mech. Comput.(EECCMC)*, pages 1–6.

- Khan, Z., Pervez, Z., and Ghafoor, A. (2014). Towards cloud based smart cities data security and privacy management. In *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, pages 806–811. IEEE.
- Kim, J. and Hahn, M. (2018). Voice activity detection using an adaptive context attention model. *IEEE Signal Processing Letters*, 25(8):1181–1185.
- King, J. M. and Faruque, I. (2016). Small unmanned aerial vehicle passive range estimation from a single microphone. In *AIAA Atmospheric Flight Mechanics Conference*, page 3545.
- Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324.
- Krug, E. G., Mercy, J. A., Dahlberg, L. L., and Zwi, A. B. (2002). The world report on violence and health. *The lancet*, 360(9339):1083–1088.
- Kurková, V. (1992). Kolmogorov’s theorem and multilayer neural networks. *Neural networks*, 5(3):501–506.
- Layouni, M., Hamdi, M. S., and Tahar, S. (2017). Detection and sizing of metal-loss defects in oil and gas pipelines using pattern-adapted Wavelets and machine learning. *Applied Soft Computing*, 52:247–261.
- Lee, C.-H., Kates, J. M., Rao, B. D., and Garudadri, H. (2017). Speech quality and stable gain trade-offs in adaptive feedback cancellation for hearing aids. *The Journal of the Acoustical Society of America*, 142(4):EL388–EL394.
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168.
- Liu, H., Wei, Z., Chen, Y., Pan, J., Lin, L., and Ren, Y. (2017). Drone detection based on an audio-assisted camera array. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 402–406. IEEE.
- Liu, T., Pei, C., Cai, R., Li, Y., and Chen, Z. (2020). A flexible and noncontact guided-wave transducer based on coils-only EMAT for pipe inspection. *Sensors and Actuators A: Physical*, 314:112213.
- Long, N., Gianola, D., Rosa, G., and Weigel, K. (2011). Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics*, 128(4):247–257.
- Lu, H., Iseley, T., Behbahani, S., and Fu, L. (2020a). Leakage detection techniques for oil and gas pipelines: State-of-the-art. *Tunnelling and Underground Space Technology*, 98:103249.
- Lu, H., Ma, X., and Azimi, M. (2020b). US natural gas consumption prediction using an improved kernel-based nonlinear extension of the Arps decline model. *Energy*, 194:116905.
- Lu, H., Ma, X., Huang, K., and Azimi, M. (2020c). Carbon trading volume and price forecasting in China using multiple machine learning models. *Journal of Cleaner Production*, 249:119386.
- Lu, H., Wu, X., Ni, H., Azimi, M., Yan, X., and Niu, Y. (2020d). Stress analysis of urban gas pipeline repaired by inserted hose lining method. *Composites Part B: Engineering*, 183:107657.

- Lu, S., Feng, J., Zhang, H., Liu, J., and Wu, Z. (2018). An estimation method of defect size from MFL image using visual transformation convolutional neural network. *IEEE Transactions on Industrial Informatics*, 15(1):213–224.
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.
- Meiri, R. and Zahavi, J. (2006). Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 171(3):842–858.
- Meister, H., Rähmann, S., Walger, M., Margolf-Hackl, S., and Kießling, J. (2015). Hearing aid fitting in older persons with hearing impairment: the influence of cognitive function, age, and hearing loss on hearing aid benefit. *Clinical Interventions in Aging*, 10:435.
- Meng, A. (2013). Death toll from Qingdao pipeline explosion rises to 62. *South China Morning Post*.
- Miller, Z. J. (2015). Drone That Crashed at White House was Quadcopter. *Time*.
- Mohamed, A., Hamdi, M. S., and Tahar, S. (2015a). An adaptive neuro-fuzzy inference system-based approach for oil and gas pipeline defect depth estimation. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 35–42. IEEE.
- Mohamed, A., Hamdi, M. S., and Tahar, S. (2015b). A machine learning approach for big data in oil and gas pipelines. In *2015 3rd International Conference on Future Internet of Things and Cloud*, pages 585–590. IEEE.
- Mohino-Herranz, M. (2017). *Emotion analysis through biological signal processing*. PhD thesis, University of Alcalá. <https://ebuah.uah.es/xmlui/handle/10017/41232>.
- Mukherjee, H., Obaidullah, S. M., Santosh, K., Phadikar, S., and Roy, K. (2018). Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. *International Journal of Speech Technology*, 21(4):753–760.
- Murdock, H. E., Gibb, D., André, T., Appavou, F., Brown, A., Epp, B., Kondev, B., McCrone, A., Musolino, E., Ranalder, L., et al. (2019). Renewables 2019 Global Status Report.
- Nakamura, N., Ogi, H., Hirao, M., et al. (2017). Emat pipe inspection technique using higher mode torsional guided wave T(0,2). *Ndt & E International*, 87:78–84.
- National Instruments (2019). Understanding FFTs and Windowing.
- Neirotti, P., De Marco, A., Cagliano, A. C., Mangano, G., and Scorrano, F. (2014). Current trends in Smart City initiatives: Some stylised facts. *Cities*, 38:25–36.
- Nguyen, P., Ravindranatha, M., Nguyen, A., Han, R., and Vu, T. (2016). Investigating cost-effective RF-based detection of drones. In *Proceedings of the 2nd Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use*, pages 17–22.

- Nievas, E. B., Suarez, O. D., García, G. B., and Sukthankar, R. (2011). Violence detection in video using computer vision techniques. In *International conference on Computer analysis of images and patterns*, pages 332–339. Springer.
- Nowicka, K. (2014). Smart City Logistics on Cloud Computing Model. *Procedia - Social and Behavioral Sciences*, 151:266 – 281.
- O’Dea, S. (2020). Number of Smartphone Users Worldwide from 2016 to 2021.
- Oppenheim, A. V. and Schaffer, R. W. (1989). Discrete-time Signal Processing.
- Pei, C., Zhao, S., Xiao, P., and Chen, Z. (2016). A modified meander-line-coil EMAT design for signal amplitude enhancement. *Sensors and Actuators A: Physical*, 247:539–546.
- Penet, C., Demarty, C.-H., Soleymani, M., Gravier, G., and Gros, P. (2012). Technicolor/INRIA/ Imperial College London at the MediaEval 2012 Violent Scene Detection Task.
- Perperis, T., Giannakopoulos, T., Makris, A., Kosmopoulos, D. I., Tsekeridou, S., Perantonis, S. J., and Theodoridis, S. (2011). Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies. *Expert systems with applications*, 38(11):14102–14116.
- Petcher, P., Potter, M., and Dixon, S. (2014). A new electromagnetic acoustic transducer (EMAT) design for operation on rail. *Ndt & E International*, 65:1–7.
- Principe, J. C., Euliano, N. R., and Lefebvre, W. C. (2000). *Neural and adaptive systems: fundamentals through simulations*, volume 672. Wiley New York.
- Qian, H. (2015). Counting the Floating Point Operations (FLOPS). MATLAB Central File Exchange. <https://es.mathworks.com/matlabcentral/fileexchange/50608-counting-the-floating-point-operations-flops>.
- Ramirez, J., Segura, J. C., Benitez, C., De La Torre, A., and Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech communication*, 42(3-4):271–287.
- Ramzan, M., Abid, A., Khan, H. U., Awan, S. M., Ismail, A., Ahmed, M., Ilyas, M., and Mahmood, A. (2019). A review on state-of-the-art violence detection techniques. *IEEE Access*, 7:107560–107575.
- Rosenblat, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.
- Safizadeh, M. and Azizzadeh, T. (2012). Corrosion detection of internal pipeline using NDT optical inspection system. *NDT & E International*, 52:144–148.
- Schedi, M., Sjöberg, M., Mironică, I., Ionescu, B., Quang, V. L., Jiang, Y.-G., and Demarty, C.-H. (2015). VSD2014: a dataset for violent scenes detection in Hollywood movies and web videos. In *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE.
- Sehgal, A. and Kehtarnavaz, N. (2018). A convolutional neural network smartphone app for real-time voice activity detection. *IEEE Access*, 6:9017–9026.



- Shah, S. C. and Kusiak, A. (2004). Data mining and genetic algorithm based gene/ SNP selection. *Artificial intelligence in medicine*, 31(3):183–196.
- Sheela, K. G. and Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering*, 2013.
- Shi, Y., Zhang, C., Li, R., Cai, M., and Jia, G. (2015). Theory and application of magnetic flux leakage pipeline detection. *Sensors*, 15(12):31036–31055.
- Sjöberg, M., Ionescu, B., Jiang, Y.-G., Quang, V. L., Schedl, M., Demarty, C.-H., et al. (2014). The MediaEval 2014 Affect Task: Violent Scenes Detection. In *MediaEval*.
- Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3.
- State of Green (2020). Smart Cities. Creating liveable, sustainable and prosperous societies. Technical report, State of Green. Connect. Inspire. Share. Think Denmark.
- Stevens, S. S., Volkman, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190.
- Stone, M. A. and Moore, B. C. (2002). Tolerable hearing aid delays. II. Estimation of limits imposed during speech production. *Ear and Hearing*, 23(4):325–338.
- Sun, H., Peng, L., Wang, S., Huang, S., and Qu, K. (2020). Development of Frequency-Mixed Point-Focusing Shear Horizontal Guided-Wave EMAT for Defect Inspection Using Deep Neural Network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14.
- Tan, F., Fu, X., Zhang, Y., and Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset selection. *Soft Computing*, 12(2):111–120.
- Trueba Sánchez, P. I. et al. (2016). Ciberadolescentes: Reduciendo la morbimortalidad a través de la eSalud.
- United Kingdom Onshore Pipeline Operators’ Association (2018). Pipeline product loss incidents and faults report.
- United Nations (2018). Revision of World Urbanization Prospects. Technical report, Department of Economic and Social Affairs. Population Dynamics.
- Urayama, R., Uchimoto, T., and Takagi, T. (2010). Application of EMAT/EC dual probe to monitoring of wall thinning in high temperature environment. *International Journal of Applied Electromagnetics and Mechanics*, 33(3-4):1317–1327.
- Van Trees, H. L. (1968). Detection, estimation and modulation, part I.
- Vattapparamban, E., GÃijvenÃğ, Ä., Yurekli, A. Ä., Akkaya, K., and UluÄšsaÃğ, S. (2016). Drones for smart cities: Issues in cybersecurity, privacy, and public safety. In *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 216–221.
- Viitanen, J. and Kingston, R. (2014). Smart cities and green growth: outsourcing democratic and environmental resilience to the global technology sector. *Environment and Planning A*, 46(4):803–819.

- Virtanen, T., Plumbley, M. D., and Ellis, D. (2018). *Computational analysis of sound scenes and events*. Springer.
- Walsh, D. (1979). Occam's razor: A principle of intellectual elegance. *American Philosophical Quarterly*, 16(3):241–244.
- Wang, T., Wang, X., Li, Z., Xue, L., Gao, Z., and Wang, Y. (2017). Comparison on failures of long-distance oil & gas pipelines at home and abroad. *Oil Gas Storage Transp*, 36(11):1258–1264.
- Wisdom, S., Okopal, G., Atlas, L., and Pitton, J. (2015). Voice activity detection using sub-band noncircularity. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4505–4509. IEEE.
- World Health Organization (2020). Deafness and hearing loss.
- World Health Organization et al. (2019). Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence. *WHO Library*, page 51.
- Xanthopoulos, P., Pardalos, P. M., and Trafalis, T. B. (2013). Linear discriminant analysis. In *Robust data mining*, pages 27–33. Springer.
- Xie, Y., Liu, Z., Yin, L., Wu, J., Deng, P., and Yin, W. (2017). Directivity analysis of meander-line-coil EMATs with a wholly analytical method. *Ultrasonics*, 73:262–270.
- Zhang, X.-D. (2020). *A Matrix Algebra Approach to Artificial Intelligence*. Springer Singapore, Singapore.
- Zhang, Y., Li, S., Wang, T., and Zhang, Z. (2013). Divergence-based feature selection for separate classes. *Neurocomputing*, 101:32–42.

Part II  
Publications

## Chapter 2

# Article 1: Energy-Efficient Acoustic Violence Detector for Smart Cities

### AUTHORS

Marta Bautista-Durán, Joaquín García-Gómez, Roberto Gil-Pita, Inma Mohino-Herranz, Manuel Rosa-Zurera

### JOURNAL

International Journal of Computational Intelligence Systems (IJCIS)  
D.O.I.: <https://doi.org/10.2991/ijcis.10.1.89>

### RANKING

JCR (2017): 2  
Quartile Rank - Computer Science, Artificial Intelligence: 54/132 (Q2)  
Quartile Rank - Computer Science, Interdisciplinary Applications: 49/105 (Q2)

### CONTRIBUTION TO THE SCOPE OF THE THESIS

In this publication, the Violent Situation Detection (VSD) issue is addressed. Violence continues to be a latent conflict in actual society, so this article proposes an energy-efficient system capable of acoustically detecting violent scenes in real time and real situations. In the solution, different experiments are carried out using genetic algorithms to select the best subset of features with a computational cost constrained in terms of the number of operations per second. A novel dataset is tested, with the objective of maximizing the probability of detection for low probabilities of false alarm. Results demonstrate the viability of the system, thanks to the low cost that some violence features require, making feasible the implementation of the proposed method in a nowadays low power microprocessor. In addition, the usefulness of MFCCs for solving the problem at hand is proved.

## Energy-Efficient Acoustic Violence Detector for Smart Cities

Marta Bautista-Durán, Joaquín García-Gómez, Roberto Gil-Pita, Inma Mohino-Herranz, Manuel Rosa-Zurera

*Signal Theory and Communications Department, University of Alcalá,  
28805 Alcalá de Henares, Madrid, Spain*

*E-mail: marta.bautista@edu.uah.es, joaquin.garciagomez@edu.uah.es*

Received 28 February 2017

Accepted 31 May 2017

### Abstract

Violence detection represents an important issue to take into account in the design of intelligent algorithms for smart environments. This paper proposes an energy-efficient system capable of acoustically detecting violence. In our solution, genetic algorithms are used to select the best subset of features with a constrained computational cost. Results demonstrate the viability of the system, thanks to the low cost that some violence features require, making feasible the implementation of the proposed method in a nowadays low power microprocessor.

*Keywords:* Violence Detection, Audio Processing, Feature Selection, Computational Cost

### 1. Introduction

Violence continues being a latent conflict in actual society. Recent researches show that 35% of women around the world have suffered physical or sexual violence during their lives<sup>1</sup> and 43% of women from the European Union declared suffering psychological violence at least once.<sup>2</sup> This fact makes violence detection and prevention to represent an important issue to take into account in the design of intelligent algorithms for smart environments. In this sense, violence can be detected through audio and video surveillance. Some works in the literature treat this problem using both audio and video processing,<sup>3,4,5</sup> and the results obtained with the combination of those sources seems to be efficient.

Main disadvantages of video can be found in terms of computational cost, intrusiveness and poor coverages. Some authors have evaluated computational cost using core hours as metric.<sup>6</sup> Furthermore, audio and video have been tested both in separate and together ways in the literature.<sup>7</sup> Their conclusions show that the system works properly using just audio source. When video information is added the performance improves slightly, but computational cost increases in a big way. Besides, an audio-

based system is economic in terms of  $\text{€}/m^2$ .

In the literature we can find other proposals where audio is used to detect violence by itself,<sup>8</sup> since violent situations are commonly accompanied by signs like arguments, shouts or an increase in the volume of the conversation. However, most of the studies up to now have been done with pretended violence from films or games, which are not applicable to real violence situations.<sup>9</sup>

In order to implement real-time audio surveillance systems in wide areas, the need of energy-efficient processing nodes arises. An energy-efficient real-time system has the restriction of consumption when it is implemented in some place where it is working in an autonomous way. In this scenario, the computational cost, related to the clock frequency of the processing units, is an important factor to take into account, and the control of the computational cost of the violence detection system is mandatory.

Bearing this in mind, this paper proposes a real-time implementation of an energy-efficient system capable of detecting a violent situation in smart environments. Since the system has to work in an autonomous way, computational cost is strictly constrained, and there is a need

to find a reduced set of features. In this sense, genetic algorithms are proposed to solve the constrained feature selection process, allowing a good tradeoff between performance and computational cost.

This paper is structured as follows. First, Section 2 introduces the implemented classification system, describing the feature extraction (Subsection 2.1), the computational cost evaluation (Subsection 2.2) and the feature selection process using genetic algorithms (Subsection 2.3). Then, Section 3 describes the results, including the description of the database, the validation method employed and the discussion of the results. To sum up, Section 4 presents the conclusions.

## 2. The Acoustic Surveillance System

The proposed system has the objective of studying solutions for audio-based violence detection in real environments and in real time, where the system has to take a decision every  $T$  seconds. The steps of the proposed acoustic surveillance system, shown in Figure 1, are being explained in detail in the following sections.

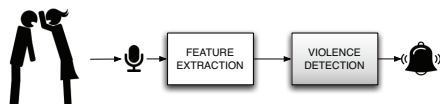


Fig. 1. Proposed system.

### 2.1. Feature extraction

There are several audio features that could exhibit a good discrimination capability for the problem at hand.<sup>8,10</sup> This section includes a brief description of the most interesting features for violence detection.

Most of the features tend to analyze some time statistics over the evaluation of a measurement along the time to get useful information from the audio. So, in order to evaluate/extract the features, the audio segments of  $T$  seconds are divided into  $M$  frames of  $L$  samples with an overlap of  $S\%$ . By default, the statistics applied to these measurements are typically the mean and the Standard Deviation (SD), although for some particular measurements more specific statistics are used.

All measurements can either be taken in the time domain or in the frequency domain. For notation purposes,

let us assume  $x_{im}$  is the  $i$ -th audio sample of the  $m$ -th time frame ( $i = 1, \dots, L$  and  $m = 1, \dots, M$ ), and  $X_{km}$  is the  $k$ -th frequency component for the  $m$ -th time frame of the Short-Time Fourier Transform (STFT), evaluated applying a windowed Discrete Fourier Transform (DFT) to the  $m$ -th time frame.

The features considered in this paper are:

- The **Mel-Frequency Cepstral Coefficients (MFCCs)**, which are a set of perceptual parameters commonly used in speech recognition,<sup>10</sup> calculated from the spectrum. They provide a compact representation of the spectral envelope. Perceptual analysis emulates human ear non-linear frequency response by creating a set of filters on non-linearly spaced frequency bands.<sup>11</sup> In the case of violence detection and considering a sampling frequency of 22,050 Hz,  $N = 25$  cepstral coefficients are calculated,<sup>12</sup> so that there will be 25 different MFCCs per frame, denoted  $MFCC_{nm}$ ,  $n = 1, \dots, 25$ .
- The **Delta Mel-Frequency Cepstral Coefficients ( $\Delta$ MFCCs)**, calculated as the time difference of standard MFCCs in two different time frames,<sup>10</sup> so that  $\Delta MFCC_{nm} = MFCC_{(n+1)m} - MFCC_{(n-1)m}$ .
- The **Pitch**, related to the fundamental frequency, determines the tone of the speech. It can be used to distinguish a person from another.<sup>11</sup> In this paper we estimate the pitch for every frame, evaluating the main peaks of the autocorrelation of the error of a linear predictor with  $P = 10$  coefficients.<sup>10</sup>
- The **Harmonic Noise Rate (HNR)** quantifies the purity of the speech in every frame. It measures the relationship between the harmonic energy produced by the vocal cords versus non-harmonic energy present in the signal.<sup>10</sup>
- The **Ratio of Unvoiced time Frames (RUF)**, is related to the presence or absence of clear or strong speech in the analyzed audio. It is obtained dividing the number of time frames with detected pitch by the total number of frames.<sup>12</sup>
- The **Short Time Energy (STE)** is the energy of the short speech segment,  $STE_m = \sum_{i=1}^L x_{im}^2$ . It is a simple and effective classifying parameter for both voiced and unvoiced frames.<sup>13</sup>
- The **Energy Entropy (EE)** expresses abrupt changes in the energy level of the audio signal. It is useful

for detecting violence due to rapid changes occurring in the tone of voice.<sup>8</sup> To evaluate this measurement, each time frame of  $L$  samples is divided into  $B$  blocks, and the energy of each block is then measured. So, EE for the  $m$ -th time frame can be evaluated using  $EE_m = -\sum_{b=1}^B \sigma_{bm}^2 \log_2 \sigma_{bm}^2$ , where  $\sigma_{bm}^2$  is the normalized energy calculated for the  $b$ -th block of the  $m$ -th frame,  $b = 1, \dots, B$ . Apart from the mean and the SD, statistics applied to the energy entropy are the ratios of maximum to mean and maximum to median values.

- The **Zero Crossing Rate (ZCR)** is one of the most widely used time-domain audio features.<sup>8</sup> It is determined by dividing the number of sign changes by the total length of the frame, so that  $Z_m = \sum_{i=1}^L |sgn(x_{im}) - sgn(x_{(i-1)m})|$ . Apart from the mean and the SD, the ratio of the maximum to mean is calculated.
- The **Spectral Rolloff (SR)** is calculated in the frequency domain and is defined as the frequency  $k_c(m)$  below which  $c\%$  of the magnitude distribution of STFT coefficients are concentrated for the  $m$ -th frame, so that  $\sum_{k=0}^{k_c(m)} |X_{km}| = c/100 \sum_{k=0}^{L/2} |X_{km}|$ . It represents the skewness of the spectral shape.<sup>8</sup> The median value is computed apart from the mean and the SD.
- The **Spectral Centroid (SC)** is defined as the center of gravity of the magnitude spectrum of the STFT,<sup>14</sup> so that  $SC_m = \sum_{k=0}^{L/2} k \cdot |X_{km}| / \sum_{k=0}^{L/2} |X_{km}|$ .
- The **Spectral Flux (SF)** represents the spectral change between successive frames,<sup>8</sup> and is determined using  $SF_m = \sum_{k=0}^{L/2} (|X_{km}| - |X_{k(m-1)}|)^2$ .

## 2.2. Computational Cost Evaluation

A energy-efficient real time system has the restriction of consumption when it is implemented in some place where it is working in an autonomous way, for instance working with a solar powered source. In this scenario, computational cost is an important aspect to consider if we want to control the consumption the node has.

In order to calculate the computational cost of our system, the number of flops that each feature requires has been calculated determining the number of Floating Point Operations Per Second (FLOPS).<sup>15</sup> The number of flops is related to the power consumption. To put this in perspective, if the system has to work autonomously and is powered by a small solar cell of  $1 \text{ dm}^2$  which spends  $1 \text{ W/dm}^2$ , and having a minimum average of 2.5 hours

of sun per day (a typical value in several winter in regions such as Spain), the average total power will be 100 mW. Low power processors, such as the ARM-Cortex-M4, typically consumes around 0.2 mW/MHz which, assuming a relationship of 1 FLOP per Hertz, gives us an idea of the amount of FLOPS that are going to be available for this kind of devices.<sup>16</sup>

The number of FLOPS of our system depends on the set of selected features, so it must take into account which ones are used for a specific design. To evaluate the impact of each feature in the selection process, we have carried out a detailed analysis of the computational cost in terms of FLOPS required to implement an energy-efficient violence detection system.

Thus, the cost of each feature has been evaluated and we propose the above equations with the objective of generalize the cost in function of some parameters explained below. As was stated above, the feature extraction process splits the audio frame of  $N_{samples}$  (so that  $T = N_{samples}/f_s$ , being  $f_s$  the sampling frequency) into  $M$  frames of  $L$  samples, with an overlap between them of  $S\%$ , so that:

$$M = \left\lceil \frac{N_{samples}}{S \cdot L} \right\rceil \quad (1)$$

Some features such as pitch-based or MFCCs have more impact in cost than others due to the amount of flops needed. Furthermore, some features share some processing blocks that do not need to be replicated for different features. Considering the measurements described in the last section, we have identified four processing blocks that are shared along more than one measurement:

- The evaluation of the STFT is shared by the MFCCs,  $\Delta$ MFCCs, the SR, the SC and the SF. Equation (2) represents the cost of the STFT matrix  $C_S$ , in terms of operations per decision, in function of the main design parameters.

$$C_S = L(M-1)(5 \log_2 L + 2) + 4L + 15 \quad (2)$$

- The evaluation of the MFCCs is shared by both the MFCCs and the  $\Delta$ MFCCs. Apart from the evaluation of the STFT, these features require some shared operations. The cost  $C_M$  associated to these operations is expressed using equation (3) in function of  $N$ , the number of MFCCs computed.

$$C_M = (L \cdot S + 1)(M(2N + 5) + 10N + 23) + N(3N + 11) + N \cdot M(2N + 7) + 29, \quad (3)$$

- The evaluation of the pitch is also shared by the HNR and the RUF. Its cost  $C_P$  can be determined using the next equation:

$$C_P = 2L \cdot M(5 \log_2 L + P + 3) + M(P(2P^2 + P + 2L + 1) - L) + 1, \quad (4)$$

where  $P$  is the number of Levinson Coefficients.

- At last, the evaluation of the energy is shared by the STE and the EE (which requires it to normalize the energy of each block), and its cost  $C_E$  can be determined using equation (5)

$$C_E = M(2L + 3) - 4 \quad (5)$$

We will use four binary variables  $b_S, b_M, b_P$  and  $b_E$  related to  $C_S, C_M, C_P$  and  $C_E$  (the number of operations associated to the described shared processing blocks) to determine whether the selected set of features does require the evaluation of one of the aforementioned blocks, respectively. The total number of operations can be expressed using equation (6):

$$C_T = b_S \cdot C_S + b_M \cdot C_M + b_P \cdot C_P + b_E \cdot C_E + \sum_{f=1}^{11} s_f \cdot C_f, \quad (6)$$

where  $C_f$  is the specific additional cost of each measurement, and  $s_f$  is a binary vector which indicates the selected measurements. The FLOPS can be easily evaluated simply taking into account that the proposed system requires a decision every  $T$  seconds.

To sum up, there are some features which are linked and depend on others, so that the computation of one allows to compute the others with practically the same cost. Because of that, we have been grouped measurements into 8 groups. These groups are:  $G_1$  (including MFCCs and  $\Delta$ MFCCs),  $G_2$  (including Pitch, HNR, and RUF),  $G_3$  (STE),  $G_4$  (EE),  $G_5$  (ZCR),  $G_6$  (SR),  $G_7$  (SC) and  $G_8$  (SF). STE and EE have been evaluated separately because the cost of the EE is not insignificant respect to the one of the STE. Table 1 describes the groups, the number of features of each measurement, the values  $b_S, b_M, b_P$  and

$b_E$  and the additional cost  $C_f$  associated to each measurement, in function of the main design parameters of each feature.

### 2.3. Constrained selection of features

As was stated above, to control the computational cost of the violence detection system, there is a need to find a reduced set of patterns that allows a good performance with an energy-efficient implementation. For this purpose, genetic algorithms have been used in the paper.

Genetic algorithms are based on the principles of genetic and natural selection, allowing to obtain the best results for solving a problem.<sup>17</sup> This method consists of exchanging randomly the features of the individuals of a population that constitute the possible solutions for the problem. In this way, the algorithm is able to resolve optimization problems.<sup>18</sup> Specifically, our problem is to determinate which features are the best to be applied to violence detection without resulting in a high cost. For that reason, a cost constraint is applied when the features are selected. There are 121 features in total, but each individual only selects a subset of them in a way that total cost is below the fixed threshold. The adaptive function has the aim of maximize the probability of detection associated to a probability of false alarm for a given detection system. In this point, two different classifiers will be applied: The Least Squares Linear Detector and the simplified version of Least Squares Quadratic Detector. They are explained in detail in the literature.<sup>12</sup>

According to the previous parameter, the individuals will be ranked and only the best individuals survive and reproduce. The population is composed of 100 individuals, 10 of them will be chosen as parents, and they will generate the remaining 90 sons by crossover. After this, mutation changes a 4 percent of the genes. This process is repeated along 30 generations and the whole process is repeated 10 times to avoid local minima.

## 3. Results

In order to validate the proposed system, a set of experiments has been carried out using a database of audio files. These audio files have been divided in segments of  $T = 5$  seconds length with a sampling frequency of  $f_s = 22,050$  Hz. Each frame is divided in windows of  $L = 512$  length and  $S = 50\%$  overlap between windows, resulting in a to-



Table 1. Dependence between grouped features.

Group	Caract	No. feats	$b_S$	$b_M$	$b_P$	$b_E$	Additional cost (No. operations)
$G_1$	MFCCs	50	1	1	0	0	$C_1 = 0$
	$\Delta$ MFCCs	50	1	1	0	0	$C_2 = N(M-2) + 1$
$G_2$	Pitch	2	0	0	1	0	$C_3 = 0$
	HNR	2	0	0	1	0	$C_4 = 9M$
	RUF	1	0	0	1	0	$C_5 = M$
$G_3$	STE	2	0	0	0	1	$C_6 = 0$
$G_4$	EE	4	0	0	0	1	$C_7 = M(\lfloor 2L/B \rfloor + 3B - 5) + 6B + 3$
$G_5$	ZCR	3	0	0	0	0	$C_8 = (6M+1)(L-1)$
$G_6$	SR	3	1	0	0	0	$C_9 = M(5N+8) + 2[M(L \cdot S - 1)/3]$
$G_7$	SC	2	1	0	0	0	$C_{10} = M(8N+L \cdot S+6) + L \cdot S+4$
$G_8$	SF	2	1	0	0	0	$C_{11} = M(9N+5) - 3N+1$

tal of  $M = 430$  frames per segment. Then feature extraction has been applied to obtain useful information from data. With the aim of selecting a reduced set of features, a genetic algorithm is used.

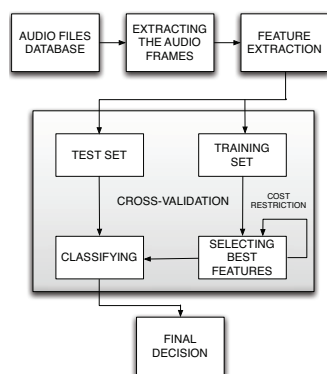


Fig. 2. Block diagram of the experiments.

This algorithm has been applied using a constraint related to the cost available in the system. Specifically, different cost thresholds measured in “Maximum number of Mega Floating Operations Per Second” (MaxMFLOPS) have been applied (1, 3, 5, 10 and 15 MaxMFLOPS). This means that the sum of costs of the selected features has to be below this values. Once the best features have been selected, a specifically trained classifier aims at giving the final decision. Figure 2 shows a block diagram describing the process carried out in the experiments.

In general, the databases used in the state-of-the-art

were not suitable for our problem, so we have used a novel database developed in a previous work.<sup>12</sup> The main characteristics of the used database are shown in Table 2.

Table 2. Summary of the database.

Parameters	Value
Total duration	27,802 s
Violent duration	3,051 s
Percentage of violence	10.97%
Number of audios	109
Minimum audio length	15 s
Maximum audio length	4,966 s

Related to the implemented validation, a tailored version of  $k$ -fold cross-validation has been used in the experiments to avoid loss of generalization of the results. The data is divided in  $k$  subsets, so that each subset is used for testing and the remaining  $k - 1$  are used for training. In our case, 109 folds with different size have been used, each fold containing data from a different audio file. In that way, we ensure that data from the same acoustic environment is not used both for training and testing at the same time, guaranteeing the generalization of the results.

As it was stated above, two genetic algorithms based feature selection strategies have been considered: the case of maximizing the probability of detection with a linear and with a quadratic detector. In each case, the same detector has been applied to classify. The probability of false alarm considered in the optimization process has been 10%. Figure 3 shows a comparative between the

costs (measured in Mega-FLOPS) required by the eight groups of features. The cost necessary to calculate the Short Time Fourier Transform (STFT) is depicted in solid colour, while the additional cost of each feature group is painted with striped bars. For instance, if the STFT has been calculated because of the group  $G_1$  (MFCCs and  $\Delta$ MFCCs), this cost can be saved in groups  $G_6$ ,  $G_7$  and  $G_8$  (spectral features). In the same way, in group  $G_4$  (EE) energy does not have to be calculated if group  $G_3$  (STE) is computed.

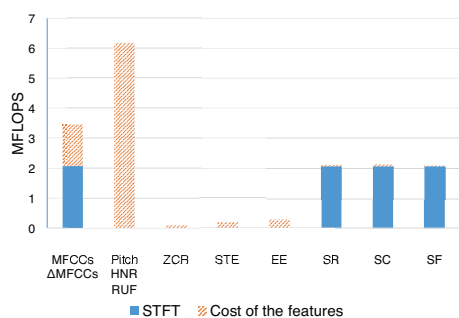


Fig. 3. Cost of the different feature groups.

In view of the results, we can appreciate that group  $G_2$  (pitch, HNR and RUF) is the most computationally expensive group, overcoming 6 millions of FLOPS. Group  $G_1$  is also too expensive, but it will provide 100 features to the experiments, aside from the calculation of the STFT, used by other groups.

Now we will evaluate the effect of the limits in the computational cost available. Figure 4 shows the probability of detection obtained for low probabilities of false alarm (under 10%) and using the linear detector, evaluated for the different cost thresholds. The same is shown in Figure 5 using the quadratic detector.

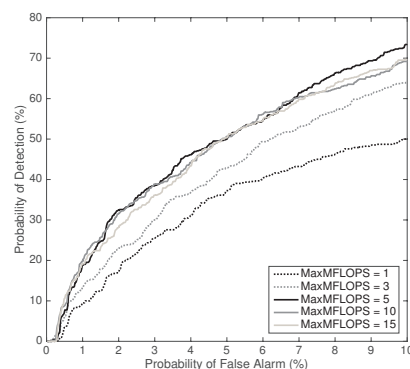


Fig. 4. Probability of Detection for Linear Detector.

The behavior is similar in both cases. With low thresholds (1 MaxMFLOPS) the probabilities of detection obtained are poor (around 50-55% for 10% of false alarm). As we increase this threshold the results are considerably improved, reaching around 75-80% of detection with 5 MaxMFLOPS cost. However, this improvement does not continue for higher costs, so it makes no sense to spend more resources in this problem.

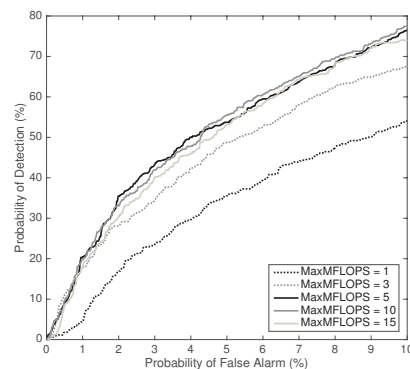


Fig. 5. Probability of Detection for Quadratic Detector.

In order to demonstrate the high accuracy of the proposed system in terms of probability of detection, we are going to make a comparison between our method and the one proposed by J. Salamon<sup>19</sup>. Applying that algorithm

Table 3. Cost, probability of detection and probability of appearance of the features groups.

MaxMFLOPS	1 MFLOPS		3 MFLOPS		5 MFLOPS		10 MFLOPS		15 MFLOPS	
Classifier	Lin.	Qua.	Lin.	Qua.	Lin.	Qua.	Lin.	Qua.	Lin.	Qua.
Average Cost (MFLOPS)	0.4	0.4	2.6	2.6	3.9	3.7	9.8	8.3	10.0	8.8
Pd (Pfa = 10%) (%)	50%	54%	64%	67%	74%	76%	69%	78%	70%	74%
$G_1$ (MFCC+ $\Delta$ MFCC)	0%	0%	0%	0%	100%	100%	100%	100%	100%	100%
$G_2$ (Pitch+HNR+RUF)	0%	0%	0%	0%	0%	0%	99%	76%	100%	82%
$G_3$ (STE)	93%	0%	19%	0%	63%	0%	80%	0%	63%	0%
Selection rate (%) $G_4$ (EE)	100%	100%	100%	100%	97%	92%	73%	89%	95%	96%
$G_5$ (ZCR)	100%	100%	100%	100%	80%	31%	25%	12%	80%	35%
$G_6$ (SR)	0%	0%	100%	100%	98%	98%	93%	98%	97%	99%
$G_7$ (SC)	0%	0%	100%	9%	82%	77%	99%	78%	97%	86%
$G_8$ (SF)	0%	0%	6%	98%	49%	57%	41%	63%	41%	70%

the results are around 65% of probability of detection for a probability of false alarm of 10%, which does not improve the ones obtained with the algorithm proposed in this experiment.

Now we will study which groups of features are more selected and useful. Table 3 displays the average cost employed, the probability of detection for a probability of false alarm of 10% and the percentages of appearance (selection rates) of the groups. It has been considered as appearance the selection of one or more features from the group.

At the beginning, the algorithm selects groups  $G_3$ ,  $G_4$  and  $G_5$  in practically 100% of the cases because of the low threshold imposed (1 MaxMFLOPS). When we increase this value to 3 MaxMFLOPS the spectral features appear. Furthermore, the MFCCs are selected with 5 or more MaxMFLOPS, and the pitch with 10 MaxMFLOPS. The case of 15 MaxMFLOPS allows the algorithm to select whatever it needs, because the sum of the total cost is lower than this value.

As it can be seen, there are some features that work better in the quadratic detector than in the linear one. Such is the case of group  $G_8$  (SF), where the difference between the appearance in both classifiers is always considerable. The opposite happens in groups  $G_3$  and  $G_5$ . In fact, the appearance of group  $G_3$  in quadratic detector is always 0%.

Additionally, the importance of some features is reflected in the table. For instance, when group  $G_1$  -MFCCs and  $\Delta$ MFCCs- appears (from 5 MaxMFLOPS onwards)

its appearance is 100% in linear and quadratic detectors, while the appearance of the features that were selected previously is significantly reduced, like in groups  $G_4$  and  $G_5$ . Because of that, MFCCs is an excellent group. The same does not happens to other expensive groups, such as group  $G_2$ , which does not improve the results when it is selected (10-15 MaxMFLOPS).

#### 4. Conclusion

The objective of this work is to develop a system capable of detect violent scenes in real time and in real situations. With this purpose, we have carried out different experiments related to audio analysis. The algorithms have been developed in order to maximize the probability of detection for low probabilities of false alarm, but subject to computational cost constraints.

The results derived from the experiments show that MFCCs are the best features for violence detection, both for linear and quadratic classifiers. Other features such as energy only show a good performance in linear classifiers and their cost is quite low compared to the rest.

Regarding to the classifiers, the results obtained are better in quadratic case (3-9% of difference respect to the linear one) for all cases with different cost thresholds. Higher cost implies better results, but a compromise of 5 MaxMFLOPS could be reached, since the results does not seem to be improved much from this value.

The cost ( $\text{€}/m^2$ ) of this audio-based system is relatively low. For instance, if we consider a typical range of 20  $m^2$  per node and each node (e.g., *Raspberry Pi*)

has a price around 100 €, the deployment costs would be around 5 €/m<sup>2</sup>.

To sum up, the experimental results show that it is viable to implement a real time system capable of detecting violence in an autonomous way. That is possible thanks to the low cost that some violence features need to be computed, which can be supported by nowadays low power microprocessors.

#### Acknowledgments

This work has been funded by the Spanish Ministry of Economy and Competitiveness (under project TEC2015-67387-C4-4-R, funds Spain/FEDER) and by the University of Alcalá (under project CCG2016/EXP-033).

#### References

1. World Health Organization, *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence*, (2013), p. 2.
2. European Union Agency For Fundamental Rights, *Violence against women: an EU-wide survey*, (2014), p. 71.
3. L. H. Chen, H. W. Hsu, L. Y. Wang, and C. W. Su, *Violence Detection in Movies*, in *Computer Graphics, Imaging and Visualization (CGIV), Eighth International Conference*, (2011), pp. 119–124.
4. M. Schedi, M. Sjöberg, I. Mironic, B. Ionescu, V. L. Quang, Y. G. Jiang and C. H. Demarty, *VSD2014: A dataset for violent scenes detection in hollywood movies and web videos*, in *13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, (Prague, 2015), pp. 1–6.
5. Acar, E., Hopfgartner, F., and Albayrak, S., *Breaking down violence detection: Combining divide-et-impera and coarse-to-fine strategies*. *Neurocomputing*, 208, pp. 225–237 (2016).
6. Lam, V., Le, S. P., Do, T., Ngo, T. D., Le, D. D., and Duong, D. A., *Computational optimization for violent scenes detection*. In *Computer, Control, Informatics and its Applications (IC3INA), 2016 International Conference on* (pp. 141-146). IEEE.
7. Gregrio Dias, L.J., *Detecting Violent Excerpts in Movies using Audio and Video Features* (2016)
8. T. Giannakopoulos, D. Kosmopoulos, A. Aristidou and S. Theodoridis, *Violence content classification using audio features*, in *Hellenic Conference on Artificial Intelligence*, (Springer Berlin Heidelberg, Greece, 2006), pp. 502–507.
9. C. H. Demarty, C. Penet, G. Gravier and M. Soleymani, *The MediaEval 2012 affect task: violent scenes detection*, in *Working Notes Proceedings of the MediaEval Workshop*, (2012).
10. I. Mohino, R. Gil-Pita, and L. Álvarez, *Stress Detection Through Emotional Speech Analysis*, in *Advances in Computer Science*, (2011), pp. 233–237
11. R. Gil-Pita, B. López-Garrido, and M. Rosa-Zurera, M., *Tailored MFCCs for sound environment classification in hearing aids*, in *Advanced Computer and Communication Engineering Technology*, (Springer International Publishing, 2015), pp. 1037–1048.
12. J. García-Gómez, M. Bautista-Durán, R. Gil-Pita, I. Mohino-Herranz and M. Rosa-Zurera, *Violence Detection in Real Environments for Smart Cities*, in *Ubiquitous Computing and Ambient Intelligence: 10th International Conference, UCAmI*, (Springer International Publishing, Spain, 2016), Part II 10, pp. 482–494.
13. M. Jalil, F. A. Butt, and A. Malik, *Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals*, in *Technological Advances in Electrical, Electronics and Computer Engineering (TAECE)*, (2013), pp. 208–212.
14. G. Tzanetakis, and P. Cook, *Musical genre classification of audio signals*, in *IEEE Transactions on speech and audio processing*, 10(5), (2002), pp. 293–302.
15. H. Qian, *Counting the Floating Point Operations (FLOPS)*, *MATLAB Central File Exchange*, No. 50608, Ver. 1.0, (2015).
16. ARM, *ARM Cortex-M4 Processor: Technical Reference Manual. Revision: r0p1*. Available at: [https://developer.arm.com/docs/100166\\_0001/00](https://developer.arm.com/docs/100166_0001/00).
17. R. L. Haupt and S. E. Haupt, *Practical genetic algorithms*. John Wiley & Sons, (2004).
18. D. E. Goldberg and J. H. Holland, *Genetic algorithms and machine learning*. *Machine learning*, 3(2), (1988), pp. 95–99.
19. Salamon, J., Jacoby, C., and Bello, J. P, *A dataset and taxonomy for urban sound research*. In *Proceedings of the 22nd ACM international conference on Multimedia* (2014), pp. 1041–1044.

## Chapter 3

# Article 2: Cost-constrained Drone Presence Detection through Smart Sound Processing

### AUTHORS

Joaquín García-Gómez, Marta Bautista-Durán, Roberto Gil-Pita, Inma Mohino-Herranz, Miguel Aguilar-Ortega, César Clares-Crespo

### BOOK

Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM)

D.O.I.: 10.5220/0007556007660772

### CONTRIBUTION TO THE SCOPE OF THE THESIS

This article shows the research related to the Drone Presence Detection (DPD) issue. Sometimes, drones lead to problems of invasion of privacy or access to restricted areas. Because of that, it is essential to develop a system capable of detecting the presence of these vehicles in real time in environments where they could be used for malicious purposes. However, the computational cost associated with that system must be limited if it has to work in an autonomously. In this manuscript, an algorithm based on smart sound processing techniques has been developed, including the typical pattern recognition stages of feature extraction, computationally-constrained feature selection, and detection. A novel dataset is tested with the objective of minimizing the error rate of the system. Results show that it is possible to detect the presence of drones with low-cost feature subsets easily supported by modern microprocessors, where MFCCs and pitch are the most relevant ones.

## Cost-constrained Drone Presence Detection through Smart Sound Processing

Joaquín García-Gómez, Marta Bautista-Durán, Roberto Gil-Pita, Inma Mohíno-Herranz,  
Miguel Aguilar-Ortega and César Clares-Crespo

*Department of Signal Theory and Communications, University of Alcalá, Alcalá de Henares 28805, Spain*

**Keywords:** Drone Detection, Smart Sound Processing, Feature Extraction, Feature Selection, Evolutionary Computation, Cost Constraints.

**Abstract:** Sometimes, drones lead to problems of invasion of privacy or access to restricted areas. Because of that, it is important to develop a system capable of detecting the presence of these vehicles in real time in environments where they could be used for malicious purposes. However, the computational cost associated to that system must be limited if it has to work in an autonomous way. In this manuscript an algorithm based on Smart Sound Processing techniques has been developed. Feature extraction, cost constrained feature selection and detection processes, typically implemented in pattern recognition systems, are applied. Results show that it is possible to detect the presence of drones with low cost feature subsets, where MFCCs and pitch are the most relevant ones.

### 1 INTRODUCTION

The use of Unmanned Aerial Vehicles, also known as drones, is on the rise in the society, mainly because of the advantages they offer. However, these vehicles usually run into problems of invasion of privacy or access to hazardous areas (e.g. airports). For this reason it is important to develop a system capable of detecting the presence of drones in particular environments where they could be used for malicious purposes, such as households, public buildings or restricted-access areas. In the state of the art there are many studies which deal with this issue, trying to detect and locate drones (Ganti and Kim, 2016). The wide range of methods includes audio, video, temperature, radar and radio frequency based detection.

Video detection systems can cover long distances, but there is a difficulty when distinguishing between drones and birds, even after including bird flight patterns which drones do not follow (Ganti and Kim, 2016). In addition, the computational cost of this kind of systems is high. Talking about the temperature-based detection, it is an efficient solution if the drone uses a propulsion engine, which usually appears in fixed-wing drones. However, most current drones are made of plastic and their electric engines do not radiate much heat.

Systems based on radar signal are useful for air-

craft detection, but the small size of the drones complicates their detection. Some manuscripts are working on this alternative (Drozdowicz et al., 2016). Related to radio frequency based methods, they are useful for the problem at hand since radio frequency is the communication mode used between drones and the remote controller (Nguyen et al., 2016). However, the use of Wi-Fi range (2.4-5 GHz) in no-license channels causes the appearance of high interferences.

Some proposals have based their study on audio information, mixed or not with video one. Some authors propose the use of an array of microphones and an infrared camera to get the information (Case et al., 2008). They try to trace the path followed by the drone through beamforming techniques. Others use only one microphone, but they are focusing on detecting a particular model of drone, so the results could not be generalizable (King and Faruque, 2016). In one manuscript, the authors analyze video information to detect the difference between frames, and in this way they track the drone movement, while they use audio information for detecting the vehicle with a threshold in frequency (Ganti and Kim, 2016). The problem is that it is not very effective when background noise is high. In addition, audio appears to be more reliable for detecting drones according to some studies (Liu et al., 2017).

This manuscript proposes a real-time implemen-

766

García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., Mohíno-Herranz, I., Aguilar-Ortega, M. and Clares-Crespo, C.  
Cost-constrained Drone Presence Detection through Smart Sound Processing.

DOI: 10.5220/0007556007660772

In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 766-772

ISBN: 978-989-758-351-3

Copyright © 2019 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

tation of an energy-efficient system capable of detecting drone presence in smart environments. We want the system to work in an autonomous way, so computational cost related to the clock frequency of the processing units will be strictly constrained. In this sense, evolutionary computation (i.e. genetic algorithms) is proposed for selecting a reduced set of features from the full set calculated previously, allowing a good tradeoff between performance and computational cost.

## 2 SMART SOUND PROCESSING (SSP) SYSTEM

In order to detect drone presence, our study will be based on an efficient system successfully used in other applications, like violence detection (Bautista-Durán et al., 2017). This is because this set includes features like pitch, which can be useful for detecting the frequency associated to the drone engine, as well as the rotation speed, size and material of the propellers. The system has the objective of studying solutions for audio-based drone detection in real environments and in real time, where the system has to make a decision every  $T$  seconds. Fig. 1 shows the system diagram, whose steps will be explained in the following sections.



Figure 1: Scheme of the system.

### 2.1 Feature Extraction

The objective of this step is to extract useful information from the audio signal in the form of features. There are several audio features that have demonstrated to be really useful in other applications, fundamentally related to speech problems (Giannakopoulos et al., 2006; Mohino et al., 2011; Gil-Pita et al., 2015). In this manuscript we will apply this type of features to the problem of drone detection. In this section a theoretical description of the features will be made. To extract the features, the audio segments of  $T$  seconds are divided into  $M$  frames of  $L$  samples with an overlap of  $S\%$ . The following features have been considered:

- The Mel-Frequency Cepstral Coefficients (MFCCs). They are  $N$  parameters calculated from the spectrum that are typically used for speech recognition. With this measurement, a compact representation of the spectral envelope is obtained. The objective is to emulate the human ear non-linear frequency response through a set of filters on non-linearly spaced frequency bands (Gil-Pita et al., 2015).
- The Delta Mel-Frequency Cepstral Coefficients ( $\Delta$ MFCCs). They are calculated differentiating the previous MFCCs in two different time frames.
- The Pitch. This feature is related to the fundamental frequency and determines the tone of the speech. It allows to distinguish a person from another. In this manuscript the pitch is evaluated in every frame through the autocorrelation of the error of a linear predictor with  $P$  coefficients (Mohino et al., 2011).
- The Harmonic Noise Rate (HNR). With this feature it is feasible to evaluate the purity of the speech. It measures the relation between the harmonic energy produced by the vocal cords and the non-harmonic energy.
- The Ratio of Unvoiced time Frames (RUF). It measures the presence or absence of clear or strong speech. The computation consists of dividing the number of time frames with detected pitch by the total number of frames.
- The Short Time Energy (STE), which is the energy of the short speech segment. It is a simple and effective parameter for both voiced and unvoiced frames (Jalil et al., 2013).
- The Energy Entropy (EE). It allows to detect changes in the energy level of the audio. It is useful for detecting a quick emergence of a drone in the environment due to rapid changes in the energy of the audio. To evaluate this measurement, each time frame is divided into  $B$  blocks, and the energy of each block is then measured.
- The Zero Crossing Rate (ZCR). It is one of the most used audio features in time domain. To calculate it, the number of sign changes is divided by the total length of the frame.
- The Spectral Rolloff (SR). It is calculated in the frequency domain and is defined as the frequency below which  $c\%$  of the magnitude distribution of Short Time Fourier Transform (STFT) coefficients are concentrated for a frame.
- The Spectral Centroid (SC) is the center of gravity of the magnitude spectrum of the STFT.
- The Spectral Flux (SF) measures the spectral changes between successive frames.

Once these features have been extracted, some statistics are applied to them (the mean and the standard deviation).

## 2.2 Feature Selection with Cost Constraints

If we want to get an energy-efficient real time system for detecting drone presence, it will have the restriction of consumption, as it will be implemented in some place to work in an autonomous way. In this scenario, computational cost is an important aspect to consider. In order to calculate the computational cost of our system, we have computed the resources that each feature requires determining the number of Floating Point Operations Per Second (FLOPS) (Qian, 2015), which is directly related to the power consumption of the device. The number of FLOPS of the system will depend on the set of selected features, so it must be taken into account which ones are used in each case (Bautista-Durán et al., 2017).

Thus, the cost of each feature has been evaluated and some equations are proposed with the objective of generalizing the cost according to some parameters that will be explained. As stated above, the feature extraction process splits the audio frame of  $N_{samples}$  (so that  $T = N_{samples}/f_s$ , being  $f_s$  the sampling frequency) into  $M$  frames of  $L$  samples, with an overlap between them of  $S\%$ , so that:

$$M = \left\lceil \frac{N_{samples}}{S \cdot L} \right\rceil \quad (1)$$

Some aspects must be taken into account for the analysis. First of all, some features will have more impact in cost than others (e.g. MFCCs or pitch-based ones). In addition, some features need to apply the same processing blocks, so their computation do not have to be repeated. Considering the measurements of Section 2.1, four processing blocks that are shared along more than one measurement have been identified:

- The STFT is shared by the MFCCs, the  $\Delta$ MFCCs, the SR, the SC and the SF.
- The MFCCs are shared by the MFCCs and the  $\Delta$ MFCCs.
- The pitch is shared by the HNR and the RUF.
- The energy is shared by the STE and the EE.

In Table 1 the four processing blocks and their equations are shown. Four binary variables  $b_1, b_2, b_3$  and  $b_4$  related to  $B_1, B_2, B_3$  and  $B_4$  (the number of operations associated to the previous processing blocks) will be defined to determine if the set of features selected requires or not the evaluation of these blocks.

Thus, the total cost  $C$  will be calculated using Equation (2).

$$C = \sum_{i=1}^4 b_i \cdot B_i + \sum_{j=1}^{11} s_j \cdot C_j, \quad (2)$$

where  $C_j$  is the additional cost of each feature and  $s_j$  is a binary value which indicates if the feature is selected or not. Taking into account that the proposed system makes a decision every  $T$  seconds, the FLOPS can be evaluated.

As there are some features which are linked and depend on others, we have grouped the measurements into 8 groups:  $G_1$  (including MFCCs and  $\Delta$ MFCCs),  $G_2$  (including Pitch, HNR, and RUF),  $G_3$  (STE),  $G_4$  (EE),  $G_5$  (ZCR),  $G_6$  (SR),  $G_7$  (SC) and  $G_8$  (SF). The groups, number of features of each measurement, values of  $b_S, b_M, b_P$  and  $b_E$ , and the equations of additional cost  $C_f$  associated to each measurement are detailed in Table 2. There we can see a typical cost of the problem at hand, considering each feature is selected individually, so the shared blocks need to be computed in each of them. The parameters used for solving the equations are:  $B = 10$  blocks,  $L = 512$  samples,  $M = 31$  frames,  $N = 25$  MFCCs coefficients,  $P = 10$  Levinson coefficients and  $S = 50\%$  overlap.

As it has been discussed, it is necessary to find a reduced set from the 117 features that allows obtaining a good performance and controlling the computational cost of the system. For this purpose, evolutionary algorithms have been implemented in the manuscript (Haupt et al., 1998). The configuration of this algorithm includes the next parameters: 100 individuals, 10 parents, 90 regenerated sons, percentage of mutation of 2%, 30 generations, 10 repetitions of the whole algorithm and minimization of the error rate as adaptive function.

## 2.3 Detectors

To evaluate the results and make a decision about the presence of drone sound, a detector has to be applied. In the present case, two different detectors have been used: the Least Squares Linear Discriminant (LSLD) and a reduced version of the Least Squares Quadratic Discriminant (LSQD). The computation of the two detectors is shown in Equations 3 and 4. (García-Gómez et al., 2016). They are obtained using the Wiener-Hopf equations. (Van Trees, 2004)

$$y = w_0 + \sum_{n=1}^L w_n x_n, \quad (3)$$

$$y = w_0 + \sum_{n=1}^L w_n x_n + \sum_{m=1}^L \sum_{n=1}^n x_m x_n v_{mn}, \quad (4)$$



Table 1: Cost of the shared processing blocks.

Block	Cost of the block (No. operations)
STFT	$B_1 = L(M-1)(5\log_2 L + 2) + 4L + 15$
MFCCs	$B_2 = (L \cdot S + 1)(M(2N + 5) + 10N + 23) + N(3N + 11) + N \cdot M(2N + 7) + 29$
Pitch	$B_3 = 2L \cdot M(5\log_2 L + P + 3) + M(P(2P^2 + P + 2L + 1) - L) + 1$
Energy	$B_4 = M(2L + 3) - 4$

Table 2: Details of the groups of features.

Group	Caract	No. feats	$b_1$	$b_2$	$b_3$	$b_4$	Additional cost (No. operations)	Typical cost (MFLOPS)
$G_1$	MFCCs	50	1	1	0	0	$C_1 = 0$	1.25
	$\Delta$ MFCCs	50	1	1	0	0	$C_2 = N(M-2) + 1$	1.26
$G_2$	Pitch	2	0	0	1	0	$C_3 = 0$	2.21
	HNR	2	0	0	1	0	$C_4 = 9M$	2.21
	RUF	1	0	0	1	0	$C_5 = M$	2.21
$G_3$	STE	2	0	0	0	1	$C_6 = 0$	0.03
$G_4$	EE	2	0	0	0	1	$C_7 = M(\lfloor 2L/B \rfloor + 3B - 5) + 6B + 3$	0.06
$G_5$	ZCR	2	0	0	0	0	$C_8 = (6M + 1)(L - 1)$	0.10
$G_6$	SR	2	1	0	0	0	$C_9 = M(5N + 8) + 2\lfloor M(L \cdot S - 1)/3 \rfloor$	0.74
$G_7$	SC	2	1	0	0	0	$C_{10} = M(8N + L \cdot S + 6) + L \cdot S + 4$	0.75
$G_8$	SF	2	1	0	0	0	$C_{11} = M(9N + 5) - 3N + 1$	0.74

where  $x_n$  and  $x_m$  are the training patterns,  $w_n$  and  $v_{mn}$  are the weights associated to them,  $w_0$  is a bias term and  $y$  is the combination of the training patterns. A threshold will be applied to this combination to obtain the binary decision about drone presence.

It is important to indicate that in the beginning more complex detectors were considered (e.g. artificial neural networks). However they were discarded because the results were not as good as expected, due to the fact that overtraining problems appear as the dataset is not large enough.

### 3 RESULTS

To validate the system we have carried out some experiments using a dataset of audio files. These audio files have been divided in segments of  $T = 1$  second, which indicates how often a decision is made. All the files have been resampled to a sampling frequency of  $f_s = 8,000$  Hz. Each frame is divided in windows of  $L = 512$  length and  $S = 50\%$  overlap between windows, resulting in a total of  $M = 31$  frames per segment. Then steps detailed in previous sections have been followed, including feature extraction, feature selection and detection.

The algorithm has been applied using a constraint related to the computational cost. Some cost thresholds measured in "Maximum number of Mega Floating Operations Per Second" (MaxMFLOPS) have been applied (0.5, 1, 1.5, 2, 2.5, 3, 3.5 and 4 MaxMFLOPS). This means that the sum of costs of the selected features has to be below these values. The up-

per limit is never reached, since the cost associated to the case of selecting all the features is below 4 MaxMFLOPS. Once the best features have been selected, a trained detector makes the final decision.

The datasets used in the state of the art are not suitable for our problem for several reasons: they just include a model of drone, or the environmental conditions do not change. Because of that, we have used a novel dataset that was developed in a previous work (García-Gomez et al., 2017). In this dataset, drones in motion and in a static position are included, as well as different models of them (Cheerson CX10, DJI Phantom 3, Eachine Racer 250, etc.). In order to make the database more challenging, similar no-drone sounds are included too (plane, helicopter, mower, etc.). The main characteristics of the used database are: total duration of 3671 seconds, duration of drone sound of 1913 seconds, percentage of drone presence of 50.08%, 36 fragments, minimum audio length of 6 seconds and maximum audio length of 316 seconds. More details about the dataset can be found in (García-Gomez et al., 2017).

The method of validation implemented has been a tailored version of  $k$ -fold cross-validation, since it allows avoiding loss of generalization of the results. The data is divided in  $k$  subsets, so that each subset is used for testing and the remaining  $k - 1$  are used for training. In the case at hand, 36 folds with different size have been used, each fold containing a different audio file. In that way, we ensure that data from the same model of drone or with the same environmental conditions are not used both for training and testing at the same time.

### 3.1 Analysis of the Computational Cost Constraints

Now we will evaluate the effect of the limits in the computational cost available, as well as the groups of features more selected and useful. Table 3 displays the error rate and the percentages of appearance (selection rates) of the groups, in function of the maximum cost established in MFLOPS, using the LSLD. The error rate is the sum of the decisions where the system says there is drone presence and it fails because there is no drone in the environment, and vice versa. It has been considered as appearance the selection of one or more features from the group. The same is displayed in Table 4 using LSQD.

At the beginning, the system selects groups  $G_3$ ,  $G_4$  and  $G_5$  in almost 100% of the cases because of the low threshold imposed (0.5 MaxMFLOPS). When we increase this value to 1 MaxMFLOPS, the spectral features appear. If the restriction is established in 1.5 MaxMFLOPS, the MFCCs start to be selected. When we reach higher values of MFLOPS (3.5), group  $G_2$  is selected, which is composed of features related to the pitch. The case of 4.0 MaxMFLOPS allows the algorithm to select whatever it needs, because the sum of all the costs is lower than this value.

In general LSQD works better than LSLD, since the error rate is lower in most cases, specially when the cost constraint is very limiting. The importance of some features is reflected in the table. For instance, when group  $G_1$  -MFCCs and  $\Delta$ MFCCs- appears (from 1.5 MaxMFLOPS onwards) its appearance is 100%. In fact, the parameter that best reflects the importance of  $G_1$  is the error rate, since it falls significantly when that group appears (in the case of LSLD, from 57.5% of error to 28.5%, and in the case of LSQD, from 41.9% to 23.4%). Something similar happens when  $G_2$  -pitch, HNR and RUF- appears (from 3.5 MaxMFLOPS onwards). Again, its selection rate is 100% and its contribution to the performance of the system is really significant (error falls from 30.1% to 15.7% with LSLD and from 23.8% to 15.5% with LSQD). The importance of pitch could be directly related to the particular frequency that drones present, which is dependent on the size of the device, the number of blades and the speed.

With regard to the rest of features,  $G_3$  seems to work well only when using LSLD because of its high selection rate. The same applies to  $G_8$ , but when using LSQD. Other features seem to be more robust to changes in the detector used ( $G_5$ ,  $G_6$  and  $G_7$ ), since they present high selection rate for both detectors.

### 3.2 Analysis of the Model of Drone and Other No-drone Sounds

Then, the error obtained in each of the models included in the drone database will be analyzed. Table 5 shows the different models of drone, the duration of each of them and the error obtained. In these results the best constraint and detector in terms of error have been selected from the previous cases (13.4% of error with 4.0 MFLOPS and LSLD).

From Table 5 it can be seen that Parrot AR is the best detected model (0% of error rate), while the worst one is the UDI 817 (50% of error). This could be because of its minor presence in the database. As it can be observed, a large proportion of the database belongs to DJI Phantom 3, which gets an error rate of 12.2%.

As mentioned previously, the dataset was developed including no-drone sounds present in smart city environments, which can be easily confused with the sound of a drone. In Table 6 the no-drone sounds, the duration of them and the error obtained are detailed.

From the results it can be observed that the most confusing sounds are the fire siren, radial saw and construction work (with error rates of 40.7%, 36.4% and 22.5%, respectively). This could be because the fundamental frequency of these sounds is in the range of the drone frequency (one or two hundreds of Hz). Likewise, other sounds like helicopter, excavator, motorbike or plane are really well detected as no-drone sounds, with error rates below 3%. This is especially interesting in the case of other aerial vehicles (helicopter, plane), since they could be more conflicting with drones as they share the same space of work (the sky) and they could appear at the same time.

## 4 CONCLUSIONS

The aim of this work is to develop a system capable of detecting the presence of drones in real time. To this end, different experiments related to Smart Sound Processing (SSP) have been carried out, including feature extraction, feature selection and detectors. The objective of the algorithms is to minimize the error rate while controlling the computational cost. This has been reached through a constraint in the number of operations per second (MFLOPS).

Related to the features selected, the results show that MFCCs and features related to pitch are the best subsets of features for the problem at hand, for both linear and quadratic detectors. Depending on the desired final error rate and on the resources of the processing device, a compromise should be reached be-

Table 3: Cost, error rate and probability of appearance of the features groups with LSLD.

MaxMFLOPS (MFLOPS)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
<b>Error Rate (%)</b>	52.3	57.5	28.5	30.4	31.9	30.1	15.7	13.4
<i>G</i> <sub>1</sub> (MFCC+ $\Delta$ MFCC)	0.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0
<i>G</i> <sub>2</sub> (Pitch+HNR+RUF)	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0
<i>G</i> <sub>3</sub> (STE)	73.9	80.8	89.1	93.7	100.0	100.0	25.7	100.0
<b>Selection Rate (%)</b>	100.0	100.0	100.0	100.0	100.0	100.0	0.0	100.0
<i>G</i> <sub>5</sub> (ZCR)	91.7	13.6	84.7	83.2	89.0	91.6	0.0	95.3
<i>G</i> <sub>6</sub> (SR)	0.0	100.0	100.0	100.0	100.0	100.0	74.3	93.9
<i>G</i> <sub>7</sub> (SC)	0.0	92.9	96.1	100.0	100.0	91.3	74.3	100.0
<i>G</i> <sub>8</sub> (SF)	0.0	70.2	35.6	40.9	50.6	53.0	15.8	22.9

Table 4: Cost, error rate and probability of appearance of the features groups with LSQD.

MaxMFLOPS (MFLOPS)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
<b>Error Rate (%)</b>	37.8	41.9	23.4	24.2	22.0	23.8	15.5	15.2
<i>G</i> <sub>1</sub> (MFCC+ $\Delta$ MFCC)	0.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0
<i>G</i> <sub>2</sub> (Pitch+HNR+RUF)	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0
<i>G</i> <sub>3</sub> (STE)	18.7	11.0	0.0	0.0	0.0	0.0	0.0	0.0
<b>Selection Rate (%)</b>	78.4	46.1	100.0	100.0	100.0	100.0	0.0	100.0
<i>G</i> <sub>5</sub> (ZCR)	100.0	100.0	95.9	96.4	95.9	100.0	0.0	88.6
<i>G</i> <sub>6</sub> (SR)	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
<i>G</i> <sub>7</sub> (SC)	0.0	92.9	74.0	69.9	61.4	67.1	95.4	91.3
<i>G</i> <sub>8</sub> (SF)	0.0	100.0	100.0	100.0	100.0	100.0	61.2	96.6

Table 5: Error Rate of the different models of drones included in the database.

Model of drone	Duration (s)	Error Rate (%)
DJI Phantom 3	1573	12.2
Cheerson CX10	284	13.0
Eachine Racer 250	171	21.6
Parrot AR	103	0.0
UDI 817	17	50.0

Table 6: Error Rate of the no-drone sound included in the database.

No-drone sound	Duration (s)	Error Rate (%)
Plane	128	3.1
Helicopter	124	0.0
Hair clipper	249	14.1
Construction work	316	22.5
Excavator	147	0.0
Motorbike	150	1.3
Mower	268	8.2
Radial saw	22	36.4
Fire siren	135	40.7
Drag racer	55	7.1

tween the two parameters. On the one hand, if the system requires high performance (13.4% of error rate), the solution should include both the MFCCs and the features related to pitch, with at least 3.5 MFLOPS. On the other hand, a worst solution in terms of error rate could be reached (23.4%), but only using 1.5 MFLOPS in the system. Regarding to the detectors,

the results are better in quadratic case, specially when the cost constraint is very restrictive.

In conclusion, the experiments developed show that it is feasible to implement a real time system capable of detecting drone presence in an autonomous way. That is possible thanks to the low cost features proposed in the manuscript, which can be supported by nowadays microprocessors.

## ACKNOWLEDGEMENTS

This work has been funded by the University of Alcalá under Project CCGP2017-EXP/060.

## REFERENCES

- Bautista-Durán, M., García-Gómez, J., Gil-Pita, R., Mohino-Herranz, I., and Rosa-Zurera, M. (2017). Energy-efficient acoustic violence detector for smart cities. *Delta*, 1:25.
- Case, E. E., Zelnio, A. M., and Rigling, B. D. (2008). Low-cost acoustic array for small uav detection and tracking. In *Aerospace and Electronics Conference, 2008. NAECON 2008. IEEE National*, pages 110–113. IEEE.
- Drozdowicz, J., Wielgo, M., Samczynski, P., Kulpa, K., Krzonkalla, J., Mordzonek, M., Bryl, M., and Jakielaszek, Z. (2016). 35 ghz fmew drone detection

- system. In *Radar Symposium (IRS), 2016 17th International*, pages 1–4. IEEE.
- Ganti, S. R. and Kim, Y. (2016). Implementation of detection and tracking mechanism for small uas. In *Unmanned Aircraft Systems (ICUAS), 2016 International Conference on*, pages 1254–1260. IEEE.
- García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., Mohino-Herranz, I., and Rosa-Zurera, M. (2016). Violence detection in real environments for smart cities. In *Ubiquitous Computing and Ambient Intelligence*, pages 482–494. Springer.
- García-Gomez, J., Bautista-Durán, M., Gil-Pita, R., and Rosa-Zurera, M. (2017). Feature selection for real-time acoustic drone detection using genetic algorithms. In *Audio Engineering Society Convention 142*. Audio Engineering Society.
- Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., and Theodoridis, S. (2006). Violence content classification using audio features. In *Hellenic Conference on Artificial Intelligence*, pages 502–507. Springer.
- Gil-Pita, R., López-Garrido, B., and Rosa-Zurera, M. (2015). Tailored mfccs for sound environment classification in hearing aids. In *Advanced Computer and Communication Engineering Technology*, pages 1037–1048. Springer.
- Haupt, R. L., Haupt, S. E., and Haupt, S. E. (1998). *Practical genetic algorithms*, volume 2. Wiley New York.
- Jalil, M., Butt, F. A., and Malik, A. (2013). Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In *Technological Advances in Electrical, Electronics and Computer Engineering (TAECE), 2013 International Conference on*, pages 208–212. IEEE.
- King, J. M. and Faruque, I. (2016). Small unmanned aerial vehicle passive range estimation from a single microphone. In *AIAA Atmospheric Flight Mechanics Conference*, page 3545.
- Liu, H., Wei, Z., Chen, Y., Pan, J., Lin, L., and Ren, Y. (2017). Drone detection based on an audio-assisted camera array. In *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*, pages 402–406. IEEE.
- Mohino, I., Gil-Pita, R., and Álvarez, L. (2011). Stress detection through emotional speech analysis. *Advances in Computer Science*, pages 233–237.
- Nguyen, P., Ravindranatha, M., Nguyen, A., Han, R., and Vu, T. (2016). Investigating cost-effective rf-based detection of drones. In *Proceedings of the 2nd Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use*, pages 17–22. ACM.
- Qian, H. (2015). Counting the floating point operations (flops), matlab central file exchange, no. 50608, ver. 1.0, retrieved june 30, 2015.
- Van Trees, H. L. (2004). *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons.

## Chapter 4

# Article 3: Analysis of the performance of Evolved Frequency Log-Energy Coefficients in Hearing Aids for different Cost Constraints and Scenarios

### AUTHORS

Joaquín García-Gómez, Inma Mohino-Herranz, César Clares-Crespo, Alfredo Fernández Toloba, Roberto Gil-Pita

### LIBRARY

AES E-Library. 145th AES Convention  
D.O.I.: <https://doi.org/10.17743/aesconv.2018.978-1-942220-25-1>

### CONTRIBUTION TO THE SCOPE OF THE THESIS

This article is the first publication of this thesis where the issue of Voice Activity Detection in Hearing Aids (VADHA) is researched. Hearing loss is a common problem in elderly people. Nowadays, hearing aids compensate these losses and make their life better, but they present some important issues (reduced battery life, requirement of real-time processing). Because of that, the algorithms implemented in these devices must work at low clock rates. Voice Activity Detection (VAD) is one of the main algorithms used in hearing aids, since it is useful for reducing environmental noise and enhancing speech intelligibility. In this paper, a VAD algorithm is tested using the QUT-NOISE-TIMIT Corpus, with different computational cost constraints and at different locations, with the objective of reducing the error rate of the system. Results show that 100 KIPS are enough to obtain low error rates, in line with other systems proposed in the literature. With such low values of instructions, the latency is close to zero, so the system does not introduce any significant delay. Besides, it has been demonstrated that the results are quite dependent on the scenario studied.



# Audio Engineering Society Convention Paper 10111

Presented at the 145<sup>th</sup> Convention  
2018 October 17 – 20, New York, NY, USA

*This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

## Analysis of the performance of Evolved Frequency Log-Energy Coefficients in Hearing Aids for different Cost Constraints and Scenarios

Joaquín García-Gómez, Inma Mohíno-Herranz, César Clares-Crespo, Alfredo Fernández-Toloba, and Roberto Gil-Pita

*Department of Signal Theory and Communications, University of Alcalá, Alcalá de Henares, Spain*

Correspondence should be addressed to Joaquín García-Gómez ([joaquin.garciagomez@uah.es](mailto:joaquin.garciagomez@uah.es))

### ABSTRACT

Hearing loss is a common problem in old people. Nowadays hearing aids compensate these losses and make their life better, but they present some important issues (reduced battery life, requirement of real-time processing). Because of that, the algorithms implemented in these devices must work at low clock rates. Voice Activity Detection (VAD) is one of the main algorithms used in hearing aids, since it is useful for reducing the environmental noise and enhancing the speech intelligibility. In this paper a VAD algorithm will be tested using QUT-NOISE-TIMIT Corpus, with different computational cost constraints and at different locations.

### 1 Introduction

Hearing loss is common in older adults [1] and is associated with social isolation and depression [2, 3]. There is evidence that hearing aids can improve the quality of life and increase the social engagement of people who suffer from these problems [4], and there are indications that this kind of devices may have a positive impact on the cognitive system after a certain time using them [5].

However, one of the problems that hearing aids involve is the lack of intelligibility when using them in a noisy environment. Nowadays most of these devices include techniques to improve the hearing experience, such as feedback cancellation, environment classification and speech enhancement [6, 7, 8]. In order to implement these mechanisms it is necessary to detect when conversations are taking place and to distinguish them from

the environmental noise. These algorithms are included in the field of Voice Activity Detection (VAD), which is the main objective of this study and is defined as the detection of presence or absence of human speech [9].

In the literature there are a lot of methods and features capable of detecting sound sources, such as Mel Frequency Cepstral Coefficients (MFCCs). However, hearing aids present some issues: the reduced battery life, the small size of the device and the need for real-time processing [10]. These requirements limit the computational capability of the device, the number of assembled components and the processing delay. Besides, all the computational capabilities are not available for developing the VAD system, since a considerable part of the DSP resources has to be used to run the algorithms that allow to compensate the hearing losses. Thus, the typical implementation of the MFCCs in hearing aids is not feasible.

Some studies of VAD in hearing aids employed features in the time and frequency domains (maximum of the autocorrelation of the LPCs, spectral slope, reflection coefficients and input signal power) [11], but using sequences with low background noise. Other studies have solved the VAD problem, obtaining good results in terms of error rate [12, 13, 14]. However, they have not restricted the computational cost in their algorithms, so it is doubtful that they could be implemented in hearing aids.

With regard to all these considerations, the aim of the paper is to develop a VAD algorithm for hearing aids capable of being used in multiple scenarios and restricted to certain constraints in terms of computational cost. With this purpose, Evolved Frequency Log-Energy Coefficients (EFLECs) will be implemented [7]. They are a set of computationally limited parameters inspired in the Mel Frequency Cepstral Coefficients. EFLECs were used for VAD in hearing aids [10], but it was a first approximation where the performance of the algorithm according to the computational cost was not studied, it was not tested in a standard database and the system was not compared with other methods from the literature.

The paper is structured as follows. Section 2 will describe the VAD system implemented, including all their stages and the computational cost associated to them. In Section 3, the developed experiments will be detailed, including the database used, the parameters studied and the results obtained. Finally, in Section 4 the conclusions will be presented.

## 2 Methods

In this section the VAD system will be described. It is composed of several stages, including the measurement extraction stage, the feature extraction stage and the detection stage.

In the measurement extraction stage the system has to obtain a set of values which contains useful information for solving the problem at hand. Other studies are based on features like spectral centroid, spectral flux, voice to white or short time energy [15]. However, most of these works have developed systems that require a great amount of resources. It is an issue when trying to implement them in hearing aids, such as Mel Frequency Cepstral Coefficients (MFCCs) based ones [16]. With this in mind, a low cost version of

the MFCCs has been developed in recent years, making feasible the implementation in hearing aids. They are known in the literature as Evolved Frequency Log-Energy Coefficients (EFLECs) [7]. The objective of these coefficients is to reach a compromise between the error classification probability and the number of instructions per second required by the system.

First of all, the computational cost required by an EFLEC-based VAD system will be analyzed. If the system uses  $M$  measurements, the number of instructions per second required by the system is obtained using equation (1) [7].

$$C_T = \frac{2F_s M}{NT} C_S + \frac{2F_s}{NT} C_D + \frac{2F_s}{N} \sum_{m=1}^M C_M(m), \quad (1)$$

where  $C_M(m)$  is the computational cost required when evaluating the  $m$ -th measurement,  $C_S$  is the computational cost required when evaluating the statistics,  $C_D$  is the computational cost required by the detector,  $F_s$  is the sampling frequency,  $T$  is the number of time frames between decisions, and  $N$  is the frame size. In this equation an overlapping factor of 50% has been considered.

The objective of the next stage is to calculate statistics from the  $M$  measurements. The mean, the square of the mean, the standard deviation and the variance have been successfully used in sound environment issues for hearing aids [10]. Using them, the classifiers can find solutions based on quadratic combinations of the features, so the performance is expected to be improved using a few more resources than using linear combinations. With this in mind, and knowing that the variance can be determined as a linear combination of the square of the mean and the mean of the squared values, we will use the set  $S_5$  from [7], which is composed of the sum of the values, the sum of the squared values, the square of the sum values, and the standard deviation. The computational complexity associated to it is  $C_S = 39$  instructions per measurement and  $N_S = 4$  features per measurement.

Concerning the last stage, the detection has been planned using two classifiers: the Least Squares Linear Detector (LSLD) and the Multilayer Perceptron (MLP), which are feedforward artificial neural networks, as in [10]. The number of instructions associated to them is represented in equations (2) and (3), respectively.

$$C_D|_{LC} = 12 + MN_S, \quad (2)$$

where  $M \cdot N_S$  is the total number of features.

$$C_D|_{MLP} = 12 + K + K(12 + (M+2)N_S), \quad (3)$$

where  $K$  represents the number of available neurons in the network.

The VAD algorithm used in the experiments is based on the EFLECs, a set of novel designed features inspired in the MFCCs. The traditional MFCCs have some disadvantages in terms of computational cost, like the large number of instructions that the evaluation of the Mel scale triangular filters and the DCT require. For solving this problem, soft computing techniques allow to optimize the design of the feature stage, extracting tailored features capable of being used in hearing aids. EFLECs might be considered as a deep learning technique, where feature parameters are determined along the training process.

EFLECs introduce some modifications to the computation of the MFCCs: the triangular filters are replaced by uniform filters, and the DCT block is removed. The performance of the system does not change in terms of error rate, but the computational cost is significantly reduced [17]. Besides, EFLECs apply an evolutive algorithm for selecting the frequency bands instead of using the Mel scale. This way the DSP load is controlled [7].

The computational cost of the VAD system is represented in equations (4) and (5) for the case of using an LSLD and an MLP, respectively [10].

$$C_T^{LSLD} = \frac{F_s}{N} \left( \frac{24}{T} + \sum_{m=1}^M 8L(m) + 24 + \frac{86}{T} \right), \quad (4)$$

where  $L(m)$  is the number of non zero coefficients of each uniform filter.

$$C_T^{MLP} = \frac{F_s}{N} \left( \frac{24}{T} + \frac{42K}{T} + \sum_{m=1}^M 8L(m) + 24 + \frac{78}{T} + \frac{8K}{T} \right) \quad (5)$$

As was stated above, most of the computational cost of the system is related to the frequency bands selected

by the system. Because of that, an optimization process based on evolutionary algorithms is required for constraining the number of instructions per second. A tailored evolutionary algorithm has been implemented in order to search for the set of frequency bands which minimizes the error rate of an LSLD, but limiting the number of instructions per second at the same time. Once this algorithm selects the best subset of bands, the MLP is trained. The same process as in [10] will be used.

### 3 Analysis of the performance in hearing aids

In this section the database used will be described, as well as the relation between the parameters studied, the relation between the error rate and the scenarios considered, and a comparison with other methods.

#### 3.1 QUT-NOISE-TIMIT Corpus

In order to present the results obtained in the developed experiments, the evaluated database will be described. There exist some datasets available for solving the VAD problem. In our case, QUT-NOISE-TIMIT Corpus has been chosen because it is one of the most relevant dataset in the field of study, it includes a large variety of environments and it is quite long [18]. It consists of 600 hours of noisy speech sequences, which are obtained mixing some background noises and speech events chosen from the TIMIT clean speech corpus.

The scenarios considered in the database are: cafe, home, street, car and reverb one. Specifically, each of them was recorded in two different locations:

- Cafe: an outdoor cafe and an indoor shopping centre food-court.
- Home: a kitchen and a living-room.
- Street: a roadside near inner-city and outer-city traffic-light controlled intersections.
- Car: a car with windows down and a car with windows up.
- Reverb: an indoor pool and an enclosed carpark.



The mixing of the background noise audios and the clean speech is made randomly, so that the segment of audio selected from the noisy recorders is not fixed by the user. The speech events are selected and included in a similar way. Besides, effects of co-talking are included in the database, since each speech event is randomly combined with the previous one with a probability of 50%, taking into account the overlap and silence between the speakers.

The full database contains 24,000 speech sequences. The sampling frequency  $F_s$  is 16 kHz. In this study a subset from the full dataset has been considered for the experiments, since it was not feasible to execute all the networks considered on such a large set of data. Specifically, 5 noise locations (cafe-foodcourtb, home-kitchen, street-city, car-windowb and reverb-pool), 2 SNRs considered as medium noise in the documentation (0 dB and 5 dB) and 6 speech sequences per location have been included (2 sequences have less than 25% of speech, 2 sequences have between 25% and 75% of speech and 2 sequences have more than 75% of speech). This way the data subset is balanced, so that there exists an average value of 50% of speech against the total duration of the audios. The 2 sessions and the 2 durations from the total set have been considered, resulting in a total of 120 sequences and 3 hours of audio.

In all the experiments detailed below a 5-fold cross validation will be implemented, so that in each iteration one of the locations is used as training subset and the rest of locations are included in the test subset. The features were computed with  $N = 128$  DFT points and window lengths of 256 samples. The decomposition is performed with 65 frequency bands and the time slot for the decision is 16 milliseconds ( $T = 4$ ), being all these values standard in typical algorithms for hearing aids.

### 3.2 Results

First of all, the relation between some parameters of the simulations will be analyzed. Specifically, we will study the cost constraint, the number of frequency bands selected and the number of neurons used in the Neural Network. The cost constraint is measured in Kilo Instructions Per Second (KIPS) and can take 19 possible values,  $C = \{10, 20, \dots, 100; 120, 140, \dots, 200; 250, 300, \dots, 400\}$  KIPS, so that the cost steps are smaller for low cost values and bigger for high values. These

values have been considered because in hearing aids the computational power of the DSP usually does not exceed 5000 KIPS, and the power consume is proportional to the DSP clock frequency. In this way, considering a DSP which consumes 1 mW/MHz and  $C = 400$  KIPS (the worst case), we obtain a consume of 0.4 mW for the VAD algorithm. This value corresponds to approximately 40% of the power consume of a typical compression/expansion algorithm of a hearing aid [7].

The number of frequency bands selected, that is to say, the number of features selected can take 30 values,  $M = \{1, 2, \dots, 30\}$ , while the number of neurons can take 9 values,  $K = \{0, 1, 2, 3, 4, 5, 10, 15, 20\}$ .

In figure 1 it can be shown three color maps, where red tones represent high error rates and blue tones represent better results (low error rates). The map in the top is related with  $C = 100$  KIPS, the map in the middle is related with  $C = 200$  KIPS and the map in the bottom is related with  $C = 400$  KIPS. White color appears when this experiment can not be implemented because the cost associated to that number of frequency bands  $M$  and neurons  $K$  exceeds the cost constraint  $C$ .

In the view of the results it can be observed that the error rate obtained is lower when the cost constraint is less restrictive. Thus, the values in the last color map are better than in the previous ones. Likewise, the amount of experiments that can not be carried out (white cells) is significantly reduced, as a greater amount of resources in terms of computational cost are available.

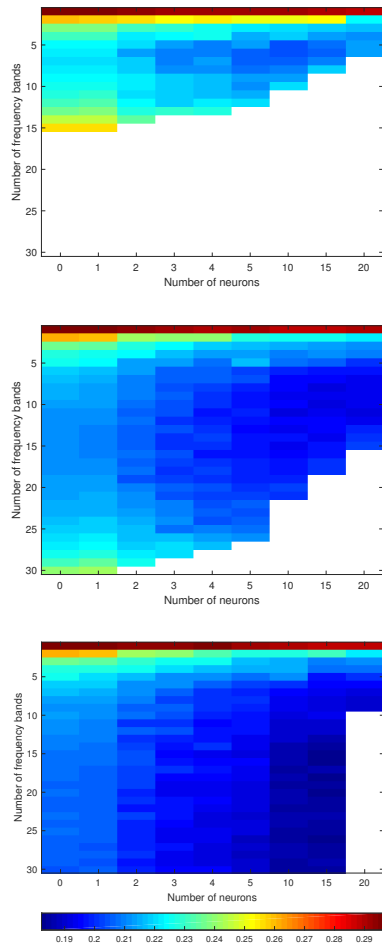
Furthermore, there exists a tendency in the behavior of the algorithm related to the number of frequency bands  $M$  and the number of neurons  $K$ . The best results are obtained in the three cost constraints  $C$  when choosing a high  $K$ , but a moderate  $M$  among all the bands available. This is due to the fact that greater amount of features does not always mean a better result. However, a larger cost usually means a better result, since the algorithm can find more expensive and useful features for the problem at hand.

Now we will study which values of  $M$  and  $K$  are related to the best result in terms of error rate, for the different  $C$  considered. Table 1 shows the mentioned values.

As would be expected, when the system is less restrictive (higher  $C$ ), the values of  $M$  and  $K$  are larger. However, we can conclude that the algorithm prefers a moderate number of frequency bands  $M$ . In fact, although

**Table 1:** Values of the parameters which optimize the error rate for the different  $C$ .

$C$ (KIPS)	$M$	$K$
10	1	0
20	2	1
30	3	2
40	3	5
50	4	5
60	5	5
70	5	5
80	6	10
90	5	10
100	6	10
120	6	10
140	6	15
160	7	20
180	12	10
200	11	10
250	17	10
300	17	15
350	16	15
400	15	15

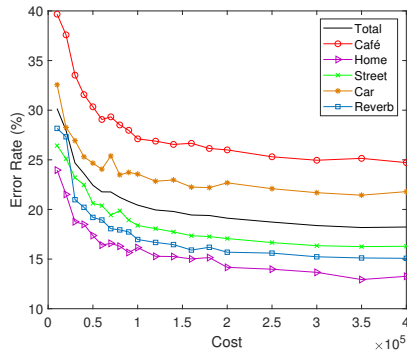
**Fig. 1:** Error rate obtained for 100 KIPS (top), 200 KIPS (middle) and 400 KIPS (bottom), according to the number of frequency bands and neurons selected.

$M$  can take until 30 bands, 17 is the maximum selected by the algorithm (250-300 KIPS).

As was mentioned above, different scenarios compose the database at hand. To continue with the analysis of the experiments, we are going to study how well the system works depending on the environment. Figure 2 represents the error rate depending on the cost constraint  $C$ , for the scenarios considered.

As it can be observed, the home scenario is the one which works better in terms of error rate, reaching less than 13% of error with high values of  $C$ . This is due to the fact that it is the scenario where there exists less background noise. For its part, the cafe scenario is the worst, getting an error close to 25% in the less restrictive case. It happens because the cafe is an environment where most people is having a conversation, so the dialogue between the main speakers is easily confused with the rest of conversations which are taking place.

Now we will observe the general tendency of the error curve according to  $C$ . The most significant improvement in terms of error rate occurs between 10 KIPS and 50 KIPS, when the error falls from 30.2% to 22.4%. In



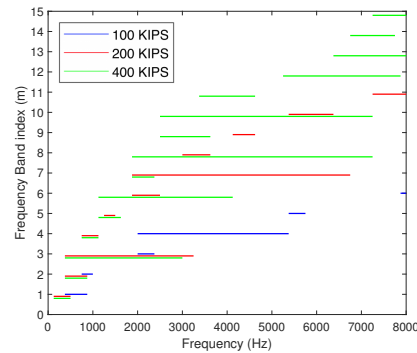
**Fig. 2:** Error rate depending on the cost constraint.

the next segment of the curve (between 50 KIPS and 100 KIPS), the error rate decreases slowly, from 22.4% to 20.4%. From here, the error remains almost constant, decreasing until 19.1% when duplicating the  $C$  (200 KIPS) and until 18.2% when quadrupling the  $C$  (400 KIPS).

It therefore seems logical that a compromise between  $C$  and performance should be reached. The final application will run in a low-cost-microprocessor capable of being executed in hearing aids, so we can establish a moderate  $C$  such as 100 KIPS. This could only mean 2-3 points in terms of error, but working at a frequency 4 times lower than before.

In figure 3 the frequency bands which provide the low error to the system are shown. Blue lines represent the case when  $C=100$  KIPS (6 frequency bands), red lines represent a restriction of  $C=200$  KIPS (11 frequency bands) and green lines are related to a  $C=400$  KIPS (15 frequency bands). The selected bands are very different from those selected in the Mel scale of the traditional MFCCs.

Now we will try to compare the results obtained in this study with the ones obtained with other methods. With this purpose, we will compare results considering medium noise, since we have included audios with SNR=0 and 5 dB. It should be noted that in the case at hand a subset of the database has been employed for training and testing the system, so it will be an approximate comparison. Nevertheless, the results should be similar, since some experiments were developed over



**Fig. 3:** Frequency bands selected for different  $C$ : 100 KIPS (blue), 200 KIPS (red) and 400 KIPS (green).

the scenarios that are out of our subset, getting a similar performance in terms of error rate.

The comparison has been made with the baseline systems of the literature. Specifically, we will compare with the studies of Sohn et al. [12], Ramírez et al. [13] and Wisdom et al. [14], which applies to QUT-NOISE-TIMIT Corpus a single-channel-based voice activity detector [19] and a two-channel-based system [20]. The results are summarized in table 2, where HTER represents the Half-Total Error Rate. This rate is calculated as the average of false alarm rate and miss rate, which at the end corresponds to the error rate calculated in this study.

**Table 2:** Comparative results with VAD baseline systems.

Method	HTER (%)
DSB + Sohn et al.	24.58
DSB + Ramírez et al.	19.87
CS-LDA (2ch) + Wisdom et al.	19.96
SDOI (1 ch) + Wisdom et al.	15.21
EFLEC (100 KIPS)	20.44
EFLEC (400 KIPS)	18.23

In view of the results, the system seems to have a performance in line with the rest of systems proposed in

the literature. It should be noted that the rest of methods do not provide a study in terms of computational cost, so they are not restricted in this respect.

#### 4 Summary

Hearing aids are devices which require algorithms constrained in the number of instructions per second due to the digital signal processor in which are implemented and the small computational resources available. The aim of this study has been to propose an optimized implementation of EFLEC for VAD in hearing aids. With this purpose, algorithms have been optimized to keep a balance between the error probability obtained and the number of instructions used. With lo

Using QUT-NOISE-TIMIT Corpus the results have shown that 100 KIPS are enough to obtain low error rates. This value supposes around a 10% of the power consume respect to a typical algorithm in a hearing aid. With such low values of instructions the latency is close to zero, so the system does not introduce any significant delay.

Besides, it has been demonstrated that the results are quite dependent on the scenario studied. Related to the error rate, the results are in line with other systems proposed in the literature, even when the comparison is not entirely accurate because a subset of the database has been used.

#### 5 Acknowledgment

This work has been funded by the Spanish Ministry of Economy and Competitiveness-FEDER under Project TEC2015-67387-C4-4-R, and by the University of Alcalá under Project CCGP2017-EXP/060.

#### References

- [1] Davis, A. C., "The prevalence of hearing impairment and reported hearing disability among adults in Great Britain," *International Journal of Epidemiology*, 18(4), pp. 911–917, 1989.
- [2] Gates, G. A. and Mills, J. H., "Presbycusis," *The Lancet*, 366(9491), pp. 1111–1120, 2005.
- [3] Strawbridge, W. J., Wallhagen, M. I., Shema, S. J., and Kaplan, G. A., "Negative consequences of hearing impairment in old age: a longitudinal analysis," *The Gerontologist*, 40(3), pp. 320–326, 2000.
- [4] Mulrow, C. D., Aguilar, C., Endicott, J. E., Tuley, M. R., Velez, R., Charlip, W. S., Rhodes, M. C., Hill, J. A., and DeNino, L. A., "Quality-of-life changes and hearing impairment: a randomized trial," *Annals of Internal Medicine*, 113(3), pp. 188–194, 1990.
- [5] Kalluri, S. and Humes, L. E., "Hearing technology and cognition," *American journal of audiology*, 21(2), pp. 338–343, 2012.
- [6] Lee, C.-H., Kates, J. M., Rao, B. D., and Garudadri, H., "Speech quality and stable gain trade-offs in adaptive feedback cancellation for hearing aids," *The Journal of the Acoustical Society of America*, 142(4), pp. EL388–EL394, 2017.
- [7] Gil-Pita, R., Ayllón, D., Ranilla, J., Llerena-Aguilar, C., and Díaz, I., "A computationally efficient sound environment classifier for hearing aids," *IEEE Transactions on Biomedical Engineering*, 62(10), pp. 2358–2368, 2015.
- [8] Gong, Z. and Xia, Y., "Two speech enhancement-based hearing aid systems and comparative study," in *Information Science and Technology (ICIST), 2015 5th International Conference on*, pp. 530–534, IEEE, 2015.
- [9] Ramirez, J., Górriz, J. M., and Segura, J. C., "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust speech recognition and understanding*, InTech, 2007.
- [10] Gil-Pita, R., García-Gomez, J., Bautista-Durán, M., Combarro, E., and Cocana-Fernandez, A., "Evolved frequency log-energy coefficients for voice activity detection in hearing aids," in *Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on*, pp. 1–6, IEEE, 2017.
- [11] Itoh, K. and Mizushima, M., "Environmental noise reduction based on speech/non-speech identification for hearing aids," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 1, pp. 419–422, IEEE, 1997.
- [12] Sohn, J., Kim, N. S., and Sung, W., "A statistical model-based voice activity detection," *IEEE signal processing letters*, 6(1), pp. 1–3, 1999.

- [13] Ramirez, J., Segura, J. C., Benitez, C., De La Torre, A., and Rubio, A., "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, 42(3-4), pp. 271–287, 2004.
- [14] Wisdom, S., Okopal, G., Atlas, L., and Pitton, J., "Voice activity detection using subband noncircularity," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4505–4509, IEEE, 2015.
- [15] Gil-Pita, R., Alexandre, E., Cuadra, L., Vicen, R., and Rosa-Zurera, M., "Analysis of the effects of finite precision in neural network-based sound classifiers for digital hearing aids," *EURASIP Journal on Advances in Signal Processing*, 2009(1), p. 456945, 2009.
- [16] Xiang, J., McKinney, M. F., Fitz, K., and Zhang, T., "Evaluation of sound classification algorithms for hearing aid applications," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 185–188, IEEE, 2010.
- [17] Gil-Pita, R., López-Garrido, B., and Rosa-Zurera, M., "Tailored MFCCs for sound environment classification in hearing aids," in *Advanced Computer and Communication Engineering Technology*, pp. 1037–1048, Springer, 2015.
- [18] Dean, D. B., Sridharan, S., Vogt, R. J., and Mason, M. W., "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," *Proceedings of Interspeech 2010*, 2010.
- [19] Rubio, J. E., Ishizuka, K., Sawada, H., Araki, S., Nakatani, T., and Fujimoto, M., "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pp. IV–385, IEEE, 2007.
- [20] Kim, H.-D., Komatani, K., Ogata, T., and Okuno, H. G., "Two-channel-based voice activity detection for humanoid robots in noisy home environments," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pp. 3495–3501, IEEE, 2008.

## Chapter 5

# Article 4: Linear detector and neural networks in cascade for voice activity detection in hearing aids

### AUTHORS

Joaquín García-Gómez, Roberto Gil-Pita, Miguel Aguilar-Ortega, Manuel Utrilla-Manso, Manuel Rosa-Zurera, Inma Mohino-Herranz

### JOURNAL

Applied Acoustics

D.O.I.: <https://doi.org/10.1016/j.apacoust.2020.107832>

### RANKING

JCR (2019): 2.44

Quartile Rank - Computer Science, Artificial Intelligence: 9/32 (Q2)

Quartile Rank - Acoustics: 49/105 (Q2)

### CONTRIBUTION TO THE SCOPE OF THE THESIS

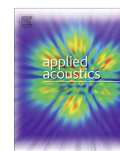
This article is the second publication of the thesis where the issue of Voice Activity Detection in Hearing Aids (VADHA) is researched. This work is a natural extension to the previous article, since additional optimization methods based on cascade-detectors are applied to reduce the computational cost while maintaining the same performance of the system or to increase the performance while maintaining the same computational cost. This is achieved by a two-stage detector. In the first stage, a linear system determines whether the detection can be easily carried out, or a second stage with a more complex neural-network-based detection is required. This way, some of the decisions are taken without using the complex detector. The results show the usefulness of this new configuration, as the system error is reduced up to 8.5% while using the same amount of resources. Moreover, it seems that extra optimizations can provide improvements in

terms of performance, especially when the environment is not very noisy. A comparison with other methods from the literature has shown that the proposals of this study get the best solutions when comparing with similar short-window algorithms (8-10 ms). Other algorithms get better results when applying window lengths higher than 20 ms to VAD, but this is not practical for hearing aid applications.



Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: [www.elsevier.com/locate/apacoust](http://www.elsevier.com/locate/apacoust)

## Linear detector and neural networks in cascade for voice activity detection in hearing aids



Joaquín García-Gómez\*, Roberto Gil-Pita, Miguel Aguilar-Ortega, Manuel Utrilla-Manso, Manuel Rosa-Zurera, Inma Mohino-Herranz

Department of Signal Theory and Communications, University of Alcalá, Ctra Madrid-Barcelona, km. 33, 600, 28805 Alcalá de Henares, Madrid, Spain

### ARTICLE INFO

#### Article history:

Received 6 March 2020

Received in revised form 5 November 2020

Accepted 24 November 2020

#### Keywords:

Voice activity detection

Hearing aids

Cascade-detectors

Computational cost constraints

Artificial neural networks

Linear detector

### ABSTRACT

Hearing loss is a common issue when people become older, resulting in problems such as depression, risk of dementia, and cognitive decline, among others. Hearing aids are computationally constrained devices that offer the possibility of solving this issue, thus improving people's quality of life. A typical algorithm that should be implemented in these devices is Voice Activity Detection. In this work, cascade detectors are applied to reduce the computational cost while maintaining the same performance or to increase the performance while maintaining the same computational cost. This is achieved by a two-stage detector. In the first stage, a linear system determines whether the detection can be easily carried out, or a second stage with a more complex neural-network-based detection is required. This way, some of the decisions are taken without using the complex detector. The results show that the system error can be reduced up to 8.5% while using the same amount of resources. Moreover, the error is the lowest among the proposals that are affordably implemented in hearing aids.

© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

### 1. Introduction

Older adults usually suffer from hearing impairment [1]. World Health Organization claims that approximately one-third of people over 65 years old have some degree of hearing loss [2], considering it as a loss greater than 40 decibels (dB) in the ear with better hearing. It makes hearing loss the third most prevalent chronic health condition affecting elderly adults. It is a problem that affects over 5% of the world population, including both adults (432 million) and children (34 million). Furthermore, it is estimated that by 2050 one-tenth of the population will suffer from disabling hearing loss.

Hearing loss is an issue that has been underdiagnosed and undertreated over time [3]. However, its health implications are very severe, including social isolation, depression, altered physical function, reduced activity participation, lower quality of life, falls, greater cognitive decline, and higher risk of dementia [4]. In this sense, hearing aids constitute a valuable tool that has demonstrated to have a positive impact on long-term cognition, as hearing aids users have shown a cognitive decline similar to elders with no hearing loss [5].

Implementing sound signal processing algorithms in hearing aids is not an easy task. Due to the real-time and consumption requirements, the implemented algorithms present several constraints that must be considered in the design process [6]. First, due to the battery life requirements, there are constraints in the computational capability of the device and the number of assembled components. Roughly speaking, the computational power of a hearing aid rarely exceeds 5 million instructions per second (MIPS), thus limiting the complexity of the algorithms that can be implemented in them. By reducing the computational power, energy consumption is reduced, and consequently battery life is increased. Second, hearing aids require low-delay real-time processing algorithms. In numerical terms, the total delay introduced by the hearing aid cannot exceed 20 ms [7]. This fact limits the length of the time frame in the time-frequency analysis, therefore limiting the frequency resolution.

In the field of hearing aids, the detection of human speech, also known as Voice Activity Detection (VAD), is essential [8]. One of the problems of hearing aids is the lack of intelligibility in noisy environments. Thanks to VAD, it is possible to differentiate between conversations and noise, and in this way the hearing experience can be improved through techniques such as feedback cancellation, environment classification, and speech enhancement [9–11]. VAD has been extensively researched in the past. In [12], the authors employed a decision-directed parameter estimation

\* Corresponding author.

E-mail addresses: [joaquin.garciagomez@uah.es](mailto:joaquin.garciagomez@uah.es) (J. García-Gómez), [roberto.gil@uah.es](mailto:roberto.gil@uah.es) (R. Gil-Pita).

<https://doi.org/10.1016/j.apacoust.2020.107832>

0003-682X/© 2020 The Author(s). Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



method for the likelihood ratio test along with a Hidden Markov Model (HMM). Another proposal [13] measured the Long-Term Spectral Estimation (LTSE) between speech and noise and compared the envelope to the average noise spectrum for detecting voice activity. The second-order non-circularity of speech and noise complex subbands is used in [14]. In general, the problem is that these proposals have not focused on hearing aid requirements, and therefore the computational resources have not been adapted to them. In a recent analysis [15], the authors compared a large number of proposals and concluded that a long temporal context and a look-ahead are beneficial for speech detection, but they require much more CPU consumption than the available for off-the-shelf hearing aids [6]. This makes infeasible the implementation of these algorithms using the available resources in digital hearing aids. In recent proposals, complex and unfeasible algorithms for hearing aid applications have been presented: in [16], Line Spectral Frequency (LSF) based features and extreme learning-based classifiers were used; in [17], a real-time VAD based on convolutional neural networks was implemented in a smartphone app; and in [18], the authors presented a model based on deep neural networks and an adaptive context attention model.

The problem of VAD in hearing aids has already been studied in the literature too. In [11], a VAD system based on neural networks was implemented using a pattern recognition scheme, including the typical stages of measurement extraction, feature extraction, and neural network-based detection. The measurements considered were the Evolved Frequency Log-Energy Coefficients (EFLECs), a less costly alternative to the traditional Mel-Frequency Cepstral Coefficients (MFCCs), which makes their calculation feasible for hearing aids [19]. These coefficients provide a trade-off between error classification probability and the number of instructions per second. Unfortunately, considering that the VAD systems must take a decision every time frame, the computational cost that these algorithms require to implement an efficient VAD is still relevant, thus affecting the battery life of the devices. Less computationally intensive VAD algorithms can be implemented using fewer EFLECs and simpler detectors, but the performance is consequently reduced.

In this article, a novel solution is proposed for implementing VAD in computationally constrained hearing aids, aiming at combining the benefits of a less computationally intensive VAD algorithm with a more complex detector. To this end, we propose the use of cascade-detectors, implementing a simple detector in a first stage (linear one) followed by a second stage based on neural networks, which must be used when the decision of the system is not clear. The main goal is to reduce the computational cost of the system while maintaining the performance of a complex VAD. We will see that it is possible to achieve the set objective so that the error rate of the system can be reduced to a certain point by applying classifiers with cascade configuration. We will compare the results with other proposals that used the same dataset (QUT-NOISE-TIMIT Corpus), concluding that our proposal outperforms the algorithms with similar computational constraints.

The paper is structured as follows. In Section 2, we will review the cost associated with EFLEC-based VAD systems in hearing aids, including the cost related to the feature extraction and the detector itself. Later, in 3, we introduce the concept of cascade-detectors and how they can be applied to VAD. Then, the optimizations applied to adjust the weights of the first part of the system will be detailed in Section 4. In Section 5, we present the experiments carried out, the obtained results, and a comparison with other proposals from the literature. Finally, we summarize the article and its results in Section 6.

## 2. Voice activity detection in hearing aids

As stated above, the objective of this paper is to combine the benefits of simple and complex classifiers for improving the relationship between performance and computational cost. For this purpose, this section is intended to briefly review standard EFLEC-based VAD systems [19].

From a machine learning perspective, VAD systems are typically composed of a feature extraction stage and a detection stage. The objective of the feature extraction is to obtain a set of  $S$  statistics from  $M$  measurements that contain useful information to distinguish speech from noise. Many features can be used to solve the VAD problem, with MFCCs being some of the most used in literature. Unfortunately, despite the good performance that can be achieved with their use, they require a significant amount of the computational resources available in current hearing aids [20]. In this regard, EFLECs have been shown to achieve equivalent performance, but with less computational complexity [19], and for that reason they will be used in the VAD system proposed in the paper.

The modifications that EFLECs introduce in comparison with MFCCs are:

- They use uniform filters instead of triangular ones, and the Discrete Cosine Transform (DCT) block is removed. These modifications highly reduce the computational cost without giving rise to an increase in error rate.
- An evolutive algorithm is implemented for selecting the limits of the frequency bands where the filters are distributed instead of using the Mel scale. This modification allows us to obtain lower error rates and, at the same time, to control the Digital Signal Processor (DSP) load [19].

Regarding the detection stage, two different detectors are considered: the Least Squares Linear Detector (LSLD) [21] and the Multilayer Perceptron (MLP) [22].

The LSLD is a detector that works properly with a very low computational complexity once it is trained. Considering a vector of  $L = S \cdot M$  features  $x = [x_1, \dots, x_L]^T$ , the detection rule can be obtained by thresholding  $y$ , a linear combination of the features:

$$y = v_0 + \sum_{i=1}^L v_i x_i \underset{H_0}{\overset{H_1}{\geq}} 0, \quad (1)$$

where  $v_i$  are the weights of the linear combination,  $H_0$  is the decision related to the presence of noise, and  $H_1$  is the decision related to the presence of speech. In the least squares approach, the weights are adjusted in order to minimize the mean square error over the design set (a set with  $P$  design vectors, each one of them corresponding to a different time frame), and this minimization leads to the Wiener-Hopf equations [23].

$$\mathbf{v} = \mathbf{t} \cdot \mathbf{Q}^T \cdot (\mathbf{Q} \cdot \mathbf{Q}^T)^{-1}, \quad (2)$$

where  $\mathbf{v}$  is a  $1 \times (L + 1)$  vector containing the weights of the linear combination  $v_i$  (including the bias),  $\mathbf{t}$  is a  $1 \times P$  vector containing the target values (+1 and -1) of the  $P$  design patterns and  $\mathbf{Q}$  is a  $(L + 1) \times P$  matrix containing a row of ones for the bias and the  $L$  features of the  $P$  design patterns.

On the other hand, MLPs are feedforward artificial neural network models that have successfully been implemented in hearing aids [11,24]. MLPs are typically designed to minimize the mean squared error at the output using backpropagation algorithms [25]. In this paper, two-layer MLPs have been trained using the Levenberg-Marquardt optimization algorithm [26]. The design

data has been randomly divided into two subsets, one for training with 80% of the data and the other with the remaining 20% for monitoring and early stopping the training process. The inputs of the neurons of the first layer are the  $L$  available features, and the second layer combines the outputs of these neurons to take the decision by thresholding. The complexity of the solution implemented by an MLP is controlled by the number of neurons in the first layer.

In general, MLPs outperform LSLD for VAD in hearing aids, and thus they have been selected as baseline detectors for the experiments carried out in the paper.

The computational cost required by a system which uses EFLECS for VAD will be analyzed in the following lines. Considering that the system uses  $M$  measurements, the number of instructions per second required to determine these EFLEC features is obtained using Eq. (3).

$$C_{VAD} = \frac{2F_s M}{NT} C_S + \frac{2F_s}{NT} C_D + \frac{2F_s}{N} \sum_{m=1}^M C_M(m), \quad (3)$$

where  $C_M(m)$  is the computational cost related to the  $m$ -th measurement,  $C_S$  is the computational cost related to the evaluation of the statistics,  $C_D$  is the computational cost related to the detector,  $F_s$  is the sampling frequency,  $T$  is the number of time frames that separate one decision from the next one, and  $N$  is the frame length. In Eq. (3), an overlapping factor of 50% is assumed.

The evaluation of each EFLEC measurement depends on two terms, as shown in Eq. (4).

$$C_M = C_F(m) + C_L, \quad (4)$$

where  $C_F(m)$  is the computational cost associated with the evaluation of the uniform filter, and  $C_L$  is the computational cost associated with the evaluation of the logarithm. The evaluation of the uniform filter requires  $C_F(m) = 3 + 4L(m)$  operations per filter, where  $L(m)$  is the number of non-zero coefficients of each filter. Related to the logarithm calculation, it uses  $C_L = 9$  instructions in a typical DSP architecture.

Once we have the measurements, the features are extracted by applying some statistics to them. The mean and the standard deviation are common statistics that have been used previously in sound issues related to hearing aids [24]. However, some classifiers present a better performance when quadratic terms are also added to the input features [27]. Thus, apart from the mean and the standard deviation, the square of the mean and the variance are considered too. The computational complexity associated with this step is  $C_S = 39$  instructions per measurement and  $S = 4$  features per measurement.

Concerning  $C_D$ , the computational cost of the detector, it is different for LSLD and MLP. The former requires the number of instructions expressed in Eq. (5) to be implemented, while the MLP requires the number of instructions expressed in Eq. (6), being  $K$  the number of available neurons in the hidden layer of the MLP.

$$C_{D|LC} = 12 + M \cdot S \quad (5)$$

$$C_{D|MLP} = 12 + K + K(12 + (M + 2)S) \quad (6)$$

Combining Eqs. (3), (4) and (6), and taking into account the above-mentioned values  $C_S = 39$  instructions per measurement and  $S = 4$  features per measurement, we can obtain the total computational cost of our VAD system using MLPs, according to Eq. (7).

$$C_{VAD}^{MLP} = \frac{F_s}{N} \left( \frac{24}{T} + \frac{42K}{T} + \sum_{m=1}^M \left( 8L(m) + 24 + \frac{78}{T} + \frac{8K}{T} \right) \right) \quad (7)$$

In Eq. (7), the terms that depend on the number of measurements  $M$ , that is, the number of frequency bands, have been grouped in the summation. This parameter is a key-value if we

want to reduce the computational cost of the VAD system. To allow the implementation of the system in hearing aids, the number of instructions per second must be limited through an optimization process. A tailored evolutionary algorithm has been applied to search for the set of frequency bands which minimize the error rate of an LSLD, but constraining the number of instructions per second at the same time. Once this algorithm selects the best set of bands, the MLP is trained. This process is described in [6].

In this approach, our starting point is the system obtained in a previous article [11] whose results were in line with the proposals of the literature that used the same dataset but restricted in computational resources terms. The question we want to answer in this proposal is: Is it possible to use a cascade configuration so that we maintain the performance of the best detectors while reducing the average computational cost?

### 3. Cascade-detectors for VAD in hearing aids

As stated in the introduction, the objective of the paper is to look for VAD solutions that combine the performance of the most complex classifiers and the computational cost of the simplest ones.

In many applications, a significant proportion of the training cases can be classified by a simple rule with some exceptions, as deduced in [1]. In that work, the authors proposed using a linear model that provides a solution for most cases, while the rest (exceptions) can be solved by a more complex model. It defines a system based on a multistage pattern recognition approach [28], so inputs rejected by the first stage are handled by a second stage. This method is also known as cascade classification.

For applying a cascade configuration, it is assumed that we have a set of pre-trained MLP based detectors from the previous approach [11] for each of the computational costs considered, where the number of features and the number of neurons are optimized according to that. To apply the cascade configurations, we use a two-stage system, in which the first stage uses a simple linear combination of a subset of measurements  $M'$  (being  $M' < M$ ) from the more complex detectors. In this step, we use a linear system since it involves a computational cost much lower than other alternatives.

Considering these two stages of the cascade configuration, the process of classification is as follows: first, a simple LSLD will use  $M'$  measurements to either decide or pass the decision to the second detector. If this first system is not able to take a clear and firm decision, then the more complex MLP based detector will take a more precise one in the second stage. The reason for using this structure is that it provides some advantages that can be understood in two ways. On the one hand, the computational cost is a critical factor in hearing aids, so a significant saving in these terms while maintaining the same performance could prove interesting. On the other hand, an improvement in the system performance could be obtained without the need to increase the computational requirements of the system.

In order to understand the proposed approach, we will consider the operations carried out by the linear detector showed in Eq. (8).

$$y_1 = w_0 + \sum_{i=1}^L w_i x_i, \quad (8)$$

where  $x_i$  are the used features,  $w_i$  are the weights associated with them, and  $w_0$  is a bias term. Please note here that this first system is not only used as a detector, but also used to determine whether the second classifier must be used. Therefore, once we have the output of the linear system  $y_1$ , a double threshold will be applied in order to decide if the pattern just contains noise ( $H_0$ ), if it can be considered as speech ( $H_1$ ) or if the decision is not clear and the sec-

ond detector must be used ( $H_2$ ). We can denote these thresholds as  $Q_1$  and  $Q_2$ , so that  $Q_1 < Q_2$ . These values will be related to the number of patterns that are classified in the first stage and the number of patterns that need to be classified in the second stage. So, the decision associated to the first system  $D_1$  is expressed according to (9).

$$D_1 = \begin{cases} \text{if } y_1 < Q_1 & \text{then } D_1 = H_0 \\ \text{if } y_1 > Q_2 & \text{then } D_1 = H_1 \\ \text{if } Q_1 \leq y_1 \leq Q_2 & \text{then } D_1 = H_2 \end{cases} \quad (9)$$

Now for the sake of simplicity and considering that  $y_1$  comes from a linear combination of the training features  $x_i$ , we shift the decision by subtracting  $(Q_1 + Q_2)/2$  from all terms in the inequalities, obtaining expression (10).

$$D_1 = \begin{cases} \text{if } y_1 - (Q_1 + Q_2)/2 < Q_1 - (Q_1 + Q_2)/2 & \text{then } D_1 = H_0 \\ \text{if } y_1 - (Q_1 + Q_2)/2 > Q_2 - (Q_1 + Q_2)/2 & \text{then } D_1 = H_1 \\ \text{if } Q_1 - (Q_1 + Q_2)/2 \leq y_1 - (Q_1 + Q_2)/2 \leq Q_2 - (Q_1 + Q_2)/2 & \text{then } D_1 = H_2 \end{cases} \quad (10)$$

After simplifications we get expression (11).

$$D_1 = \begin{cases} \text{if } y_1 - (Q_1 + Q_2)/2 < -(Q_2 - Q_1)/2 & \text{then } D_1 = H_0 \\ \text{if } y_1 - (Q_1 + Q_2)/2 > (Q_2 - Q_1)/2 & \text{then } D_1 = H_1 \\ \text{if } -(Q_2 - Q_1)/2 \leq y_1 - (Q_1 + Q_2)/2 \leq (Q_2 - Q_1)/2 & \text{then } D_1 = H_2 \end{cases} \quad (11)$$

Now we can identify a new variable  $y'_1 = y_1 - (Q_1 + Q_2)/2$  (we will now have a modified bias value  $w'_0 = w_0 - (Q_1 + Q_2)/2$ ) and we denote  $Q = (Q_2 - Q_1)/2$ , which will be always a positive value ( $Q > 0$ ) since we stated that  $Q_1 < Q_2$ . Considering these two new variables, expression (11) becomes into expression (12).

$$D_1 = \begin{cases} \text{if } y'_1 < -Q & \text{then } D_1 = H_0 \\ \text{if } y'_1 > Q & \text{then } D_1 = H_1 \\ \text{if } -Q \leq y'_1 \leq Q & \text{then } D_1 = H_2 \end{cases} \quad (12)$$

Thus, if the computational cost is very restricted, the value of  $Q$  will be small, so that most of the patterns will be processed by this first linear system. On the contrary, when the computational cost is less restricted, the value of  $Q$  will be higher, and many other patterns will be given to a more complex classifier (MLP based one).

After normalizing with  $Q$ , expression (13) is obtained:

$$D_1 = \begin{cases} \text{if } \frac{y'_1}{Q} < -1 & \text{then } D_1 = H_0 \\ \text{if } \frac{y'_1}{Q} > 1 & \text{then } D_1 = H_1 \\ \text{if } -1 \leq \frac{y'_1}{Q} \leq 1 & \text{then } D_1 = H_2 \end{cases} \quad (13)$$

We can denote a new variable  $y''_1$  as:

$$y''_1 = \frac{y'_1}{Q}, \quad (14)$$

so that the decision is taken using expression (15).

$$D_1 = \begin{cases} \text{if } y''_1 < -1 & \text{then } D_1 = H_0 \\ \text{if } y''_1 > 1 & \text{then } D_1 = H_1 \\ \text{if } -1 \leq y''_1 \leq 1 & \text{then } D_1 = H_2 \end{cases} \quad (15)$$

If we combine Eq. (8) and Eq. (14) we obtain the following:

$$y''_1 = \frac{w_0 - \frac{Q_1 + Q_2}{2}}{Q} + \sum_{i=1}^L \frac{w_i x_i}{Q}, \quad (16)$$

where  $y''_1$  is the normalized output of the linear detector, according to the variable  $Q$ , which scales the weights of the detector to fulfill the computational cost requirement.

At this point we can define the normalized bias  $w''_0$  and the normalized weights  $w''_i$  according to Eq. (17) and Eq. (18).

$$w''_0 = w_0/Q - (Q_1 + Q_2)/(2Q) \quad (17)$$

$$w''_i = w_i/Q \quad (18)$$

Thus,  $y''_1$  can be expressed as follows:

$$y''_1 = w''_0 + \sum_{i=1}^L w''_i x_i \quad (19)$$

In the end, we can always properly select the normalized weights  $w''_i$  of the linear combination so that the decision in the first stage is implemented using expression (15). Later we will refer to different optimizations of the system, including the optimization of these weights  $w''_i$ .

Being  $y''_1$  the normalized output of the linear system and  $y_2$  the output of the MLP-based detector trained according to [11], then the scheme followed by the system to take the decision  $D$  can be expressed using expression (20).

$$D = \begin{cases} \text{if } y''_1 < -1 & \text{then } D = H_0 \\ \text{if } y''_1 > 1 & \text{then } D = H_1 \\ \text{if } -1 \leq y''_1 \leq 1 & \text{then } \begin{cases} \text{if } y_2 < 0 & \text{then } D = H_0 \\ \text{if } y_2 > 0 & \text{then } D = H_1 \end{cases} \end{cases} \quad (20)$$

Considering that a part of the decisions will need the implementation of the second detector, the average computational cost  $\bar{C}$  is calculated according to Eq. (21).

$$\bar{C} = \alpha \cdot C_1 + (1 - \alpha) \cdot (C_1 + C_2), \quad (21)$$

where  $C_1$  and  $C_2$  are the computational costs associated with the first system (linear) and the second detector (MLP), respectively,  $\alpha$  is a value between 0 and 1 that indicates the proportion of decisions which are classified just using the first system, while  $(1 - \alpha)$  indicates the proportion of decisions which require the second detector. We will describe  $C_1$  and  $C_2$  in the following lines.

As stated in the previous section, the cost of a VAD classifier includes three terms: the cost related to the calculation of the measurements  $C_M$ , the cost related to the computation of the statistics  $C_S$  and the cost related to the detectors  $C_D$ . In the following optimization process, the terms  $C_S$  and  $C_M$  do not have to be considered, as the features have been previously computed. In the following optimization process, the terms  $C_S$  and  $C_M$  do not have to be considered, as the features have been previously computed. Applying these considerations to Eq. (3) we obtained  $C_1$  in Eq. (22).

$$C_1 = \frac{2F_s}{NT} C_D^{LIN} \quad (22)$$

Taking into account that the computational cost of the linear detector is  $C_D^{LIN} = 12 + M' \cdot S$ , where  $M'$  is the new number of measurements considered, and  $S = 4$  features per measurement, the computational cost associated to  $C_1$  is shown in Eq. (23).

$$C_1 = \frac{2F_s}{NT} (12 + 4M') \quad (23)$$

The cost  $C_2$  is equal to  $C_{VAD}^{MLP}$  from Eq. (7), so we can obtain the average computational cost of the cascade configuration  $\bar{C}$  combining previous equations:

$$\bar{C} = \frac{F_s}{N} \left[ \alpha \left( \frac{24}{T} + \frac{8M'}{T} \right) + (\alpha - 1) \left( \frac{24}{T} + \frac{42K}{T} + \sum_{m=1}^M (8L(m) + 24 + \frac{78}{T} + \frac{8K}{T}) \right) \right] \quad (24)$$

Now we will calculate which value of  $\alpha$  allows us to get an average computational cost  $\bar{C}$  equal to the total cost available  $C_T$ . With this purpose, we must equal these two values. Replacing  $\bar{C}$  by  $C_T$  in Eq. (21):

$$C_T = \alpha_f \cdot C_1 + (1 - \alpha_f) \cdot (C_1 + C_2), \quad (25)$$

where  $\alpha_f$  is a new value of  $\alpha$  that represents a frontier value. Thus, when  $\alpha$  is upper than  $\alpha_f$ , the cascade system is worthwhile and

computational resources used are lower than the available resources (i.e.,  $\bar{C} < C_T$ ). We can obtain  $\alpha_f$  from Eq. (26):

$$\alpha_f = \frac{C_1 + C_2 - C_T}{C_2} \quad (26)$$

If  $\alpha_f$  is upper than 1, it is not possible to apply the cascade-detectors because there are not enough computational resources in the system.

We will resume the proposed system following the scheme shown in Fig. 1, where the cascade configuration is shown. The system follows the steps described below:

1. The system requires an average computational cost  $\bar{C}$ , so we select a subset of features ( $M'$ ) from the full set ( $M$ ) that allows us to fulfill that requirement.
2. Once the linear detector has been applied, we evaluate if the value of the output  $y_1''$  is clear or not to take a decision, according to Eq. (20).
  - If the decision is clear, it is taken.
  - If the decision is not clear, we apply the original detector based on an MLP, which will take the decision through the output  $y_2$ .

#### 4. Determining the weights of the first linear system

As demonstrated in the previous section, the computational cost of the cascade configuration can be controlled, once we choose the MLP based detector in the second stage, but several parameters must be fitted in order to make the system work properly. In this sense, we must define a methodology to determine these parameters and the configuration, so that the global performance of the cascade system is optimized.

Two parameters need to be determined: firstly, we should determine how many measurements ( $M'$ ) and which of the available measurements ( $M$ ) should be used so that the performance is optimized. Secondly, the normalized weights of the linear combination in the first system ( $w_i''$ ) must be determined.

Concerning the selected measurements, the VAD systems explored in [11] did not use a large number of features. As a reference, the best configuration using 0.2 MIPS selected 11 EFLEC bands (that is 11 measurements). Since the number of combinations selecting subsets of 11 measurements is not very high, all the possible combinations were explored in the experiments.

The weights of the linear system ( $w_i''$ ) must be determined for each combination. Considering Eq. (18), we can see that these coefficients have a common term  $Q$ , which will be related to  $\alpha$ , the number of patterns that are classified in the first stage of the cascade system. So, taking into account that the value of  $\alpha$  can be determined from the costs of each stage and the average cost, then

for a given set of weights  $w_i''$  we can determine the associated value of  $Q$ .

This fact leaves us with the problem of determining the original weighting coefficients  $w_i$ . In a first approach, we explored the possibility of optimizing these values using a heuristic algorithm (such as a genetic algorithm). However, the obtained results were not very satisfactory since there were many convergence problems and many local minima in the optimization process. We then opted for starting the optimization from a set of approximated coefficients.

To analyze the effectiveness and the importance of this optimization in the final performance of the cascade-based detector, we explored three different optimization processes.

1. Optimization 1. One of the tasks of the linear system is to be able to separate speech from noise, that is, to implement a VAD system. So, as a first approach, we considered obtaining the coefficients  $v_i$  of an LSD system trained using a subset of features from the original set with Eq. (2), and directly assigning them to the coefficients of the linear system in the first stage of the cascade solution, that is,  $w_i = v_i, \forall i = 0, \dots, S \cdot M'$ . Regarding the thresholds, in this optimization it is assumed that  $-Q_1 = Q_2 = Q$ , that is to say,  $w_0' = w_0$ , and the value of the output of the LSD  $y_1$  is considered approximately symmetric around zero. The idea here is that if the classifier is correctly designed, the probability that a pattern with an output value close to 0 is well classified will be low, and thus these values will be more accurately classified in the second stage of the cascade system.
2. Optimization 2. It is assumed that the best linear combination for solving the tasks of the first stage is a rough one. We are neglecting that the objective of the first stage is not only to classify properly, but also to determine whether the second stage of the cascade system must be used. So, as a second approach, we studied the possibility of determining the best bias of the linear combination independently of the weights assigned to each feature. Thus, in this second case  $w_i = v_i, \forall i = 1, \dots, S \cdot M'$ . In this optimization, the equation  $-Q_1 = Q_2 = Q$  is not assumed, and therefore the value of  $w_0'$  must be estimated. This optimization implies an increase in the computational time of the design process, but since it implies the estimation of a unique parameter ( $w_0'$ ), we can make a full sweep of this value to look for the best case.
3. Optimization 3. As a third and more general case, we considered also modifying the remaining weights of the linear combination. In addition to the optimization carried out with the cascade-detector in the last case, an additional enhancement based on the free-derivative method is applied to the weights of the linear discriminant. The weights  $w_i, \forall i = 1, \dots, S \cdot M'$  of the first detector are now modified using the Nerled-Mead simplex method to directly minimize the average error rate of the

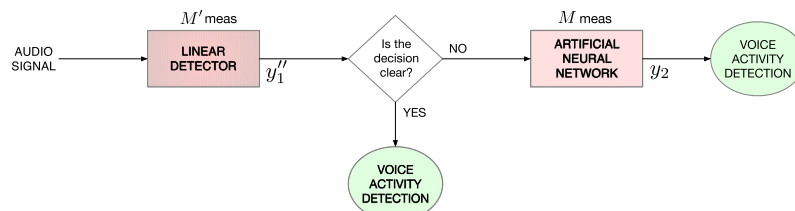


Fig. 1. Cascade configuration with a linear detector and an artificial network.

global system (that is, considering both detectors) over the design set. This additional optimization is applied to minimize the probability of converging to a local minimum.

## 5. Experiments and results

In this section, we will detail the dataset used in the proposal (Section 5.1) along with the experiments carried out and the results obtained (Section 5.2).

### 5.1. Dataset

Before detailing the different experiments carried out, it is worth looking at the acoustic dataset we have used. There exist various datasets available in the literature for testing VAD algorithms. In our case, the QUT-NOISE-TIMIT corpus has been chosen [29], as it is one of the most important in the field of study, so as to compare the proposed VAD with other proposals tested with the same dataset. Furthermore, the size of the dataset (number of scenarios and length of the signals) is large enough to avoid problems of generalization during training.

The dataset was developed by mixing some background noises and speech events, which were part of the TIMIT clean speech corpus [30]. In addition, the reverberant response of the environment was added to locations that required such effect. As a result, this dataset presents approximately 600 h of noisy speech sequences. In Table 1 the different scenarios (5 in total) and specific locations of each one are detailed.

It must be kept in mind that the recorders with both voice and noise are not fixed previously. The mixing of the background noise audios and the clean speech is executed randomly to allow each user to generate its own dataset. Effects of co-talking are also included, since each speech event is combined with the previous one with a probability of 50% (if not, silence appears), so it succeeds in simulating a hypothetical conversation between speakers.

The sampling frequency of the audios is  $F_s = 16$  kHz. It contains 24,000 speech sequences. The large amount of time and resources required for the full processing of the data has forced us to select a reduced subset from the full dataset. We have tried to include in the subset all the variety of signals that the dataset provides, as it is detailed below:

- The ten noise locations are included.
- Four Signal-to-Noise Ratios (SNRs) are included: 0 dB and 5 dB (considered as medium noise level), and  $-5$  dB and  $-10$  dB (considered as high noise level). We do not include the low noise case (10 dB and 15 dB) since it is the least problematic and the least realistic one.
- The three types of conversations provided by the dataset are included. The first type of conversation has less than 25% of speech in relation to the whole sequence, the second one has between 25% and 75% of speech, and the third one has more

than 75% of speech. This way, the data subset is balanced, so that there exists an average value of 50% of speech in the total duration of the audios.

- Sequences with the two possible durations are included: 1 and 2 min.
- In the full subset, each noise record was repeated in the same scenario. This eliminates the possibility of particular one-day sound effects. Both the sequences of the first and the second record have been included.

This results in a subset with a total of 240 sequences. If half the audios last 1 min and the other half last 2 min, it results in a dataset of 360 min (6 h).

A 10-fold cross-validation has been implemented in all the experiments. One of the locations is used as the training subset and the remaining locations are part of the test subset. The features were calculated with  $N = 128$  Discrete Fourier Transform (DFT) points (8 ms) with an overlapping of 50% (one time frame every 4 ms). The detection is performed every 16 ms, that is, every four time frames ( $T = 4$ ). All these values are standard in algorithms for hearing aids [31,32].

### 5.2. Experiments and results

In this study, we try to reduce the probability of error when detecting voice activity in noisy conversations, as it has been detailed previously. The behavior and importance of some parameters have been tested, with computational cost being the most important one. Because of that, a cost constraint related to the required operations,  $\bar{C}$ , has been included. It can take 15 possible values: from 10 to 100 KIPS, in steps of 10 KIPS, and from 120 to 200 KIPS, in steps of 20 KIPS, being KIPS equal to thousands of instructions per second. To determine this range of values, we have compared the power consumption of the hearing aid in a simple case in which only the compression and the VAD algorithms are implemented. Using a real hearing aid with a DSP working at 1.92 MHz, we estimated an average power consumption of 0.87 mW in the case of only implementing the main compression algorithm, which required 1100 KIPS, and an average power consumption of 0.90 mW in the case of also implementing the proposed VAD algorithm running with 200 KIPS. Thus, it is clear that the consumption associated with the selected range of KIPS values will be suitable for the processor of a typical hearing aid.

The main objective of this proposal is to present the effect that cascade-detectors have on the performance of VAD algorithms. As mentioned previously, the idea is to keep the same average cost as in the non-cascade detectors but with a better performance of the system. This is achieved by taking the features from the MLP based detector and reducing the average computational cost  $\bar{C}$ , forcing the linear system to take some of the decisions.

Five previous values of  $\bar{C}$  have been considered concerning the value of  $C_2$ , except in the first configurations, where it was possible to consider between 1 to 4 previous values. The reason for limiting the number of previous average costs  $\bar{C}$  is that in the preliminary experiments we noted that the performance decreases if we force the system to use very complex features with a low number of resources. That is, the cascade configuration does not work properly if the number of features used in the first stage is too high ( $\alpha$  close to 1). In total, 14 cascade configurations have been tested in the experiments. The parameters that characterize each of them are shown in Table 2.

This way, in configuration A the features from the original MLP-detector with  $C_2 = 20$  KIPS are taken into account, and average computational costs of  $\bar{C} = 10$  and 20 KIPS are considered. In the case of configuration B, the features from the original MLP-

**Table 1**  
Scenarios and locations where the dataset was recorded.

Scenario	Location A	Location B
Cafe	Outdoor cafe	Indoor shopping center food-court
Home	Kitchen	Living room
Street	Roadside near inner-city	Outer-city traffic-light controlled intersections
Car	Car with windows down	Car with windows up
Reverb	Indoor pool	Enclosed carpark

**Table 2**  
Computational cost values of the different cascade configurations considered.

Configuration	$C_2$ (KIPS)	$\bar{C}$ (KIPS)
A	20	10, 20
B	30	10, 20, 30
C	40	10, 20, 30, 40
D	50	10, 20, 30, 40, 50
E	60	10, 20, 30, 40, 50, 60
F	70	20, 30, 40, 50, 60, 70
G	80	30, 40, 50, 60, 70, 80
H	90	40, 50, 60, 70, 80, 90
I	100	50, 60, 70, 80, 90, 100
J	120	60, 70, 80, 90, 100, 120
K	140	70, 80, 90, 100, 120, 140
L	160	80, 90, 100, 120, 140, 160
M	180	90, 100, 120, 140, 160, 180
N	200	100, 120, 140, 160, 180, 200

detector with  $C_2 = 30$  KIPS are taken into account, and average computational costs of  $\bar{C} = 10, 20$  and  $30$  KIPS are considered. The same applies to the rest of the configurations.

To explain clearly the obtained results, we will focus at the beginning in the case of having audios with a medium SNR (0 dB and 5 dB), in Section 5.2.1. This way, we will appreciate the effect of applying the new proposal in a case that is neither ideal nor very noisy. Later, we will study how the values change when having a complex environment, that is, when the SNR is lower (-5 dB and -10 dB). It will be explained along Section 5.2.2. Lastly, a comparison with other proposals from the literature is included in Section 5.2.3.

### 5.2.1. Medium noise environment

In order to analyze the behavior of these configurations, we will show the most significant results in Table 3. In this table, the main parameter is the probability of error  $P_e$ , which can be defined as the number of detections that predict an incorrect result (e.g., these predictions which predict voice activity when there is not it, and vice versa) divided by the total number of time slots analyzed. The probability of error of the regular detector  $P_{er}$ , the lowest probabilities of error of the optimizations 1, 2 and 3 ( $P_{e,c_1}$ ,  $P_{e,c_2}$  and  $P_{e,c_3}$ ) and the computational cost of the simple MLP based detector  $C_2$  associated to those results, for each average computational cost  $\bar{C}$ , are represented. For each configuration, the optimizations which get the best results are shown in bold.

**Table 3**  
Most significant results obtained for the different average computational costs  $\bar{C}$ , including the lowest probabilities of error in each of the optimizations and the values of  $C_2$  associated with them; SNR = 0 and 5 dB (medium noise environment).

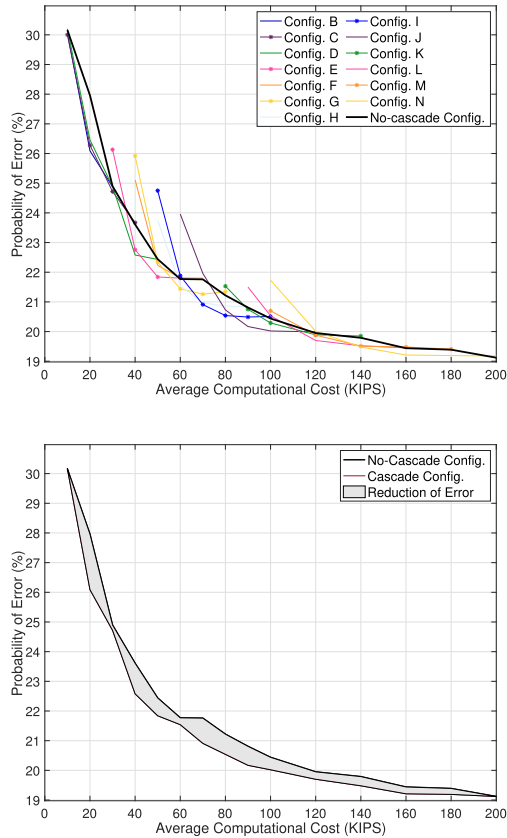
$\bar{C}$	$P_{er}$ (%)	Opt. 1		Opt. 2		Opt. 3	
		$C_2$	$P_{e,c_1}$ (%)	$C_2$	$P_{e,c_2}$	$C_2$	$P_{e,c_3}$ (%)
10	30.16	30	30.27	30	<b>30.00</b>	40	<b>30.00</b>
20	27.96	40	26.30	30	<b>26.07</b>	30	26.09
30	24.90	30	24.68	40	<b>24.29</b>	40	24.72
40	23.61	50	23.19	50	22.66	50	<b>22.58</b>
50	22.44	60	21.99	60	21.99	60	<b>21.84</b>
60	21.77	80	21.60	60	21.76	80	<b>21.44</b>
70	21.76	90	21.08	90	21.17	100	<b>20.91</b>
80	21.22	100	20.56	100	20.55	100	<b>20.54</b>
90	20.81	120	20.39	100	20.45	120	<b>20.17</b>
100	20.44	120	20.05	120	20.04	120	<b>20.02</b>
120	19.95	160	19.69	160	<b>19.68</b>	160	19.70
140	19.79	160	<b>19.48</b>	160	<b>19.48</b>	200	<b>19.48</b>
160	19.44	200	19.29	200	19.29	200	<b>19.21</b>
180	19.39	200	<b>19.16</b>	200	19.17	200	19.19
200	19.12	200	19.13	200	19.13	200	19.18

First of all, it can be shown that in almost all the configurations there are some improvements in terms of error. One of the most relevant improvement is presented when applying  $\bar{C} = 20$  KIPS and optimization 2 with  $C_2 = 30$  KIPS, where the error is reduced from 27.96% to 26.07%. The detection is improved almost 2% in absolute terms, turning into 6.72% in relative error. Another remarkable improvement can be found when using  $\bar{C} = 70$  KIPS and optimization 3 with  $C_2 = 100$  KIPS, where the error falls from 21.76% in the original detector to 20.91%. This means a reduction of almost 1%, which in relative terms means 3.91% of improvement in the system.

Furthermore, this system could provide a reduction in terms of cost but keeping the same probability of error. For instance, instead of using the original detector with  $\bar{C} = 60$  KIPS, which provides an error of 21.77%, we could obtain practically the same result with optimization 3. Using  $\bar{C} = 50$  KIPS, we get an error of 21.84% (just 0.07% worse) but reducing the cost in 10 KIPS (16.67% of relative saving). Another example is found when we have an average computational cost of  $\bar{C} = 120$  KIPS. While the original detector gets 19.95% of error, we obtain 20.17% of error applying the optimization 3 with  $\bar{C} = 90$  KIPS. It represents a relative reduction of 25% in terms of cost, at the expense of a performance loss of 0.22%.

Now we will try to extract some conclusions about the three different optimizations applied to the original detector. As mentioned before, in Table 3 the optimizations which get the best results in each configuration are shown in bold. Looking at this, we can study if any optimization works better than the others. The last value of  $\bar{C}$  (200 KIPS) does not allow us to obtain any conclusion, as no improvement has been reached with the limits considered in the experiments. It can be seen that optimization 3 gets the best results (10 of 14) in most cases, followed by optimization 2 (3 of 14). In two of them there are one or more ties. In conclusion, the application of additional optimizations to optimization 1 seems to be useful. It can be deduced that the optimization based on the free-derivative method can provide us some additional improvement without increasing the cost.

In Fig. 2 the probability of error  $P_e$  as a function of the average computational cost  $\bar{C}$  (in KIPS) is represented graphically. The solid black line represents the performance when cascade-detectors are not applied (single MLP based approach). Thus, a regular detector based on MLPs with some cost constraints is applied. On the other hand, colored lines represent the different configurations included in Table 2 when using optimization 3 from the algorithm, which has proved to be the best. The final result when using this approach



**Fig. 2.** Probability of Error obtained using a detector without cascade implementations (bold black line) and applying the optimization 3 of the cascade-detectors (colored lines), as a function of the average computational cost. In the top chart, the results obtained in each configuration (A, B, ..., N) are shown, while the bottom one shows an approximation to the final improvement if we take into account the best results from the different configurations applying the optimization 3. Medium SNR (0 dB and 5 dB) is tested in this figure.

is shown in the bottom chart, where the grey area represents the improvement in system performance when configuration 3 is applied. As it can be seen at a glance, the probability of error is lower when using them.

However, a general trend in the results is that the performance typically improves more stably for values of  $\bar{C}$  close to the value of  $C_2$  considered. It can be clearly seen in Fig. 2, where the performance of the system decreases when values of  $\bar{C}$  are far away from the original value of  $C_2$ . At this point, we must think about how much the value of the average cost  $\bar{C}$  should be reduced with respect to the value of the original detector  $C_2$ , because the results can fall if this difference is huge. With this purpose, we define the Relative Cost Reduction (RCR) as the relative difference between costs. This parameter is defined in Eq. (27).

$$RCR = \frac{C_2 - \bar{C}}{C_2} \cdot 100(\%) \quad (27)$$

Now the idea is to take the points that have obtained better results in the previous configurations to find a value of RCR that ensures us it is well worth applying cascade-detectors. These results are detailed in Table 4, where  $\min(P_{e,c})$  represents the minimum error from the three optimizations.

We can define the limit value of RCR (in bold in Table 4) for a given configuration as the value until which this configuration gets better results than the others. It can be seen that in the cascade-detector with  $C_2 = 200$  KIPS, results are improved from  $\bar{C} = 160$  until 200 KIPS, since this detector has proved to be the best at this range. The resulting RCR is 20%. In the case of the cascade-detector with  $C_2 = 160$  KIPS, results are improved from  $\bar{C} = 120$  until 140 KIPS, resulting in a limit value of RCR = 25%.

These mentioned limit values of RCR give us an idea of how these cascade-configurations must be applied to get better results in our experiments. The reduction of the average cost  $\bar{C}$  should not reach 20–25% of the original cost  $C_2$ . It means that in other experiments, it seems reasonable to place more cascade configurations in low values of  $\bar{C}$  than in higher ones, so that they will be spaced in a non-linear way. Thus, we ensure that with reductions in cost around 20–25% the results are improved, and additional useless configurations are not simulated.

### 5.2.2. High noise environment

Now we will see how the system works when the SNR is worse than before, taking values of  $-5$  dB and  $-10$  dB, depending on the tested audio. In Table 5 the best results for each computational cost  $\bar{C}$  are shown, similarly to Table 3.

As could be expected, the values of the probability of error are worse as the voice of the conversation is more masked by the background noise. However, the most important issue in this proposal is to see if the cascade configuration continues improving the results that we would get without applying it. As shown in the table, the performance generally improves when using any of the three optimizations. One of the most relevant improvement appears when applying  $\bar{C} = 30$  KIPS and optimization 1 with  $C_2 = 80$  KIPS, where the error is reduced from 40.88% to 37.40%. The detection is improved by almost 3.5 points in absolute terms, which means an improvement of 8.51% in relative error.

Furthermore, it is possible to reduce the computational cost of the system while keeping the prior probability of error, as it happened in Section 5.2.1. For example, when using the original detector with  $\bar{C} = 140$  KIPS, we get an error of 35.70%. We could get the same value using optimization 1 and  $\bar{C} = 70$  KIPS, which reports an error of 35.69%. This way, it is feasible to save half of the instructions per second and obtain the same error.

**Table 4**  
Values of the parameter RCR for each value of  $\bar{C}$ .

$\bar{C}$ (KIPS)	$\min(P_{e,c})$ (%)	$C_2$ (KIPS)	RCR (%)
10	30.00	30	<b>67</b>
20	26.07	30	33
30	24.29	40	<b>25</b>
40	22.58	50	20
50	21.84	50	0
60	21.44	80	<b>25</b>
70	20.91	100	<b>30</b>
80	20.54	100	20
90	20.17	120	<b>25</b>
100	20.02	120	17
120	19.68	160	<b>25</b>
140	19.48	160	13
160	19.21	200	<b>20</b>
180	19.16	200	10

**Table 5**  
Most significant results obtained for the different average computational costs  $\bar{C}$ , including the lowest probabilities of error in each of the optimizations and the values of  $C_2$  associated with them. SNR = -5 and -10 dB (high noise environment).

$\bar{C}$	$P_{er}$ (%)	Opt. 1		Opt. 2		Opt. 3	
		$C_2$	$P_{e,c1}$ (%)	$C_2$	$P_{e,c2}$	$C_2$	$P_{e,c3}$ (%)
10	42.82	30	<b>41.78</b>	20	42.27	30	41.97
20	41.86	60	38.47	50	<b>38.43</b>	60	38.53
30	40.88	60	<b>37.40</b>	80	37.49	80	37.53
40	39.54	80	37.07	80	<b>37.01</b>	80	37.12
50	38.45	60	36.48	80	36.55	80	<b>36.36</b>
60	37.40	80	<b>35.93</b>	80	35.99	80	36.15
70	37.05	100	<b>35.69</b>	100	35.76	100	35.75
80	36.87	90	<b>35.40</b>	90	35.47	100	35.73
90	36.50	100	35.50	120	<b>35.45</b>	100	35.55
100	36.20	120	<b>35.40</b>	120	35.42	120	35.41
120	36.03	140	35.29	140	35.30	140	<b>35.27</b>
140	35.70	160	35.20	160	35.22	160	<b>35.10</b>
160	35.33	200	<b>34.80</b>	200	35.09	200	34.99
180	34.71	200	34.27	200	34.33	200	<b>34.17</b>
200	34.30	200	34.13	200	34.25	200	<b>34.10</b>

Regarding the different optimizations applied, it is not easy to determine which one works better in this case. From Table 6, it can be seen that Optimization 1 seems to work better when having low values of computational cost, while Optimization 3 provides better results when the available resources are higher. However, it could be concluded that applying additional optimizations (opt. 2 and 3) is useful when the noise is medium or low, but it might not be advantageous when the background noise is higher (low SNR).

Fig. 3 shows a graph similar to the one represented in Fig. 2 but testing low values of SNR. It can be seen that the scope for improvement is bigger than before, as the error is worst too. The usefulness of the configurations is noticeable, given the reduction of error achieved (the grey area in the bottom graph).

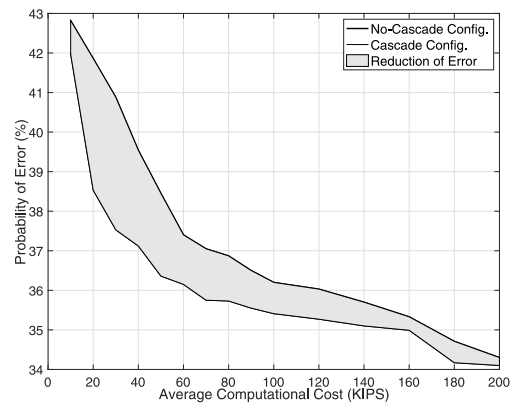
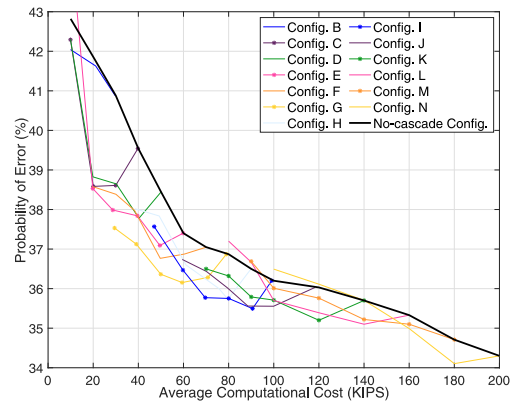
In Table 6, we have summarized the best result obtained in each computational cost, as well as the value of RCR defined in the previous section. With the exceptions of the lowest computational costs, it is shown that a suitable value for RCR could be between 20% and 30%, as it was previously deduced.

5.2.3. Comparison with other proposals

To conclude with this section, we compare the results with previous approaches. Table 7 includes the results obtained with different methods found in the literature applied to the same database. Specifically, we compare our results with three different proposals [12–14], in terms of the error rate for the same SNR. In [12] the authors tested a system based on a Hidden Markov Models (HMM), in [13] Long-Term Spectral Estimation (LTSE) was applied,

**Table 6**  
Values of the parameter RCR for each value of  $\bar{C}$ .

$\bar{C}$ (KIPS)	$\min(P_{er})$ (%)	$C_2$ (KIPS)	RCR (%)
10	41.78	30	67
20	38.43	50	60
30	35.40	60	50
40	37.01	80	50
50	36.36	80	37
60	35.93	80	25
70	35.69	100	30
80	35.40	90	11
90	35.45	120	25
100	35.40	140	28
120	35.27	160	25
140	35.10	200	30
160	34.80	200	20
180	34.17	200	10



**Fig. 3.** Probability of Error obtained using a detector without cascade implementations (bold black line) and applying the optimization 3 of the cascade-detectors (colored lines), as a function of the average computational cost. In the top chart, the results obtained in each configuration (A, B, ..., N) are shown, while the bottom one shows an approximation to the final improvement if we take into account the best results from the different configurations applying the optimization 3. Low SNR (-5 dB and -10 dB) is tested in this figure.



**Table 7**  
Comparative results with VAD baseline systems.

Method	$P_e$ (%), medium SNR	$P_e$ (%), low SNR	Window length
HMM [12]	24.48	35.16	10 ms
LTSE [13]	19.87	32.51	32 ms
CS-LDA (2 ch) [14]	19.96	29.63	128 ms
SDOI (1 ch) [14]	15.21	30.80	128 ms
EFLEC + MLP (100 KIPS) [11]	20.44	36.20	8 ms
EFLEC + MLP (200 KIPS) [11]	19.12	34.30	8 ms
EFLEC + Prop. Cascade System (80 KIPS)	20.54	35.40	8 ms
EFLEC + Prop. Cascade System (160 KIPS)	19.21	34.80	8 ms

and in [14] Circularity Spectrum with Linear Discriminant Analysis (CS-LDA) was used, as well as an analysis based on the Summed Degree of Impropriety (SDOI). Some of the results obtained in previous approaches have been found in [14]. From this comparison, we can highlight the following issues:

- The method that performs better under this database is the SDOI when having a medium value of SNR, and the CS-LDA when testing a low value of SNR. Unfortunately, these methods, as many of the other methodologies, suppose the use of considerably large time windows (128 ms, and a high level of overlapping). Implementing these methods implies that the VAD process might use a time-frequency analysis different from the main one used to implement the algorithm that overcomes the hearing losses, highly increasing the computational cost and making the method unpractical for hearing aid applications. Please, note here that the maximum delay of the hearing aid device cannot exceed 20 ms [7], which implies using window lengths typically shorter than 10 ms.
- The method in [12] is computationally efficient for being implemented in hearing aids since it fulfills the short window length requirement. To ensure a fair comparison, we have followed a similar process for counting the number of instructions to the one used in this approach. Considering the evaluation of the maximum likelihood (ML) criterion and the update of the level of energy for noisy frames, this system requires around 290 KIPS. This value is greater than the computational cost required by our system.
- The proposed methods represent the best solution among those with short windows. In the case of low SNR values, the LTSE method works better than the cascade system, but it is still using windows of 32 ms length, which is not a suitable value in hearing aids. The cascade system reduces the required computational resources by 20% compared with an MLP based VAD.

## 6. Conclusions

In this approach, we have tried to investigate the effect that a set of cascade-detectors has on the performance of a VAD system. It is a system thought to be used in hearing aids, and in this way, the computational cost has been adapted to the typical values of that field.

The results yield some interesting conclusions. First of all, the usefulness of the new configuration has been proved, as we have succeeded in reducing the probability of error of the system while maintaining the same computational cost. Numerically it has reached a typical value of 2–3% of relative improvement, reaching 7% in some cases for medium SNR values, while an 8.5% of relative improvement has been reached when using low SNR values.

Regarding the different configurations proposed, it seems that a cascade configuration can be useful, but extra optimizations can

provide improvements in terms of performance, especially when the environment is not very noisy.

We have estimated that the average computational cost  $\bar{C}$  of the system should not be reduced by more than 20%–30% of the original value of the more complex detector  $C_2$ , as the results are worse if the same number of features continues to be used for lower computational costs.

A comparison with other methods from the literature has shown that the proposals of this study get the best solutions if we compare with similar short-window algorithms (8–10 ms). Other algorithms get better results when applying window lengths higher than 20 ms to VAD, but this is not practical for hearing aid applications.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Funding: This work was supported by the Spanish Ministry of Science and Innovation/FEDER under Project RTI2018-098085-B-C42 and by the University of Alcalá.

## References

- [1] Davis A. The prevalence of hearing impairment and reported hearing disability among adults in great britain. *Int J Epidemiol* 1989;18(4):911–7.
- [2] W.H. Organization, Deafness and hearing loss; 2020. URL <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>.
- [3] Lin FR. Hearing loss in older adults: who's listening?. *Jama* 2012;307(11):1147–8.
- [4] Amieva H, Ouvrard C, Meillon C, Rullier L, Dartigues J-F. Death, depression, disability, and dementia associated with self-reported hearing problems: A 25-year study. *J Gerontol Ser A* 2018;73(10):1383–9.
- [5] Amieva H, Ouvrard C, Giulioli C, Meillon C, Rullier L, Dartigues J-F. Self-reported hearing loss, hearing aids, and cognitive decline in elderly adults: A 25-year study.
- [6] Gil-Pita R, García-Gómez J, Bautista-Durán M, Combarro E, Cocana-Fernandez A. Evolved frequency log-energy coefficients for voice activity detection in hearing aids. In: 2017 IEEE international conference on fuzzy systems (FUZZ-IEEE). IEEE. p. 1–6.
- [7] Stone MA, Moore BC. Tolerable hearing aid delays. II. Estimation of limits imposed during speech production. *Ear Hear* 2002;23(4):325–38.
- [8] Ramirez J, Górriz JM, Segura JC. Voice activity detection, fundamentals and speech recognition system robustness. In: Robust speech recognition and understanding. IntechOpen; 2007.
- [9] Lee C-H, Kates JM, Rao BD, Garudadri H. Speech quality and stable gain trade-offs in adaptive feedback cancellation for hearing aids. *J Acoust Soc Am* 2017;142(4):EL388–94.
- [10] Gong Z, Xia Y. Two speech enhancement-based hearing aid systems and comparative study. In: 2015 5th international conference on information science and technology (ICIST). IEEE; 2015. p. 530–4.
- [11] García-Gómez J, Mohino-Herranz I, Clares-Crespo C, Fernández-Toloba A, Gil-Pita R. Analysis of the performance of evolved frequency log-energy coefficients in hearing aids for different cost constraints and scenarios. In: Audio Engineering Society Convention 145; 2018. URL <http://www.aes.org/e-lib/browse.cfm?elib=19837>.
- [12] Sohn J, Kim NS, Sung W. A statistical model-based voice activity detection. *IEEE Sig Process Lett* 1999;6(1):1–3.
- [13] Ramirez J, Segura JC, Benitez C, De La Torre A, Rubio A. Efficient voice activity detection algorithms using long-term speech information. *Speech Commun* 2004;42(3–4):271–87.
- [14] Wisdom S, Okopal G, Atlas L, Pitton J. Voice activity detection using subband noncircularity. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. p. 4505–9.
- [15] Graf S, Herbig T, Buck M, Schmidt G. Features for voice activity detection: a comparative analysis. *EURASIP J Adv Sig Process* 2015;2015(1):91.
- [16] Mukherjee H, Obaidullah SM, Santosh K, Phadikar S, Roy K. Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. *Int J Speech Technol* 2018;21(4):753–60.
- [17] Sehgal A, Kehtarnavaz N. A convolutional neural network smartphone app for real-time voice activity detection. *IEEE Access* 2018;6:9017–26.
- [18] Kim J, Hahn M. Voice activity detection using an adaptive context attention model. *IEEE Signal Process Lett* 2018;25(8):1181–5.

- [19] Gil-Pita R, Ayllón D, Ranilla J, Llerena-Aguilar C, Díaz I. A computationally efficient sound environment classifier for hearing aids. *IEEE Trans Biomed Eng* 2015;62(10):2358–68.
- [20] Xiang J-J, McKinney MF, Fitz K, Zhang T. Evaluation of sound classification algorithms for hearing aid applications. In: *Acoustics speech and signal processing (ICASSP)*. 2010 IEEE International Conference on. IEEE; 2010. p. 185–8.
- [21] Van Trees HL. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons; 2004.
- [22] Rosenblatt F. *Principles of neurodynamics: Perceptions and the theory of brain mechanisms*.
- [23] Ye J. Least squares linear discriminant analysis, in. In: *Proceedings of the 24th international conference on Machine learning*. ACM. p. 1087–93.
- [24] Gil-Pita R, Alexandre E, Cuadra L, Vicen R, Rosa-Zurera M. Analysis of the effects of finite precision in neural network-based sound classifiers for digital hearing aids. *EURASIP J Adv Sig Process* 2009;2009(1):456945.
- [25] Bishop CM, Nasrabadi NM. *Pattern recognition and machine learning*, vol. 1. New York: Springer; 2006.
- [26] Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. *J Soc Industr Appl Math* 1963;11(2):431–41.
- [27] Kosmatopoulos EB, Polycarpou MM, Christodoulou MA, Ioannou PA. High-order neural network structures for identification of dynamical systems. *IEEE Trans Neural Netw* 1995;6(2):422–31.
- [28] Pudil P, Novovicova J, Blaha S, Kittler J. Multistage pattern recognition with reject option. In: *Proceedings, 11th IAPR international conference on pattern Recognition, Conference B: Pattern Recognition Methodology and Systems*, vol. II. IEEE; 1992. p. 92–5.
- [29] Dean DB, Sridharan S, Vogt RJ, Mason MW. The qut-noise-timit corpus for the evaluation of voice activity detection algorithms. *Proc Interspeech* 2010.
- [30] Fisher WM. The darpa speech recognition research database: specifications and status. In: *Proc. DARPA Workshop on Speech Recognition*; Feb. 1986; 1986. p. 93–9.
- [31] Chong K, Gwee B, Chang JS. A low energy fft/ift processor for hearing aids. 2017 IEEE international symposium on circuits and systems 2007:1169–72.
- [32] Ayllón D, Gil-Pita R, Rosa-Zurera M. Rate-constrained source separation for speech enhancement in wireless-communicated binaural hearing aids. *EURASIP J Adv Sig Process* 2013;2013(1):187.

## Chapter 6

# Article 5: Smart Sound Processing for Defect Sizing in Pipelines Using EMAT Actuator Based Multi-Frequency Lamb Waves

### AUTHORS

Joaquín García-Gómez, Roberto Gil-Pita, Manuel Rosa-Zurera, Antonio Romero-Camacho, Jesús Antonio Jiménez-Garrido, Víctor García-Benavides

### JOURNAL

Sensors. Special issue “State-of-the-Art Sensors Technology in Spain 2017”  
D.O.I.: <https://doi.org/10.3390/s18030802>

### RANKING

JCR (2018): 3.076  
Quartile Rank - Instruments & Instrumentation: 13/61 (Q1)  
Quartile Rank - Physics, Applied: 42/148 (Q2)  
Quartile Rank - Engineering, Electrical & Electronic: 87/266 (Q2)

### CONTRIBUTION TO THE SCOPE OF THE THESIS


In this publication, the issue of Pipeline Defect Assessment (PDA) is addressed. Pipeline inspection is a topic of particular interest to the companies. Defect sizing is especially important, since it allows them to avoid subsequent costly repairs in their equipment. A solution for this issue is using ultrasonic waves sensed through Electro-Magnetic Acoustic Transducer (EMAT) actuators. The main advantage of this technology is the absence of the need to have direct contact with the surface of the material under investigation, which must be a conductive one. Specifically interesting is the meander-line-coil-based Lamb wave generation, since the directivity of the waves allows a study based in the

circumferential wrap-around received signal. However, the variety of defect sizes changes the behavior of the signal when it passes through the pipeline. Because of that, it is necessary to apply advanced techniques based on smart sound processing. These methods involve extracting useful information from the signals sensed with EMAT at different frequencies to obtain nonlinear estimations of the depth of the defect, and to select the features that better estimate the profile of the pipeline. The proposed technique has been tested using both simulated and real signals in steel pipelines, obtaining promising results in terms of Root Mean Square Error (RMSE). Furthermore, it has been demonstrated the importance of applying a multi-frequency study for defect sizing problem, the relevance of some features from the signals (e.g., energy and amplitude), and the absence of the need to increase significantly the complexity of the classifiers to get a good estimation in the problem at hand.



Article

## Smart Sound Processing for Defect Sizing in Pipelines Using EMAT Actuator Based Multi-Frequency Lamb Waves

Joaquín García-Gómez <sup>1</sup>, Roberto Gil-Pita <sup>1,\*</sup> , Manuel Rosa-Zurera <sup>1</sup>, Antonio Romero-Camacho <sup>2</sup>, Jesús Antonio Jiménez-Garrido <sup>2</sup> and Víctor García-Benavides <sup>2</sup>

<sup>1</sup> Department of Signal Theory and Communications, University of Alcalá, Ctra. Madrid-Barcelona, km. 33,600, 28805 Alcalá de Henares, Spain; joaquin.garciagomez@uah.es (J.G.-G.); manuel.rosa@uah.es (M.R.-Z.)

<sup>2</sup> Innerspec Technologies Europe S.L, Av. de Madrid 2, 28802 Alcalá de Henares, Spain; aromero@innerspec.com (A.R.-C.); jjimenez@innerspec.com (J.A.J.-G.); vgarcia@innerspec.com (V.G.-B.)

\* Correspondence: roberto.gil@uah.es; Tel.: +34-91-885-6751

Received: 15 January 2018; Accepted: 5 March 2018; Published: 7 March 2018

**Abstract:** Pipeline inspection is a topic of particular interest to the companies. Especially important is the defect sizing, which allows them to avoid subsequent costly repairs in their equipment. A solution for this issue is using ultrasonic waves sensed through Electro-Magnetic Acoustic Transducer (EMAT) actuators. The main advantage of this technology is the absence of the need to have direct contact with the surface of the material under investigation, which must be a conductive one. Specifically interesting is the meander-line-coil based Lamb wave generation, since the directivity of the waves allows a study based in the circumferential wrap-around received signal. However, the variety of defect sizes changes the behavior of the signal when it passes through the pipeline. Because of that, it is necessary to apply advanced techniques based on Smart Sound Processing (SSP). These methods involve extracting useful information from the signals sensed with EMAT at different frequencies to obtain nonlinear estimations of the depth of the defect, and to select the features that better estimate the profile of the pipeline. The proposed technique has been tested using both simulated and real signals in steel pipelines, obtaining good results in terms of Root Mean Square Error (RMSE).

**Keywords:** EMAT actuators; Lamb waves; pipeline inspection; defect sizing; smart sound processing

### 1. Introduction

Ultrasonic techniques have demonstrated over the years to be really useful for Non-Destructive Testing (NDT) examinations [1–3]. Conventional ultrasounds are primarily generated taking advantage of the piezoelectric effect. Although it is an efficient way of generating ultrasounds, a proper coupling between the transducer and test specimens is needed, which is a disadvantage. Therefore, materials inspected by conventional ultrasounds are covered with a thin layer of fluid. EMAT (Electro-Magnetic Acoustic Transducer) actuators are able to generate and receive ultrasonic waves without the need to have thorough contact with the surface of the material under investigation [4]. This technology is capable of generating multiple types of waves: Lamb, shear, longitudinal and Rayleigh. Besides, when EMAT technique is implemented with a meander-line-coil, the waves are generated in a directional way [5,6]. This fact is interesting since it allows differentiating between circumferential and axial scans.

A highlighted application of this technology is the pipeline inspection [7]. On the one hand, some manuscripts have focused on the defect detection and location in its circumferential path, mainly using shear waves [8–10]. However, it is important to obtain not only the position of the defect, but also its residual thickness. For the companies it is interesting to know this parameter, since it is a vital factor to make the decision to replace a section of the pipeline [11]. On the other hand, there have

been some proposals of sizing techniques applied to pipeline inspection, mainly based on the analysis of the physical mode [12], but the distortion caused by the defects over the different modes strongly varies with the shape of the defect [12–14]. Smooth defects usually reflect less energy than abrupt ones, independently of the residual thickness of the pipe caused by the defect. Thus, the amplitude of the received echo is strongly related not only to the depth of the defect but also to its hardness, and both the amplitude and the time of arrival of the wrap-around signal vary with the length of the defect.

In general, the studies have followed the same line with regard to the information extracted from the wave modes. The most relevant and used parameters are the amplitude and the phase from the received signal [14,15]. Once this information is obtained, the use of Smart Sound Processing (SSP) techniques is suitable for solving sizing problems using EMAT guided waves. These methods involve extracting useful information from the sensed acoustic signals and applying nonlinear techniques to obtain estimations of useful parameters. Other proposals have applied this type of techniques in their studies, including: Artificial Neural Networks [16], Neural Networks with large number of neurons [17] or Adaptive Neuro-Fuzzy Inference Systems [18]. Using this type of methods is interesting to excite the coil at multiple frequencies, as the behavior of the Lamb modes is different depending on this parameter, and SSP allows to combine all this information and get better defect estimation results.

In this sense, this paper studies pipeline inspection mixing both the EMAT and SSP techniques. Specifically, EMAT-based Lamb waves will be generated at multiple frequencies. Axial scans will be developed and the circumferential path followed by waves will allow the analysis of the wrap-around signals received. These ultrasound signals will be related to the behavior of the pipeline depending on its profile, conditions and damage. Once these signals are measured, it will be possible to apply SSP techniques in order to get useful information from the amplitude and phase of the multi-frequency signals. In particular, feature selection techniques and Neural Networks-based estimators will be applied. Following this process, it will be feasible to obtain an approximated characterization of the residual thickness along the pipeline.

## 2. Materials and Methods

In this section the sensorization of Lamb waves through EMAT actuators will be described. Initially, EMAT technology and the fundamentals of Lamb waves will be introduced. A brief hardware description will be made at the end of the section as well.

### 2.1. Lamb Wave Generation Using EMAT Actuators

EMAT transducers consist of a coil wire and a magnet. The alternating electrical current flowing through the coil wire placed in a uniform magnetic field ( $B$ ) near the surface of a ferromagnetic material, induces surface currents (Eddy Currents,  $J$ ) in the material. The field generated by electrical coils interacts with the field generated by the magnet producing a Lorentz force ( $F$ ) according to Equation (1).

$$F = J \times B \quad (1)$$

The disturbance is applied to the lattice of the material, producing an elastic wave. In a reciprocal process (reception of an ultrasonic wave), the interaction of elastic waves in the presence of a magnetic field induces currents in the EMAT receiver coil circuit. In Figure 1 a comparative between the generation of ultrasonic waves using conventional ultrasound methods and using EMAT technology is represented.

The advantages of using EMAT over piezoelectric transducers are: as the transduction process occurs within an electromagnetic depth skin, it is a couplant free technique; it is insensitive to surface conditions, being capable of inspecting rough, dirty (oily/wet), oxidized or uneven surfaces; inspection can be carried out on flat, curved or complex surfaces; it allows high speed inspections (up to 60 m/s), high temperature inspections and low temperature inspections, and it can generate Lamb, Shear

Horizontal (SH), Shear Vertical (SV), Longitudinal and Rayleigh waves due to its good selectivity in frequency. On the other hand, the challenges of EMAT are the high level of power required, the bigger size of the transducers and the lower Signal to Noise Ratio (SNR). Besides, the material under inspection needs to be conductive.

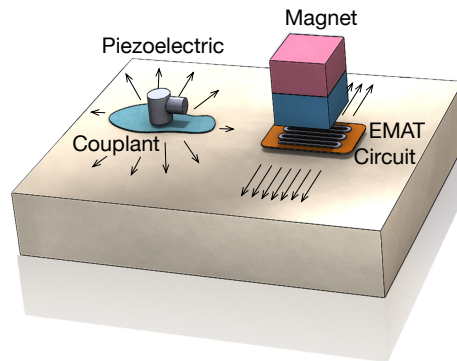


Figure 1. Conventional Ultrasound vs EMAT.

Guided Wave Testing is a NDT technique that employs ultrasonic stress waves that propagate along a structure while guided by its boundaries. Guided waves permit covering long distances from a single point with a limited number of sensors, being very effective for rapid scanning of pipelines and tanks. On relatively thin structures, it is possible to generate volumetric guided waves that fill up the material and permit a complete, volumetric inspection. The most common types of volumetric waves are SH and Lamb.

Lamb waves travel throughout the material with both vertical and forward motion in an elliptical pattern. These waves are dispersive by nature, and very sensitive to thickness variations. They can be classified in symmetric (also known as longitudinal) and asymmetric (also known as flexural) modes. The introduction of boundary conditions makes Lamb wave problems inherently more difficult than the more conventional bulk waves. Unlike the finite number of modes present in a bulk wave problem, there are an infinite number of modes associated with a given Lamb wave application. That is, a finite body can support an infinite number of different Lamb wave modes. Now the generation of the Lamb wave modes will be described. With this purpose, Lamé parameters will be defined. Lamé parameters are two material-dependent quantities denoted by  $\lambda$  (Lamé's first parameter) and  $\mu$  (Lamé's second parameter). They are defined by Equations (2) and (3).

$$\lambda = \frac{Ev}{(1+\nu)(1-2\nu)} \quad (2)$$

$$\mu = \frac{E}{2(1+\nu)} \quad (3)$$

where  $E$  is the Young's modulus, which measures the stiffness of a material, and  $\nu$  is the Poisson's ratio, which is an elastic constant that measures how an elastic, linear and isotropic material is narrowed when it is longitudinally stretched.

Then the elastic wave equation needs to be taken into account.

$$\mu \nabla^2 \mathbf{u} + (\lambda + \mu) \nabla \nabla \cdot \mathbf{u} = \rho \left( \frac{\partial^2 \mathbf{u}}{\partial t^2} \right), \quad (4)$$

where  $\rho$  represents the density of the material under inspection. Applying the Helmholtz decomposition, the displacement field  $\mathbf{u}$  can be split into a rotational component  $\nabla \times \mathbf{H}$  and an irrotational component  $\nabla \phi$ :

$$\mathbf{u} = \nabla \phi + \nabla \times \mathbf{H} \quad (5)$$

Then, the system of partial differential equations can be rewritten as:

$$c_L \nabla^2 \phi = \frac{\partial^2 \phi}{\partial t^2} \quad (6)$$

$$c_T \nabla^2 \mathbf{H} = \frac{\partial^2 \mathbf{H}}{\partial t^2}, \quad (7)$$

where  $c_L = \sqrt{(\lambda + 2\mu)/\rho}$  and  $c_T = \sqrt{\mu/\rho}$  represent the sound velocity for the longitudinal and transversal modes, respectively.

To continue with the analysis, an infinitely plate extended in the  $x$  and  $y$  directions will be assumed. Furthermore, it is considered that the wave propagates in the  $x$  direction, the fields are uniform in the  $y$  direction and boundary conditions at  $z = -h/2$  and  $z = +h/2$ , where  $h$  is the thickness of the plate, are considered traction free.

Assuming that the particle displacement is zero in the  $y$  direction ( $u_y = 0$ ) and the only rotation is about the  $y$  axis ( $H_x = H_z = 0$ ) in Equation (4), the Lamb wave equations are obtained [19,20].

$$\mu \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2} \right) u_x + (\lambda + \mu) \frac{\partial}{\partial x} \left( \frac{\partial u_x}{\partial x} + \frac{\partial u_z}{\partial z} \right) = \rho \left( \frac{\partial^2 u_x}{\partial t^2} \right) \quad (8)$$

$$\mu \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2} \right) u_z + (\lambda + \mu) \frac{\partial}{\partial z} \left( \frac{\partial u_x}{\partial x} + \frac{\partial u_z}{\partial z} \right) = \rho \left( \frac{\partial^2 u_z}{\partial t^2} \right) \quad (9)$$

Applying the restriction in the frontiers:

$$\mu \left( \frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x} \right) \Big|_{z=\pm \frac{h}{2}} = 0 \quad (10)$$

$$\lambda \frac{\partial u_x}{\partial x} + (\lambda + 2\mu) \frac{\partial u_z}{\partial z} \Big|_{z=\pm \frac{h}{2}} = 0 \quad (11)$$

Focusing on Lamb waves, they are composed of two waves (one longitudinal and one transversal) traveling at different angles  $\theta_L$  and  $\theta_T$ , where the first one represents the longitudinal angle and the second one the transversal one. Therefore, the wave number  $k$  is related to the component of the waves that propagates in the  $x$  direction at velocity  $c_p$ :

$$k_L \cos \theta_L = k_T \cos \theta_T = k = \frac{2\pi f}{c_p} = \frac{\omega}{c_p}, \quad (12)$$

where  $k_L = 2\pi f/c_L$ ,  $k_T = 2\pi f/c_T$ ,  $\omega$  is the angular velocity,  $c_L$  is the longitudinal component of the velocity and  $c_T$  is the transversal component of the velocity.

On the other hand, the displacement of each independent wave in the  $z$  axis can be obtained using Equations (13) and (14).

$$\alpha_L = k_L \sin \theta_L = k_L \sqrt{1 - \cos^2 \theta_L} = \sqrt{k_L^2 - k^2} = \sqrt{\frac{\omega^2}{c_L^2} - k^2} = \omega \sqrt{\frac{1}{c_L^2} - \frac{1}{c_p^2}} \quad (13)$$

$$\alpha_T = k_T \sin \theta_T = k_T \sqrt{1 - \cos^2 \theta_T} = \sqrt{k_T^2 - k^2} = \sqrt{\frac{\omega^2}{c_T^2} - k^2} = \omega \sqrt{\frac{1}{c_T^2} - \frac{1}{c_p^2}} \quad (14)$$



where  $\alpha_L$  and  $\alpha_T$  represent the longitudinal and transversal displacements of the wave, and  $\theta_L$  and  $\theta_T$  are the angles related to these displacements. Taking into account that the wave is reflected in the surfaces, applying the boundary conditions and simplifying the equations, the dispersion equation of the Lamb modes is obtained. Equation (15) refers to the symmetric modes and Equation (16) refers to the asymmetric ones.

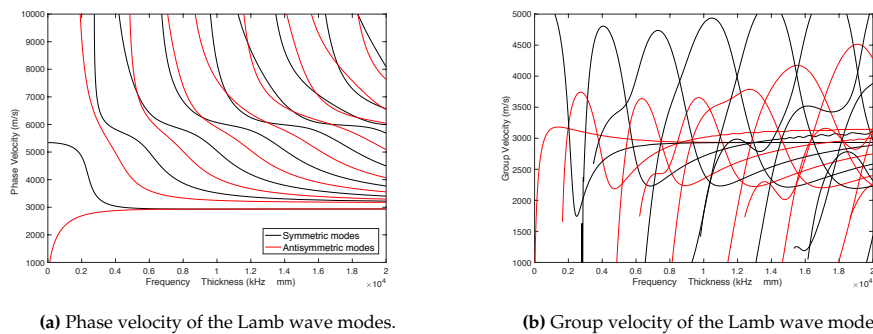
$$4k^2\alpha_L\alpha_T \sin\left(\frac{\alpha_L h}{2}\right) \cos\left(\frac{\alpha_T h}{2}\right) + \sin\left(\frac{\alpha_L h}{2}\right) \cos\left(\frac{\alpha_T h}{2}\right) (\alpha_T^2 - k^2)^2 = 0 \quad (15)$$

$$4k^2\alpha_L\alpha_T \cos\left(\frac{\alpha_L h}{2}\right) \sin\left(\frac{\alpha_T h}{2}\right) + \cos\left(\frac{\alpha_L h}{2}\right) \sin\left(\frac{\alpha_T h}{2}\right) (\alpha_T^2 - k^2)^2 = 0 \quad (16)$$

From the previous equations, it can be figured out that there exists a relation between the excited frequency  $f$ , the thickness of the pipe  $z$  and the phase velocity  $c_p$ . More specifically, each mode will move at different  $c_p$  depending on the other above-mentioned parameters. A similar relation can be obtained using the group velocity  $c_g$ , which is defined in Equation (17).

$$c_g = \frac{\partial\omega}{\partial k} \quad (17)$$

Graphs showed in Figure 2 were obtained by means of the previous equations for different values of the product frequency by thickness, where phase velocity is represented in Figure 2a and group velocity in Figure 2b. As it can be observed, the relation between the velocities and the frequency is non-linear, so there exists dispersion in the Lamb wave propagation.



(a) Phase velocity of the Lamb wave modes.

(b) Group velocity of the Lamb wave modes.

**Figure 2.** Phase velocity and group velocity depending on the product frequency by thickness.

In Figure 2, black lines increasingly represent symmetric modes from left to right (S0, S1, S2...), while red lines represent in the same way antisymmetric modes (A0, A1, A2...). These graphs correspond to a steel pipe with the following parameters: Young's modulus  $E = 200 \cdot 10^9$  N/m<sup>2</sup>, Poisson's ratio  $\nu = 0.3$  and density  $\rho = 7700$  kg/m<sup>3</sup>.

Now the methodology followed to generate and receive signals in the pipeline will be described. The transducer consists on a meander-line-coil which generates two signals per loop in the test piece (one per meander). These waves will be characterized by the wavelength which depends on the separation of the meanders.

The following equations are valid for one mode and then the same procedure will be applied iteratively for all the modes which appear at a set frequency. Thus, wave equation is set depending on the group and phase velocities. Considering  $f$  as the excited frequency, the transmitted signal  $s(x, t)$  will be generated according to Equation (18).

$$s(x, t) = \sin \left( 2\pi f \left( \frac{x}{c_p} \pm t \right) + \phi \right), \quad (18)$$

where  $\phi$  is a phase term that controls the phase of the transmitted signal.

In a real case, the transmitted signal includes an envelope  $w(t)$  that generates the transmitted wave packet  $p(x, t)$ . This envelope limits the transmission time, and allows controlling the length of the transmitted pulse. Typically, the length of this envelope is described in function of  $C$ , the number of cycles included in the wave packet. That is, the length of  $w(t)$  will be  $C/f$ , where  $1/f$  is the time period corresponding to the excited frequency. This envelope will travel at an average velocity of  $c_g$ , and in general its shape will change with the distance due to dispersion effects.

The inclusion of the time envelope  $w(t)$  in the transmitted signal causes the signal to be wider in the frequency domain. Thus, the number of cycles  $C$  is related to the transmitted bandwidth, so that the lower the number of cycles, the wider the transmitted bandwidth. For instance, if a signal with  $f = 300$  kHz and  $C = 4$  cycles is transmitted, the 3 dB transmission bandwidth ranges from 252 kHz to 345 kHz. This must be taken into consideration, since the phase velocity at these frequencies might not vary linearly, causing dispersion in the wave packet.

Therefore, once the envelope is considered, the transmitted wave packet  $p(x, t)$  will be expressed using Equation (19).

$$p(x, t) = s(x, t) \cdot w \left( \frac{x}{c_g} \pm t \right) = \sin \left( 2\pi f \left( \frac{x}{c_p} \pm t \right) + \phi \right) \cdot \hat{w} \left( \frac{x}{c_g} \pm t \right) \quad (19)$$

Please note there that instead of using the transmitted envelope  $w(t)$  we are using  $\hat{w}(t)$ , which changes its shape in function of the distance due to dispersion effects.

It is necessary to consider that under EMAT technology the excitation signal is generated in a set of  $N$  loops of a coil, separated by a distance  $L$ , which will generate the propagation wave  $y(x, t)$  using Equation (20).

$$y(x, t) = \sum_{m=1}^{2N} (-1)^m p(x - m \cdot L/2, t) \quad (20)$$

Please note here that each loop generates two signals (one per meander), and that the sign of their contribution to the propagation wave  $y(x, t)$  is included in the term  $(-1)^m$ . Besides, the measure is sensed at a distance  $D$ , in another set of  $N$  loops separated by a distance  $L$ . Therefore, the received signal  $z(t)$  will be expressed using Equation (21).

$$z(t) = \sum_{n=1}^{2N} (-1)^n y(D - n \cdot L/2, t) \quad (21)$$

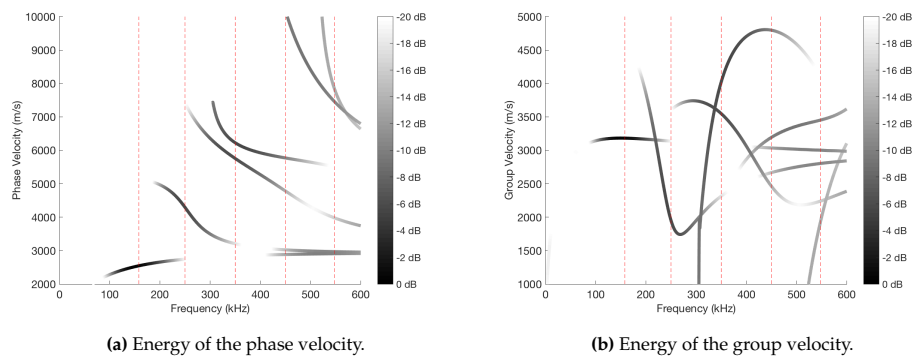
Again, the sign of each meander is represented by the term  $(-1)^n$ . Going back to the previous equations, the total signal sensed from each mode  $z(t)$  is obtained.

$$z(t) = \sum_{m,n=1}^{2N} (-1)^{m+n} p(D - (m+n) \cdot L/2, t) \quad (22)$$

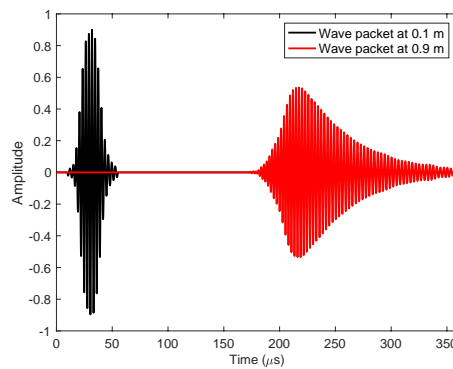
$$z(t) = \sum_{m,n=1}^{2N} (-1)^{m+n} \sin \left( 2\pi f \left( \frac{D - (m+n) \cdot L/2}{c_p} \pm t \right) + \phi \right) \cdot \hat{w} \left( \frac{D - (m+n) \cdot L/2}{c_g} \pm t \right) \quad (23)$$

The signal received from each mode  $z(t)$  has different values of  $c_p$  and  $c_g$ , as it was concluded from Figure 2. Thus, each mode arrives at the receiver with different amplitude and envelope, depending on the attenuation of each mode and the difference of phase when the signal is received in the coil. Therefore, the amount of energy of the received signal can vary at different frequencies.

In order to find out more about the behavior of the modes in a set frequency range, a frequency sweep was made between 0 and 600 kHz with one coil and  $C = 4$  cycles per wave packet. Figure 3 shows the phase velocity (Figure 3a) and group velocity (Figure 3b), where black color means the energy is maximum at that frequency. It is important to indicate that dispersion has been taken into account to carry out the experiments, since the signal has been decomposed with the envelope window  $\hat{w}(t)$  through the Fourier Transform. Thus, the velocities and delays of the different frequencies which are part of the same pulse have been considered. As an example of the effects of dispersion over the wave packet, Figure 4 shows the dispersion suffered by the wave packet when traveling 0.8 m in the pipe (S0 mode,  $f = 300$  kHz,  $C = 4$  cycles).



**Figure 3.** Phase velocity and group velocity of the different modes represented according to the energy in a range of frequencies.



**Figure 4.** Dispersion suffered by the wave packet when traveling 0.8 m in the pipe (S0 mode,  $f = 300$  kHz,  $C = 4$  cycles).

The coil used in the experiments has the following parameters: distance between loops  $L = 16.26$  mm and  $N = 3$  loops. It can be observed in the graphs that the maximum energy appears in  $f = 158$  kHz and mode A0. However, there exists a certain periodicity in the energy of the received signals. It implies that the same coil could be used to excite other frequencies, even if it has been designed to get the maximum energy in a set frequency. In fact, different frequencies will be excited in the experiments. Specifically, the frequencies indicated with red dashed lines in Figure 3 will be used, because the energy and the excited modes are different in each of them.

## 2.2. Hardware Description

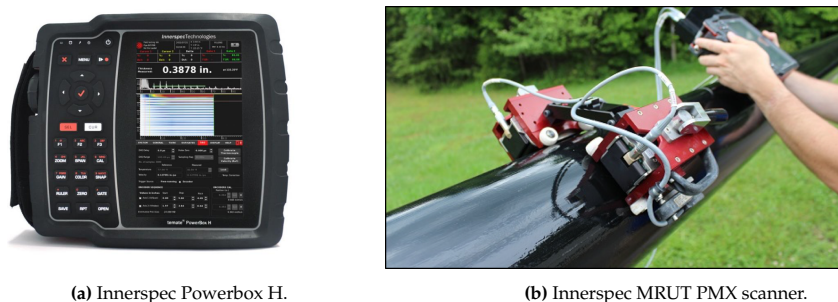
The technology (Innerspec PowerBox H and the MRUT PMX scanner) and the pipe mock-ups needed to perform the empirical validation of the modeling results were provided by Innerspec Technologies S.L [21], a company which provides NDT solutions using EMAT technology.

The inspection instrument used is the Innerspec PowerBox H, a hand-held battery operated instrument. It is designed for ultrasonic applications that require very high voltages and/or long bursts of energy such as non-contact techniques (EMAT, Air-Coupled) and inspection of highly-attenuating materials. The instrument is capable of generating up to 1200V or 8kW of peak power at speeds of up to 300 Hz.

Guided waves can be used to cover distances ranging from a few millimeters to tens of meters. The two most common techniques for in-service inspections with guided waves are Long Range UT (LRUT) and Medium Range UT (MRUT). All the results showcased within this manuscript were obtained with the MRUT PMX scanner, which is used in both attenuation and reflection mode to cover shorter distances (0.1–5 m). The sensors are mounted on scanners to inspect long stretches of pipes or tanks. It typically works with frequencies from 100 kHz to 1 MHz, and can detect small pits ( $\times 10$  more sensitivity than using LRUT).

The MRUT PMX scanner allows to scan axially with a single or double sensor on the pipe to measure attenuation and/or velocity changes in the signal due to corrosion, cracks or other defects around the circumference of the pipe. It is ideal for quick inspections of exposed pipe at speeds up to 150 mm/s (6 in/s).

Figure 5 shows the hardware equipment used in the manuscript.



(a) Innerspec Powerbox H.

(b) Innerspec MRUT PMX scanner.

Figure 5. Measuring equipment used in the experiments.

## 3. Effects of the Defect Over the Lamb Waves

The modeling of the pipe by means of the ultrasonic waves is a non-trivial problem. The changing shape of the defects makes difficult to draw general conclusions about the relation between the defect and the received signals. The distortion caused by the defects over the different modes strongly varies with the shape of the defect [12,13]. For instance, the amplitude of the signal, the time of arrival (group velocity  $c_g$ ) and the phase velocity  $c_p$  of the wrap-around signal vary with the dimensions of the defect.

To study the relation between these parameters and the shape of the defects, the Finite Element Method (FEM) included in the Partial Differential Equations Toolbox of *Matlab* has been used. A database of 418 defects has been generated using this simulation tool. The defects have been characterized with three parameters: length ( $l$ ), depth ( $d$ ) and slope ( $s$ ). Figure 6 depicts the defect dimensions using the three aforementioned parameters. If  $s > l$  the defect is discarded. In the case studied, the thickness of the pipe is  $z = 9.27$  mm. Table 1 shows the range of values that the parameters can take.

Likewise, the simulated coil has the following parameters: distance between loops  $L = 16.26$  mm,  $N = 3$  loops and distance to the receiver  $D = 0.7$  m. Besides, each pulse of the signal contains 4 cycles in all the experiments.

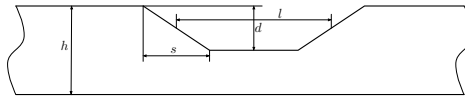


Figure 6. Model of the simulated defects.

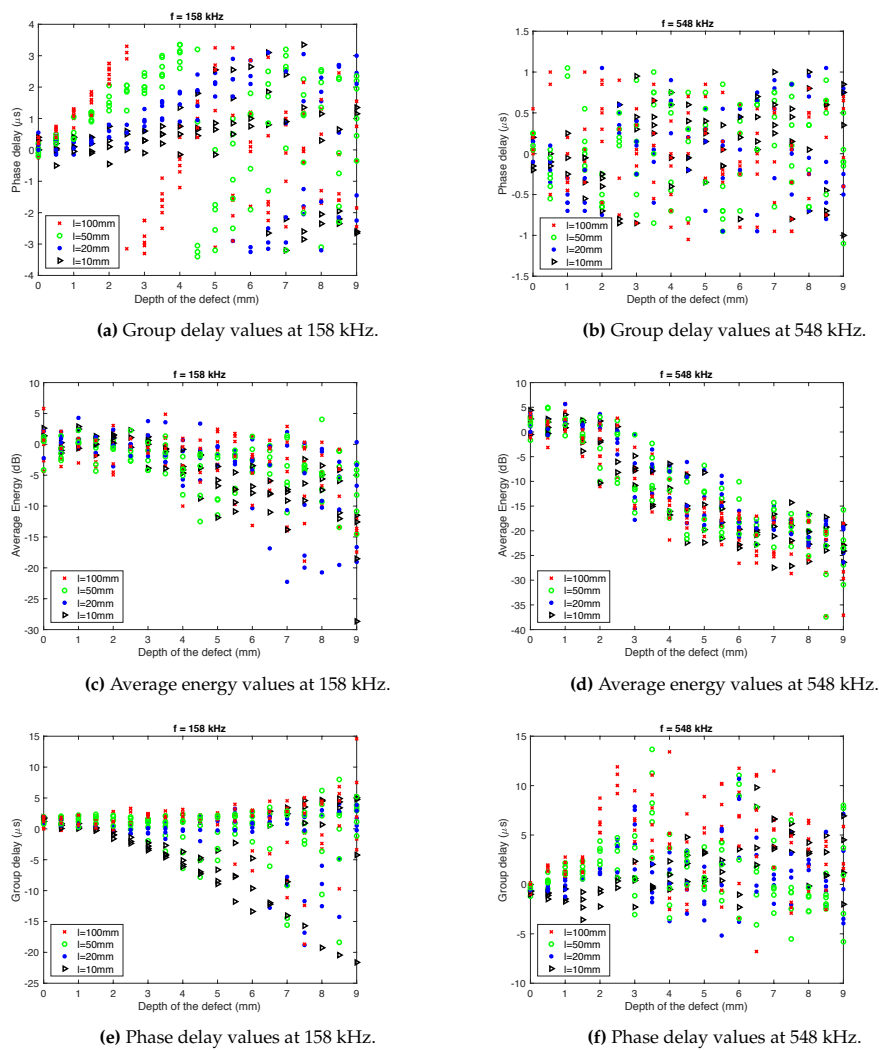


Figure 7. Feature values depending on the depth of the defect.

Figure 7 shows how the depth of the defect affects to the values of group delay (Figure 7a,b), average energy (Figure 7c,d) and phase delay (Figure 7e,f), at two frequencies:  $f = \{158, 548\}$  kHz. For simplicity, the width of the defect was not modeled, that is to say, that dimension of the pipe was not considered. Each case has a few points since defects with different slope have been considered.

**Table 1.** Range of values for the different parameters of the defects.

Parameter	Number of Possible Values	Values
Frequency (kHz)	5	158, 250, 350, 450, 548
Length (mm)	4	10, 20, 50, 100
Depth (mm)	19	0, 0.5, 1, ..., 9
Slope (mm)	7	1, 3, 5, 10, 50, 100

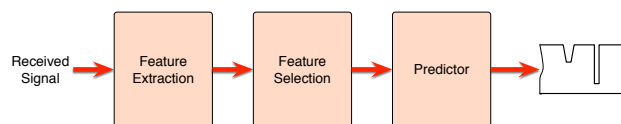
As it can be shown in the graphs, there exists a tendency in some of the considered features. Focusing on the average energy, calculated in  $f = 548$  kHz, it is clear to see that, as the length of the defect increases, the average energy decreases drastically, especially between 2 and 6 mm of defect depth. This is exactly what would be expected when there exists a leak in the pipeline and the energy of the signal is scattered through it.

In the case of the group delay, considering the measurement taken at  $f = 158$  kHz, it can be observed that the signal tends to be delayed (positive delay) when the defect increases. This does not happen when the length of the defect is very small ( $l = 10\text{--}20$  mm), since the signal arrives earlier than in the non-defect case. In any event, the aim of this modeling work was to evaluate whether these features contain useful information to tackle the problem addressed.

There exists a difficulty of reaching a conclusion about the relation between the calculated features and the profile of the pipe. Because of that, it is necessary to apply advanced techniques which bring more information about what are the best features or how they should be mixed.

#### 4. Smart Sound Processing (SSP) for Defect Sizing

It is necessary to apply SSP methods to solve the defect sizing problem in pipes that is being coped in the current manuscript. This type of methods usually follows the process described in Figure 8. First of all, it is important to extract useful information from the signals on the form of features. Once this is done, the next step is to select the ones that best work to solve the problem at hand. Finally, a predictor will construct a model capable of predicting the solution in an unknown case, such as a new pipeline.



**Figure 8.** Scheme of an SSP system.

It is important to extract useful information from the received signal in order to be capable of detecting and sizing defects present in the pipes under inspection. With this purpose, different features were elicited:

- Maximum Amplitude (dB). This measure indicates the value of the maximum peak received from the signal. It is determined by looking for the value of the maximum peak around the expected point, which is the position of the maximum of the reference signal form in case of absence of any defects,  $t_0$ ).
- Phase Delay ( $\mu\text{s}$ ). This measure represents the time taken between the pulse shipment and its reception at the same pipe location. It is determined measuring the time difference between

the position of the maximum of the reference signal (signal without defect)  $t_0$  and its closest maximum in the sensed signal. There may be considerable uncertainty in this feature if the delay is higher than half the period, since the nearest peak becomes the maximum of the signal.

- Average Energy (dB). It represents the average energy of the pulse. The interval considered for its calculation started 30  $\mu$ s before  $t_0$  and ends up 30  $\mu$ s after it.
- Group Delay ( $\mu$ s). In order to calculate this measure, the centroid of the average energy of the pulse has been considered. It has been estimated using the centroid of the pulse around the expected maximum ( $\hat{t}_g$ ), with Equation (24).

$$\hat{t}_g = \frac{\sum_{t=t_0-3 \cdot 10^{-5}}^{t_0+3 \cdot 10^{-5}} tz(t)^2}{\sum_{t=t_0-3 \cdot 10^{-5}}^{t_0+3 \cdot 10^{-5}} z(t)^2} \quad (24)$$

All these features have been extracted directly from the received signal at  $D = 0.7$  m. Two additional features, extracted from the reflected signal, were included to study the importance of the analysis of the echos in the sizing problem.

- Maximum Amplitude of the echo (dB).
- Average Energy of the echoes (dB).

It is remarkable the fact that the reflected echoes are not always depicted in the gathered signals, since the position of their contribution depends on the relative position of the defect in the pipeline with respect to the EMAT actuator. Therefore, in some cases the echo is overlapped with the transmitted signal, and cannot be clearly identified. In the simulations two scenarios would be modeled: on the one hand, the case where the reflected echoes are present and on the other hand, the case where they are not, to study the importance of these echo dependent features in the performance of the sizing estimator.

In the problem at hand, 5 excitation frequencies were used. In total, 30 features were extracted taking into account that 6 features were obtained for each excitation frequency. Some of them will work better than others, so it is important to select the best ones and reduce the total amount of them, in order to properly estimate the size of the defect.

To select the features which better work in this experiment, a feature selection process was applied through evolutionary techniques. Evolutionary Algorithms (EAs) are inspired in natural evolution laws and allow to find the optimum solution from the solutions (denoted individuals) obtained in previous iterations [22]. In this paper, a tailored EA has been applied, searching for the best subset of features and trying to minimize the Root Mean Square Error (RMSE) of a Least Squares Linear Discriminant (LSLD). The use of more complex prediction methods has been avoided in the feature selection process, since the EA requires training and testing the predictor a large number of times. The considered constraints are to limit the number of frequencies used as well as the number of features selected.

The EA, which is described schematically in Figure 9, is composed of several steps:

1. A population of  $N_p$  individuals is generated. Each solution consists of a binary vector with a length equal to the total number of features. Thus, ones indicate the features which are selected in the individual, while zeros indicate the features which remain outside.
2. The candidates of the population are restricted to the considered constraints. All of them are modified in order to randomly change the value of some bits until one or the two constraints are fulfilled, depending on the case.
3. A LSLD is designed with the subset of features of each candidate solution. The RMSE of the defect depth is calculated, which is the fitness function in this experiment. With this value, the population is ranked, keeping the best individuals in the top of the ranking.
4. After that, a selection process is applied, which consists in keeping the best 10% of the solutions, removing the remaining ones.

5. The removed solutions (90% of the population) are regenerated by crossover between the best candidates.
6. Mutations are applied to the new population. With this step, 1% of the bits is changed. Furthermore, this step is not applied to the best solution in order to ensure the convergence of the algorithm.
7. This process is repeated from step 2 during  $N_g$  generations. The final best solution will be the best solution obtained in the last iteration.

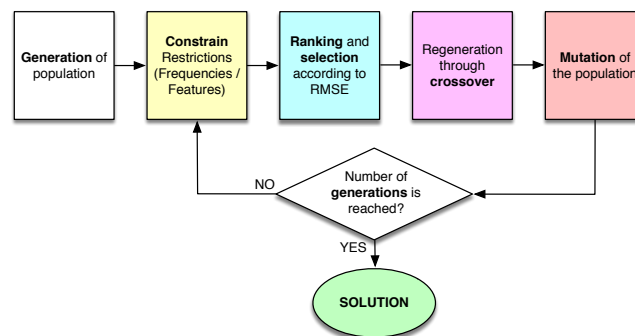


Figure 9. Scheme of the evolutionary algorithm applied in the experiments.

Sometimes the EAs do not reach a high convergency. In order to improve it, an elimination tournament of small EAs has been implemented. It consists in joining the winners of the EAs in pairs during several rounds ( $N_r = 6$  in our case), until the best individual reaches the end and, consequently, it becomes the best solution to the problem. 32 small EAs were considered with a population of 100 individuals and 8 generations each, except for the last EA, where 16 generations are configured for convergency [22].

Once the best features have been selected for this specific problem, a non-linear predictor needs to be applied to get the final profile of the pipeline and to know the performance of the developed model. Neural Networks were applied, specifically the Multi Layer Perceptron (MLP) [23]. A perceptron is a neuron with a set of adjustable weights and an activation function by steps [24]. Levenberg-Marquardt algorithm, a method where the minimization function is a sum of quadratic terms [25], was applied for training purposes. The number of hidden neurons was a parameter in the experiments.

A  $k$ -fold cross validation has been applied in the generated database, being  $k = 12$  [26]. This method allows to divide the database in  $k$  groups so that the full process is repeated  $k$  times, using one group as test subset and the remaining  $k - 1$  groups as training subset. The advantage of this method is that the obtained results are more generalizable. It has been applied in both the feature selection process and the training of the neural network-based predictor.

## 5. Results

The research work showcased in this manuscript was carried out using both real and simulated measurements. This section of the paper will be opened discussing the results obtained during the simulation experiments.

### 5.1. Defect Sizing in Simulated Pipelines

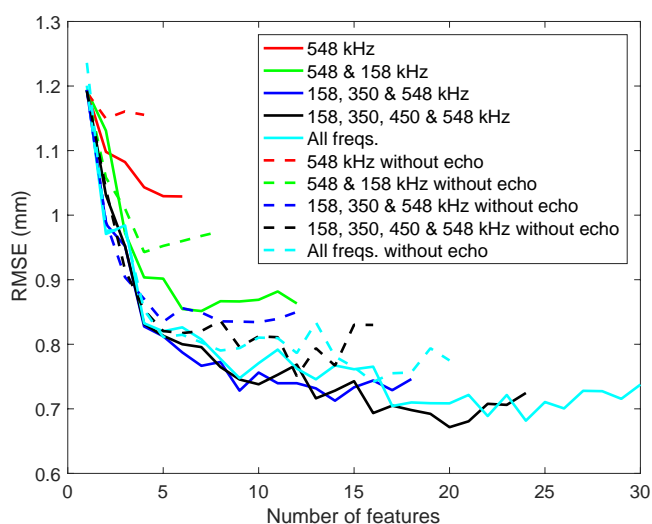
An experiment was developed using the synthetic database described in Table 1. In this case, the RMSE was computed to evaluate the performance of the predictors. The objective is to know how well the received signal is estimated at different frequencies. The  $k$ -fold cross validation described in



Section 4 has been applied in this experiment, so all the results presented here have been obtained with this method.

It is important to indicate that noise from the real measurements was introduced to the signal, the amplitude and the temporal resolution in order to make the experiments as real as possible. A real pipe free from defects was taken in order to study the measuring system tolerance to noise. It was obtained an SNR of 32.93 dB, a variation in time with a standard deviation of 2.1434 samples at 10 MHz (0.2  $\mu$ s) and a variation in scale of 0.80 dB.

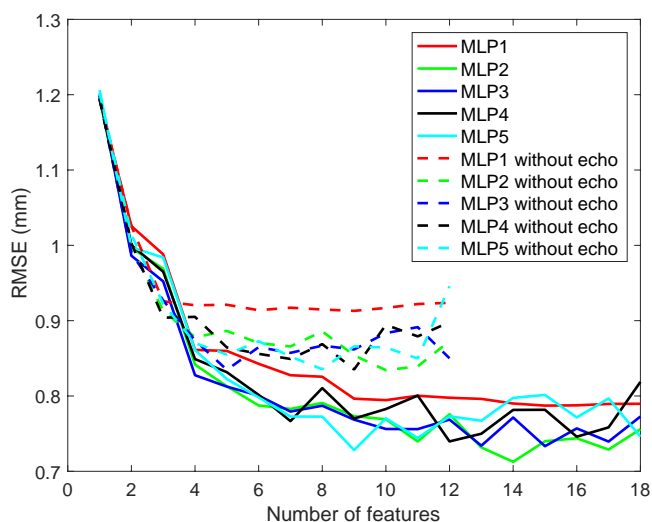
As stated above, the relevance of extracting features from echoes related to reflections has been also studied. Figure 10 shows the RMSE obtained by the different predictors, depending on the number of selected features. The different curves show the performance depending on the combination of frequencies considered, in the case of only using features from the wrap-around echo, and in the case of including features from the reflected echo. The combination of frequencies and the classifier used in each case are those that provide the best results in terms of RMSE. It can be seen that the results improve (less RMSE) as the number of evaluated frequencies is increased. It happened especially until 158, 350 and 548 kHz frequencies were evaluated. The results slightly improve when the number of frequencies studied is increased, so there would be no need to make the system more complex to reduce a few millimeters the final rate.



**Figure 10.** RMSE obtained with neural networks predictor depending on the number of features selected at different frequencies.

In terms of the reflected echo features, if they are not used (dashed lines) the tendency of the graphs is similar to the previous one. In general, the results are not significantly different from those using echo features, especially when all frequencies are used. Because of that, the features extracted from the echo signals are not essential to evaluate the problem and it is possible to get a good result in case the return of the signal cannot be evaluated.

With regard to the used predictors, Figure 11 shows the RMSE depending on the number of features employed and the number of neurons configured in the MLP (1, 2, 3, 4 and 5). Although the results are good using just one neuron, they can be improved by using two. However, if the number of neurons is further increased (3–5), the result is not much better than before.



**Figure 11.** RMSE obtained depending on the number of selected features considering different number of neurons in the MLPs.

In all the cases above it is clear that as the number of features is higher, the results improve. It happens especially in the range between 1 and 5 features, where the error falls considerably in Figures 10 and 11.

Now the features selected at different frequencies from the proposed ones are going to be studied. With this purpose the manuscript will focus on the case of mixing eight features from three frequencies: 158, 350 and 548 kHz (dark blue line). The results do not improve substantially when more frequencies are added or when the number of features is increased. In Tables 2 and 3 the ratios of selection of the features are shown. The first one includes the case of considering the reflected echo features and the second the case of not considering them. The RMSE associated to these two cases are 0.79 mm and 0.87 mm, with and without reflected echo features, respectively.

**Table 2.** Ratios of selection of the features, including echo features.

Feature	Frequency (kHz)		
	158	350	548
Maximum Amplitude	0%	92%	0%
Phase Delay	0%	0%	0%
Average Energy	100%	100%	100%
Group Delay	0%	0%	58%
Maximum Amplitude of Echo	8%	50%	0%
Average Energy of Echos	92%	100%	100%

**Table 3.** Ratios of selection of the features without echo features.

Feature	Frequency (kHz)		
	158	350	548
Maximum Amplitude	100%	100%	92%
Phase Delay	0%	0%	0%
Average Energy	100%	100%	100%
Group Delay	8%	100%	100%

The results show that the most important feature is the average energy of the received pulse of the signal, since it is selected in all the frequencies considered. It confirms that the energy is the most affected parameter when the signal travels through a defect. This reasoning applies for both the wrap around echo and the reflected echo.

Other features which work properly are the maximum amplitude and the group delay, but mainly when the features extracted from the reflected echo are not taken into consideration. It means that they are good features, but not as much as the energy from the reflected echo. Furthermore, features like phase delay are not relevant for the study, because they are not selected to get a better result. In fact, this feature has 0% of selection in all cases of study. This is caused by the problems with the uncertainty in measuring this parameter.

These results demonstrate again that the reflected echo features are not essential for the problem. Using them the model fits just 0.1 mm better with the target.

### 5.2. Defect Sizing in a Real Pipeline

Now the results obtained during the experimental trials in a pipe mock-up are presented. Data has been collected using MRUT PMX scanner on a pipe, as mentioned before. The inspection was performed moving the sensor axially (sending the waves circumferentially) on a pipe which includes three flat defects with different depths. An image of the pipe is shown in Figure 12, while its specifications are shown in Table 4.



Figure 12. Image of the real pipe.

Table 4. Specifications of the real pipe.

Parameter	Value
Material	Structural Steel S355NH
Thickness (mm)	9.27
Depth of Defect 1 (mm)	6.18
Depth of Defect 2 (mm)	4.63
Depth of Defect 3 (mm)	1.85

The inspections were carried out at different frequencies. The distortions introduced by the defect over the analyzed signals are shown in Figure 13. Axial scans sending the waves circumferentially have been carried out every millimeter of the pipeline, and the main wave packages of the wrap-around echos have been stored. Different behaviors are depicted depending on the depth of the defects and the excited frequency,  $f = 158$  kHz (Figure 13a) and  $f = 548$  kHz (Figure 13b).

The graphs included in Figure 13 show that the defects change the amplitude and the phase of the received signals. Furthermore, these distortions are different depending on the size of the defects and the excited frequency  $f$ .

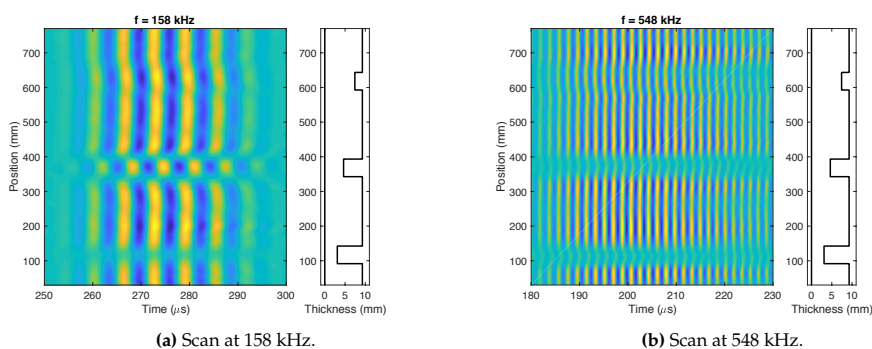


Figure 13. Axial scan of the real pipeline at two frequencies.

The profile of the real pipe will be modeled using the features explained above. In this experiment, the predictor trained was used along with the synthetic database and it was also tested with the signals sensed in the real laboratory trials. Therefore, it has not been necessary to apply the cross validation technique, as the training and test subsets were clearly defined. Three cases were developed: considering 4 features at 158 kHz, 4 features at 548 kHz and mixing all of them. The 2 features from the reflected echo (maximum amplitude of the reflected echo and average energy of the reflected echo) have not been considered because the return from the signal was sometimes overlapped with the excitation pulse, so it was not possible to extract any information from it.

In Figure 14 and Table 5 the results of the described experiments are shown. The model that better fits with the pipeline is the predictor which considers eight features and the two frequencies. It is the one which better distinguishes between defect and non-defect areas, since the estimation of the defect depth is very close to zero in non-defect areas. In fact, the RMSE is the best of the three predictors (1.48 mm). When mixing the information from the two frequencies the results are quite improved. With the other models the estimation was not so good, especially when there were no defects in the profile.

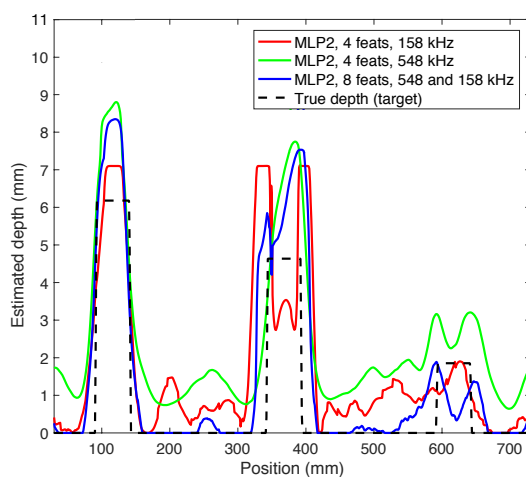


Figure 14. Estimation of the real pipeline model.

**Table 5.** Results of real database.

Predictor	Number of feats	Frequency (kHz)	RMSE (mm)
MLP 2 neurons	4	158	1.84
	4	548	1.80
	8	158, 548	1.48

## 6. Discussion

In this work the defect detection and sizing in steel pipelines have been studied. With this purpose, Lamb waves have been generated with EMAT-based techniques. These have been analyzed with SSP methods, in order to try to model the pipeline. After the experiments described above have been carried out, the following conclusions have been drawn:

- The shape of the defect causes differences in the received signal. It is not feasible to obtain an analytical solution for all the cases.
- The extracted features are useful for the pipeline sizing problem, since the results are good in terms of RMSE. The average energy and the maximum amplitude from the signals are particularly relevant for the study.
- It is important to excite the waves at several frequencies because the behavior and velocity of the Lamb modes is totally different depending on this parameter.
- Related to the predictors, a large number of neurons in the MLPs is not required. When it is increased above two or three neurons, the results do not improve significantly.

In the future it would be possible to increase the amount and variety of defects in real pipelines. Besides, more features could be applied, such as some from the signal in successive wrap-arounds.

## 7. Conclusions

Pipeline inspection problem can be approached in many different ways. Lamb wave generation through EMAT actuators proves to be a very effective and useful one. However, the amount of information provided by the wrap-around signals needs to be processed by advanced techniques, such as smart sound processing algorithms. Thanks to them, it is feasible to get good estimation results of the pipeline defects, in both real and simulated signals. In the manuscript it has been demonstrated the importance of applying a multi-frequency study for defect sizing problem, the relevance of some features from the signals (e.g., energy and amplitude) and the absence of the need to greatly increase the complexity of the classifiers to get a good estimation in the problem at hand.

**Acknowledgments:** This work has been funded by Innerspec Technologies Europe S.L through the “Chair of modeling and processing of ultrasonic signals” (CATEDRA2007-001), and by the Spanish Ministry of Economy and Competitiveness/FEDER under Project TEC2015-67387-C4-4-R.

**Author Contributions:** J.G., R.G. and M.R. designed and performed the experiments, as well as analyzed the results; J.G. and R.G. wrote the paper; A.R., J.A.J. and V.G. provided the hardware equipment and the real pipeline measurements.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Blitz, J.; Simpson, G. *Ultrasonic Methods of Non-Destructive Testing*; Springer Science & Business Media: New York, NY, USA, 1995; Volume 2.
2. Silk, M.G. *Ultrasonic Transducers for Nondestructive Testing*; CRC Press: Boca Ratón, FL, USA, 1984.
3. Vanaei, H.; Eslami, A.; Egbewande, A. A review on pipeline corrosion, in-line inspection (ILI), and corrosion growth rate models. *Int. J. Press. Vessels Piping* **2017**, *149*, 43–54.
4. Green, R.E. Non-contact ultrasonic techniques. *Ultrasonics* **2004**, *42*, 9–16.

5. Zhai, G.; Jiang, T.; Kang, L. Analysis of multiple wavelengths of Lamb waves generated by meander-line coil EMATs. *Ultrasonics* **2014**, *54*, 632–636.
6. Park, J.H.; Kim, D.K.; Kim, H.j.; Song, S.J.; Cho, S.H. Development of EMA transducer for inspection of pipelines. *J. Mech. Sci. Technol.* **2017**, *31*, 5209–5218.
7. Salzbürger, H.J.; Niese, F.; Dobmann, G. EMAT pipe inspection with guided waves. *Welding World* **2012**, *56*, 35–43.
8. Andruschak, N.; Saletes, I.; Filleter, T.; Sinclair, A. An NDT guided wave technique for the identification of corrosion defects at support locations. *NDT E Int.* **2015**, *75*, 72–79.
9. Clough, M.; Fleming, M.; Dixon, S. Circumferential guided wave EMAT system for pipeline screening using shear horizontal ultrasound. *NDT E Int.* **2017**, *86*, 20–27.
10. Wang, S.; Huang, S.; Zhao, W.; Wei, Z. 3D modeling of circumferential SH guided waves in pipeline for axial cracking detection in ILI tools. *Ultrasonics* **2015**, *56*, 325–331.
11. Singh, R. *Pipeline Integrity: Management and Risk Evaluation*; Gulf Professional Publishing: Houston, TX, USA, 2017.
12. Demma, A. The Interaction of Guided Waves With Discontinuities in Structures. Ph.D. Thesis, University of London, London, UK, 2003.
13. Cobb, A.C.; Fisher, J.L. Flaw depth sizing using guided waves. In *AIP Conference Proceedings*; AIP Publishing: Melville, NY, USA, 2016; Volume 1706, p. 030013.
14. Nurmaliya; Nakamura, N.; Ogi, H.; Hirao, M. EMAT pipe inspection technique using higher mode torsional guided wave T (0, 2). *NDT E Int.* **2017**, *87*, 78–84.
15. Hirao, M.; Ogi, H. An SH-wave EMAT technique for gas pipeline inspection. *NDT E Int.* **1999**, *32*, 127–132.
16. Layouni, M.; Hamdi, M.S.; Tahar, S. Detection and sizing of metal-loss defects in oil and gas pipelines using pattern-adapted wavelets and machine learning. *Appl. Soft Comput.* **2017**, *52*, 247–261.
17. Mohamed, A.; Hamdi, M.S.; Tahar, S. A machine learning approach for big data in oil and gas pipelines. In Proceedings of the 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud), Rome, Italy, 24–26 August 2015; pp. 585–590.
18. Mohamed, A.; Hamdi, M.S.; Tahar, S. An adaptive neuro-fuzzy inference system-based approach for oil and gas pipeline defect depth estimation. In Proceedings of the SAI Intelligent Systems Conference (IntelliSys), London, UK, 10–11 November 2015; pp. 35–42.
19. Luangvilai, K. *Attenuation of Ultrasonic Lamb Waves With Applications To Material Characterization and Condition Monitoring*; Georgia Institute of Technology: Atlanta, GA, USA, 2007.
20. Su, Z.; Ye, L. *Identification of Damage Using Lamb Waves: From Fundamentals To Applications*; Springer Science & Business Media: New York, NY, USA, 2009; Volume 48.
21. Garcia, V.; Boyero, C.; Jimenez, J.A. Corrosion detection under pipe supports using EMAT Medium Range Guided Waves. In Proceedings of the 19th World Conference on Non-Destructive Testing, Munich, Germany, 13–17 June 2016; pp. 1–9.
22. Gil-Pita, R.; García-Gomez, J.; Bautista-Durán, M.; Combarro, E.; Cocana-Fernandez, A. Evolved frequency log-energy coefficients for voice activity detection in hearing aids. In Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 9–12 July 2017; pp. 1–6.
23. Weisz, L. Pattern Recognition Statistical Structural And Neural Approaches. *Pattern Recogn.* **2016**, *1*, 2.
24. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386.
25. Hagan, M.T.; Menhaj, M.B. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Netw.* **1994**, *5*, 989–993.
26. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-validation. In *Encyclopedia of Database Systems*; Springer: New York, NY, USA, 2009; pp. 532–538.



## Part III

# Conclusions and future lines

# Chapter 7

## Conclusions

On the basis of the results obtained along the experiments of this thesis, the main contributions and conclusions are summarized and analyzed in this chapter. The more general contributions are summarized in Section 7.1, while the specifics of each application are explained in Sections 7.1.1 (Violence Situation Detection), 7.1.2 (Drone Presence Detection), 7.1.3 (Voice Activity Detection in Hearing Aids), and 7.1.4 (Pipeline Defect Assessment). Later, future lines of research that remain open are presented in Section 7.2. Finally, a list of the publications produced during this thesis is presented in Section 7.3.

### 7.1 Summary of the conclusions

In this thesis, we have tried to solve different issues that are present in societies nowadays, and that will undoubtedly require a solution in future projects for smart cities. Studies in the literature have faced these issues before, but in general they have not limited the complexity of their methodologies and the source of data to process, so it is difficult to determine if they would fulfill the requirement of sustainability that must be present in these spaces of the future, where systems must consume as less amount of energy as possible and provide a great autonomy. Along this thesis, sound sources (including both acoustic and ultrasonic signals) combined with energy-efficient systems have proved to be useful, delivering promising results. In the following lines, the conclusions obtained for each of the applications are presented.

#### 7.1.1 Conclusions for VSD

The main contributions of this thesis regarding VSD are:

- A dataset composed of audios from different platforms has been created. In these signals, the mild signs of violence (shouts, increase in the volume of the conversation, etc.) have been added, so it is possible to detect these situations before more severe consequences appear (e.g., gunshots, the appearance of blood in scene, fatalities, etc.). This fact partially answers RQ1.
- A study related to the number of FLOPS required for computing different acoustic features in both frequency and time domains has been carried out. From the set calculated in this thesis, the pitch, the Harmonic Noise Rate and the Ratio of Unvoiced Time Frames are the most computationally costly. With this study, RQ3 is partially answered.
- The MFCCs are the best features for VSD, since the probability of detection improves around 10 points when the system includes them in the feature selection



process (from 64% to 74% using LSLD, and from 67% to 76% using LSQD). It does not happen when pitch and its related features are part of the subset of features, since the probability of detection barely increases when using LSQD (from 76% to 78%), and it is even worse when using LSLD (from 76% falls to 69%). Therefore, high costly features do not necessarily provide better results.

- Regarding the detectors, LSQD always works better than LSLD (between 2 and 9 points).
- In general, higher cost implies better results, but in this application a compromise of 4-5 MFLOPS could be reached as the results do not improve much from this point. This result partially answers RQ2.

### 7.1.2 Conclusions for DPD

Regarding DPD, the contributions provided by this thesis are the following:

- Sounds from five different models of drones and other no-drone sounds present in any city have been included in a new dataset. This way, a challenging dataset has been created and can contribute to further researches. This fact partially answers RQ1.
- Applying a similar computational cost study to the VSD one (RQ3), the pitch and its related features seem to be useful for the problem at hand, as the error rate is significantly reduced when they are selected by the system (from 30.1% to 15.7% using LSLD, and from 23.8% to 15.5% using LSQD). The reason why these features are more useful in the DPD issue than in VSD is that pitch is directly related to the frequency of the sound wave, so it allows the system to differentiate between the frequency of the drones and the frequency dominant in the rest of the sounds of the dataset. In the case of drones, its frequency is in the range of hundreds of Hz, depending on its size, the number of blades and the speed.
- Regarding the detectors, LSQD almost always works better than LSLD, especially when computational restrictions are high.
- With a compromise of 3.5-4 MFLOPS, satisfactory results are obtained in this application. This result answers RQ2.
- The system makes a very clear distinction between drone sounds and other no-drone sounds like helicopter (0.0% of error), excavator (0.0%), motorbike (1.3%), plane (3.1%), or mower (8.2%). It gets worse results with other sounds included in the dataset, such as fire siren (40.7% of error), radial saw (36.4%), or construction work (22.5%).

### 7.1.3 Conclusions for VADHA

In the field of VADHA, the most remarkable contributions of this thesis are:

- Novel EFLECs have proved to be an attractive alternative to MFCCs for VADHA in the widely used dataset QUT-NOISE TIMIT (RQ1), as the results are in line with other proposals from the literature, but highly reducing the resulting computational cost of the system in comparison with the traditional coefficients.

- During the optimization process, in which a balance between the number of features (frequency bands) and the number of neurons of the MLPs has to be reached (RQ3), it has been observed that the system chooses a moderate number of features and a high number of neurons, even selecting the maximum of 20 neurons for some computational cost constraints. This is because a higher number of features does not always imply a better result, while a higher number of neurons can improve the performance, especially in such a large dataset like this.
- From the different scenarios provided in the dataset, the worst results are obtained in the café. This is because it is the environment where most people are having a conversation simultaneously in an enclosed area, so the discussion between the main speakers is easily confused with the rest of the conversations that are taking place.
- The use of cascade-detectors can be useful in algorithms for hearing aids, whose computational resources are very limited, since the performance can be improved while keeping the same computational cost. Using a LSLD in cascade with MLPs, a 2–3% of relative improvement has been obtained, reaching 7% for medium SNR values and 8.5% when using low SNR values. This fact partially answers RQ4.
- In addition, cascade-detectors allow us to keep the same performance, but obtaining some computational cost savings, a very precious resource in the field of hearing aids. Thanks to this system, relative cost reductions up to 25% can be reached when testing medium SNR values, and up to 50% when using low SNR values. RQ4 is fully answered with this result.
- A limit must be set concerning the reduction of the average computational cost of the system, since forcing it to use the original number of features of the more complex MLP detector can carry a reduction in the performance of the whole system. The average computational cost of the system should not be reduced by more than 20–30% of the original cost.
- The proposed method gets the lowest error rate among the methods from the literature that use short window lengths (lower than 20 ms) and that, therefore, are suitable for hearing aids. When the cascade-detector system is restricted to 160 KIPS, it obtains 19.21% of error for medium SNR values and 34.80% for low SNR values. This fact partially answers RQ2.

#### 7.1.4 Conclusions for PDA

The main conclusions of this thesis regarding PDA are:

- The received signal when inspecting a pipeline with ultrasonic guided-waves is different depending on the shape of the defects. In general, it is difficult to obtain an overall solution for all the cases. This has been observed after creating an experimental dataset with the Finite Element Method (FEM) included in Matlab. More than 400 defects have been characterized according to three parameters: length, depth and slope (RQ1).
- In general, the proposed features are useful for addressing PDA, as the results are promising in terms of RMSE. While the average energy and the maximum amplitude seem to be the most relevant ones, the features obtained from echoes when reflections of the signal take place slightly improve the results. Hence, the latter are not essential for solving the problem. This conclusion partially answers RQ5.

- In this field, researchers do not always share the features applied in their systems, as it is a very competitive industry. However, in this thesis six different features have been proposed and explained: maximum amplitude, phase delay, average energy, group delay, maximum amplitude of the echo and average energy of the echoes. This fact also answers RQ5.
- The results in simulated pipelines show that the performance of the system improves as the number of evaluated frequencies is increased. The RMSE decreases from 1.02 mm when using only one frequency (548 kHz) to 0.71 mm when using three frequencies (158, 350 and 548 kHz). Thus, it is important to excite the waves at several frequencies. However, the inclusion of more than three frequencies does not improve the results, quite the opposite. Regarding the number of neurons used in the MLPs, the system works properly when the predictor is simple (1 or 2 neurons), but the results do not improve or even worsen when this predictor is more complex (3 to 5 neurons).
- Three different defects from a real pipeline with 9.27 mm of thickness have been estimated adequately through machine learning (RQ5), getting a RMSE of 1.48 mm. This result has been obtained using a MLP with 2 neurons, 8 computed features (no echo ones) and 2 excited frequencies (158 and 548 kHz).
- There is a clear correspondence between the results obtained in simulated and real pipelines. With the same parameters, in the experimental measurements the RMSE was 0.98 mm, which is not far from the 1.48 mm got in the real situation. That gives an idea of the usefulness of simulation software in PDA when the number of available real measurements is not large.

## 7.2 Future lines

After the study carried out along this thesis, some attractive solutions have been provided in the different issues. However, systems are always capable of being improved, and new challenges arise. Taking into account the goals achieved in this thesis, several research lines could be addressed in the future:

- The next natural step would be to implement the different systems proposed in a real-time environment. In the PDA, the system has already been implemented in the hardware used by the collaborating company, but it would also be interesting to do something similar in the VSD and the DPD issues, whose systems could be tested in an autonomous node using a low-cost solar-powered microprocessor (e.g., Raspberry Pi), and in the application of VAD, where the algorithms of detection could be implemented in a real hearing aid.
- Expanding the created datasets, or testing the systems with new state-of-the-art datasets that could be uploaded in the following years, would be interesting future lines of research. In general, the more data are tested, the more generalizable and robust the system becomes. In the PDA application, it began to be made as it was developed a software where other users could expand the dataset with the measurements of their pipelines, and even label them by manually correcting the obtained profile. This feedback is very useful to the final system, as the algorithm continues learning when new labeled data is provided.
- Some systems could be tested together in a multi-class classification task. If the VSD and DPD datasets are merged, we could convert the detection approach into

a classification approach, and it could be tested if the same microprocessor could solve correctly both issues at the same time, thus reducing the final implementation costs.

- It would also be appealing to improve the performance of the systems with a more specific methodology in each field. In VSD, it would be interesting to introduce a word detector system capable of identifying insults and rude language, which can be directly related to violent situations. In the field of DPD, it would be useful to find new features that allow the system to better differentiate between the drone sound and the no-drone sounds with which the system is often confused (fire siren, radial saw, construction work, etc.). Related to VADHA, it would be interesting to make the system more effective and robust to noisy environments. In PDA, other guided-waves with different behavior in their modes could be tested apart from Lamb waves, such as SH waves.
- Furthermore, the use of computationally restrictive deep learning techniques could be another future line of research. Similarly to the cascade-detectors in VADHA, the viability of computing deep learning features and later reducing the average computational cost of the system could be explored.

### 7.3 List of publications

In the following lines, a list of the published work during the course of this thesis is presented. The accepted and published works in both journals and proceedings that support the main contributions of the thesis or that are related to the content of it are listed below:

- **International journals**
  - Bautista-Durán, M., García-Gómez, J., Gil-Pita, R., Mohino-Herranz, I., and Rosa-Zurera, M. (2017). Energy-Efficient Acoustic Violence Detector for Smart Cities. *International International Journal of Computational Intelligence Systems (IJCIS)*, 10(1), 1298-1305. JCR 2.000 (2017 Impact Factor, Q2 from Artificial Intelligence, and Q2 from Computer Science, Interdisciplinary Applications).
  - García-Gómez, J., Gil-Pita, R., Rosa-Zurera, M., Romero-Camacho, A., Jiménez-Garrido J. A., and García-Benavides V. (2018). Smart Sound Processing for Defect Sizing in Pipelines Using EMAT Actuator Based Multi-Frequency Lamb Waves. *Sensors*, 18(3), 802. JCR 3.076 (2018 Impact Factor, Q1 from Instruments & Instrumentation, Q2 from Physics, Applied, and Q2 from Engineering, Electrical & Electronic).
  - García-Gómez, J., Gil-Pita, R., Aguilar-Ortega, M., Utrilla-Manso, M., Rosa-Zurera, M., Mohino-Herranz, I. (2021). Linear detector and neural networks in cascade for voice activity detection in hearing aids. *Applied Acoustics*, 175, 107832. JCR 2.440 (2020 Impact Factor, Q2 from Computer Science and Artificial Intelligence, and Q2 from Acoustics).
  - Mohino-Herranz, I., Gil-Pita, R., García-Gómez, J., Rosa-Zurera, M., and Seoane, F. (2020). A Wrapper Feature Selection Algorithm: An Emotional Assessment Using Physiological Recordings from Wearable Sensors. *Sensors*, 20(1), 309. JCR 3.073 (2019 Impact Factor, Q2 from Instruments & Instrumentation, Q2 from Physics, Applied, and Q2 from Engineering, Electrical & Electronic).

- **International conferences**

- García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., Mohino-Herranz, I., and Rosa-Zurera, M. (2016). Violence detection in real environments for smart cities. In *Ubiquitous Computing and Ambient Intelligence*.
- Bautista-Durán, M., García-Gómez, J., Gil-Pita, R., Sánchez-Hevia, H. A., Mohino-Herranz, I., and Rosa-Zurera, M. (2017). Acoustic Detection of Violence in Real and Fictional Environments. In *6th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*.
- García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., and Rosa-Zurera, M. (2017). Feature Selection for Real-Time Acoustic Drone Detection Using Genetic Algorithms. In *Audio Engineering Society Convention 142*. Audio Engineering Society.
- Gil-Pita, R., García-Gómez, J., Bautista-Durán, M., Combarro, E., and Cocaña-Fernández, A. (2017). Evolved frequency log-energy coefficients for voice activity detection in hearing aids. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*.
- García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., Romero-Camacho, A., Jiménez-Garrido, J. A., and García-Benavides, V. (2018). Smart Sound Processing for Residual Thickness Estimation using Guided Lamb Waves generated by EMAT. In *27th ASNT Research Symposium*.
- García-Gómez, J., Mohino-Herranz, I., Clares-Crespo, C., Fernández-Toloba, A., and Gil-Pita, R. (2018). Analysis of the performance of Evolved Frequency Log-Energy Coefficients in Hearing Aids for different Cost Constraints and Scenarios. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- Mohino-Herranz, I., García-Gómez, J., Utrilla-Manso, M., and Rosa-Zurera, M. (2018). Precision maximization in anger detection in interactive voice response systems. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- Clares-Crespo, C., García-Gómez, J., Fernández-Toloba, A., Gil-Pita, R., Rosa-Zurera, M., and Utrilla-Manso, M. (2018). Combining the Signals of Sound Sources Separated from Different Nodes after a Pairing Process Using an STFT-Based GCC-PHAT. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- Fernández-Toloba, A., Sánchez-Hevia, H. A., Espino-SanJosé, R., Clares-Crespo, C., García-Gómez, J., and Gil-Pita, R. (2018). Solar Powered Autonomous Node for Wireless Acoustic Sensor Networks Based on ARM Cortex M4. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
- García-Gómez, J., Bautista-Durán, M., Gil-Pita, R., Mohino-Herranz, I., Aguilar-Ortega, M., and Clares-Crespo, C. (2019). Cost-constrained Drone Presence Detection through Smart Sound Processing. In *8th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*.
- García-Gómez, J., Gil-Pita, R., Romero-Camacho, A., Jiménez-Garrido, J. A., García-Benavides, V., Clares-Crespo, C., and Aguilar-Ortega, M. (2019). Ultrasonic Thickness Estimation using Multimodal Guided Lamb Waves. In *ASNT Research Symposium 2019*.
- Aguilar-Ortega, M., Mohino-Herranz, I., Utrilla-Manso, M., García-Gómez, J., Gil-Pita, R., and Rosa-Zurera, M. (2019). Multi-microphone acoustic events

detection and classification for indoor monitoring. In *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*.

- **Other publications**

- Clares-Crespo, C., Gil-Pita, R., Rosa-Zurera, M., García-Gómez, J., and Mohino-Herranz, I. (2019). Mixture Models Applied to the Estimation of Mixing Parameters in Multi-channel Blind Source Separation Algorithms. *International Journal of Signal Processing Systems*, 7(3).