



TESI DOCTORAL

Títol	PERCEPTUALLY-BASED SIGNAL FEATURES FOR ENVIRONMENTAL SOUND CLASSIFICATION
Realitzada per	Xavier Valero González
en el Centre	Enginyeria i Arquitectura La Salle
i en el Departament	Tecnologies Audiovisuals
Dirigida per	Dr. Francesc Alías Pujol

Preface

This PhD was initiated when the author received a grant from the European Commission to carry out an internship at the Joint Research Centre of the European Commission (Ispra, Italy, October 2007 - October 2008). There, he participated in the development of an exploratory research project consisting in the application of pattern recognition techniques for the identification of the environmental noise sources that affect citizen in their day-to-day life. The author continued his research in La Salle-Universitat Ramon Llull, broadening the research theme (considering a wider range of environmental sounds) so as to widening the potential final applications of the initially developed technique. In 2010, he obtained the Diploma of Advanced Studies. During the next two years, the major contributions of the PhD were achieved, writing a total of 3 journal publications (2 accepted and 1 with minor revisions review at the time of submitting this thesis) and 9 papers in conference proceedings. It should be mentioned that some of these contributions received several quality acknowledgments: by the European Acoustics Association, for the papers presented in Euronoise'09 and Euroregio'10, and by the Institute of Noise Control Engineering, for the work in Internoise'12.

Two research stays in European research groups have been carried out within this period. The first one took place in the Acoustics group of the Information Technology Department of Ghent University (Ghent, Belgium), thanks to a mobility grant that the author received from the EU COST Action TUD0840 on Soundscapes of European Cities and Landscapes. During this stay, in collaboration with the researchers of the aforementioned group, the efforts were focused on developing unified learning approaches to come up with the recognition of sound events in urban soundscapes. The outcome of the obtained results was published as invited paper in Internoise'12.

The second stay was carried out in the *Institut de Recherche en Informatique de Toulouse* (IRIT) (Toulouse, France), granted by the Government of Catalonia within the program *Comunitat de Treball dels Pirineus*. The conducted research work addressed the problem of specifically recognising water sound events for medical purposes (assisting doctors in the diagnostic and follow-up of dementia illnesses). At the moment of the writing of this thesis, a paper was being prepared to publish the results that had been obtained using medical data. In addition, the author attended to the Summer School organized by the European Acoustics Association in Ljubljana (Slovenia), where several well-known professors and experts from the acoustics field gave lectures on different hot topics on acoustics.

During the development of the PhD thesis, a total of 13 articles have been written, from which 11 have been published and only two are pendent of acceptance. Next, they are all listed in inverse chronological order (from newer to older).

1. P. Guyot, X. Valero, J. Pinquier, F. Alías, "Two-step detection of water sound events for the diagnostic and monitoring of dementia", in peer review.

2. X. Valero, F. Alías, "Narrow-Band Autocorrelation Function Features for Automatic Acoustic Environment Recognition", in peer review.
3. X. Valero, F. Alías, "Hierarchical Classification of Environmental Noise Sources Considering the Acoustic Signature of Vehicle Pass-bys", in *Archives of Acoustics*, (ISSN 0137-5075), December 2012. *Impact Factor: 0.847*
4. X. Valero, F. Alías, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification", in *IEEE Transactions on Multimedia*, vol. 14, no. 6, (DOI: 10.1109/TMM.2012.2199972), December 2012. *Impact Factor: 1.935*
5. X. Valero, F. Alías; "Acoustic signal analysis by means of bio-inspired Cepstral coefficients and its application to the recognition of soundscapes", in *Proc. Acústica Évora 2012*, Évora (Portugal), October 2012 (in Spanish).
6. X. Valero, F. Alías; "Gammatone Wavelet Features for Sound Classification in Surveillance Applications", in *Proc. 20th European Signal Processing Conference (EUSIPCO 2012)*, (ISSN 2076-1465), pp. 1658-1662, Bucharest (Romania), August 2012.
7. X. Valero, F. Alías; "Classification of Audio Scenes Using Narrow-Band Autocorrelation Features", in *Proc. 20th European Signal Processing Conference (EUSIPCO 2012)*, (ISSN 2076-1465), pp. 2015-2019, Bucharest (Romania), August 2012.
8. X. Valero, D. Oldoni, F. Alías, D. Botteldooren; "Support Vector Machines and Self-Organizing Maps for the recognition of sound events in urban soundscapes" , in *Proc. Inter-Noise 2012*, New York (USA), August 2012.
9. X. Sevillano, X. Valero, F. Alías; "Audio and video cues for geo-tagging online videos in the absence of metadata", in *Proc. 10th Workshop on Content-Based Multimedia Indexing (CBMI2012)*, (ISSN: 1949-3983), pp. 1-6, Annecy (France), June 2012.
10. X. Valero, F. Alías, "Automatic monitoring of environmental noise sources", in *Proc. Tecniacustica'11*, Cáceres (Spain), October 2011 (in Spanish).
11. X. Valero, P. Farré, F. Alías, "Comparison of machine learning techniques for the automatic recognition of soundscapes", in *Proc. Forum Acusticum'11*, Aalborg (Denmark), July 2011. ISSN 2221-3767
12. X. Valero, F. Alías, "Applicability of MPEG-7 low level descriptors to environmental sound source recognition", in *Proc. Euroregio'10*, Ljubljana (Slovenia), September 2010.
13. X. Valero, F. Alías, S. Kephelopoulos, M. Paviotti, "Pattern Recognition and separation of road noise sources by means of ACF, MFCC and probability density estimation", in *Proc. Euronoise'09*, Edinburgh (UK), October 2009.

Abstract

This thesis faces the problem of automatically classifying environmental sounds, i.e., any non-speech or non-music sounds that can be found in the environment. Broadly speaking, two main processes are needed to perform such classification: the signal feature extraction so as to compose representative sound patterns and the machine learning technique that performs the classification of such patterns. The main focus of this research is put on the former, studying relevant signal features that optimally represent the sound characteristics since, according to several references, it is a key issue to attain a robust recognition. This type of audio signals holds many differences with speech or music signals, thus specific features should be determined and adapted to their own characteristics. In this sense, new signal features, inspired by the human auditory system and the human perception of sound, are proposed to improve the representation and classification of environmental sound signals.

Firstly, in the spectral signal analysis domain, Cepstral coefficients computed with the biologically inspired Gammatone filters are proposed and adapted to environmental sound classification, obtaining the so-called Gammatone Cepstral Coefficients (GTCC). The experimental results show an increase in the classification rates when GTCC are used instead of the standard-de-facto Mel Frequency Cepstral Coefficients (MFCC) to describe any of the different tested environmental sound sets. The improvement is attributed to a better representation of the spectral signal details, especially when those appear at low frequency bands.

Secondly, the temporal signal analysis domain is introduced according to the specific characteristics of different environmental sounds. On the one hand, the Gammatone Wavelet coefficients (GTW) are proposed for surveillance-related sounds parameterisation, since they merge the optimum spectral analysis of Gammatone filters with the ability to catch the short duration and impulsive events of Wavelet time-frequency transform. On the other hand, the Narrow-Band Autocorrelation Function (NB-ACF) features are proposed for soundscape signal parameterisation, since they are able to take into account the complex characteristics of such signals that are composed of multiple coexisting sound events. In this case, the NB-ACF features are able to represent non-spectrally overlapped sounds thanks to the detailed analysis (consisting in the parameterisation of the Autocorrelation Function with five perceptually-based descriptors) that is performed in each spectral band. NB-ACF features, especially when combined with Gammatone filter banks, notably outperform MFCC, regardless of the machine learning technique employed.

Finally, a particular case is studied, which deals with the classification of the environmental noise sources that affect human's health and quality of life. Preliminary works flag out the difficulty to distinguish among road vehicle noise sources (car, truck, motorbike). With the goal of improving the classification of such noise sources, a hierarchical classification system that takes into account the different vehicle pass-by phases is proposed. The vehicle pass-by phases refer to the perceptually distinguishable phases in which a vehicle pass-by might be divided into: the approaching, the passing-by and the receding. The proposed scheme, working with Gaussian Mixture Models, is able to yield comparable classification accuracies with respect to a traditional approach employing Hidden Markov Models (a machine learning technique that inherently takes into account the signal time evolution) but with dramatically lower computational cost requirements.

Resumen

Esta tesis plantea el problema de la clasificación automática de sonidos ambientales, es decir, cualquier sonido diferente al habla o a la música que se encuentre en el medio ambiente. En términos generales, se requieren dos grandes procesos para llevar a cabo dicha clasificación: la extracción de descriptores de la señal con el fin de componer patrones representativos de cada tipo de sonido y la técnica de aprendizaje máquina que efectúa la clasificación de dichos patrones. El objetivo principal de esta investigación se centra en el primer proceso, estudiando descriptores de las señales que representen de manera óptima las características de cada sonido, ya que, según varias referencias, es un punto clave para lograr un reconocimiento robusto. Este tipo de señales de audio poseen diferencias significativas con respecto a las señales del habla o de la música. Por lo tanto, los descriptores de la señal deben ser determinados y adaptados a sus características propias. En este sentido, se proponen descriptores inspirados por el sistema auditivo y la percepción sonora humana para mejorar la representación y clasificación de las señales sonoras ambientales.

En primer lugar, en el análisis del dominio espectral de la señal, se proponen y adaptan a la clasificación de sonido ambiental unos coeficientes Cepstrales computados con filtros biológicamente inspirados *Gammatone*, obteniendo los llamados *Gammatone Cepstral Coefficients* (GTCC). Los resultados experimentales muestran un incremento en las tasas de clasificación cuando usamos los GTCC en lugar de los clásicos *Mel Frequency Cepstral Coefficients* (MFCC) para describir cualquiera de los conjuntos de sonidos ambientales testeados. La mejora es atribuida a una mejor recopilación de la información espectral de la señal, especialmente cuando los detalles o particularidades aparecen en bandas bajas de frecuencia.

En segundo lugar, la información del dominio temporal de la señal es introducida acorde con las características específicas de cada conjunto de sonidos ambientales. Por un lado, se proponen los coeficientes *Wavelets Gammatone* (GTW) para parametrizar sonidos relacionados con aplicaciones de vigilancia, dado que conjugan el óptimo análisis espectral de los filtros *Gammatone* con la capacidad de captar eventos impulsivos o de corta duración de la transformada espectro-temporal de *Wavelet*. Por otro lado, se proponen los descriptores *Narrow-Band Autocorrelation Function* (NB-ACF) para parametrizar señales de paisajes sonoros, dada su capacidad para extraer las complejas características de dichos paisajes sonoros compuestos por múltiples y coexistentes eventos sonoros. En este caso, los descriptores NB-ACF son capaces de representar sonidos espectralmente no superpuestos gracias al análisis detallado (consistente en la parametrización de la función de autocorrelación mediante cinco parámetros perceptuales) realizado independientemente en cada banda espectral. Los NB-ACF superan a los MFCC independientemente de la técnica de aprendizaje máquina utilizada, especialmente cuando éstos son calculados un banco de filtros *Gammatone*.

Por último, se estudia el caso particular de la clasificación de fuentes de ruido ambiental que afectan a la salud y calidad de vida de las personas. En trabajos preliminares se detectó la dificultad de distinguir entre fuentes de ruido de tráfico (coche, camión, moto). Con el objetivo de mejorar la clasificación de dichas fuentes de ruido, se propone un sistema de clasificación jerárquica que considera las distintas fases del paso de un vehículo. Las fases de paso se refiere a las fases en que se puede dividir el paso de un vehículo y que son perceptualmente distinguibles: aproximación, paso y alejamiento. El esquema propuesto, que usa modelos de mezcla de Gaussianas, proporciona una precisión en la clasificación comparable a una aproximación clásica con modelos ocultos de Markov (técnica de aprendizaje que contempla intrínsecamente la evolución temporal de la señal) pero con unos requisitos computacionales notablemente inferiores.

Resum

Aquesta tesi planteja el problema de la classificació automàtica de sons ambientals, és a dir, qualsevol so diferent a la parla o la música que es trobi en el medi ambient. En termes generals, es necessiten dos grans processos per dur a terme aquesta classificació: l'extracció de descriptors del senyal per tal de conformar patrons representatius de cada tipus de so i la tècnica d'aprenentatge màquina que efectua la classificació dels esmentats patrons. L'objectiu principal d'aquesta investigació se centra en el primer procés, estudiant descriptors dels senyals que representin de manera òptima les característiques de cada so, ja que, segons diverses referències, es tracta d'un punt clau per aconseguir un reconeixement robust. Aquest tipus de senyals d'àudio tenen diferències significatives respecte als senyals de la parla o de la música. Per tant, s'hauran de determinar descriptors del senyal adaptats a les seves característiques pròpies. En aquest sentit, es proposen descriptors inspirats pel sistema auditiu i la percepció sonora humana per millorar la representació i classificació dels senyals sonors ambientals.

En primer lloc, en l'anàlisi del domini espectral del senyal, es proposen i s'adapten a la classificació de so ambiental uns coeficients Cepstrals computats amb filtres biològicament inspirats *Gammatone*, anomenats *Gammatone Cepstral Coefficients* (GTCC). Els resultats experimentals mostren un increment en les tasses de classificació quan usem els GTCC enlloc dels clàssics *Mel Frequency Cepstral Coefficients* (MFCC) per descriure qualsevol dels conjunts de sons ambientals testejats. La millora és atribuïda a una millor recopilació de la informació espectral del senyal, especialment quan els detalls o particularitats apareixen en bandes baixes de freqüència.

En segon lloc, la informació del domini temporal del senyal és introduïda segons les característiques específiques de cada conjunt de sons ambientals. D'una banda, es proposen els coeficients *Wavelets Gammatone* (GTW) per parametritzar sons relacionats amb aplicacions de vigilància, donat que conjuguen l'òptim anàlisi espectral dels filtres *Gammatone* amb la capacitat de captar esdeveniments impulsius o de curta durada de la transformada espectre-temporal de *Wavelet*. D'altra banda, es proposen els descriptors *Narrow-Band Autocorrelation Function* (NB-ACF) per parametritzar senyals de paisatges sonors, donada la seva capacitat per extreure les complexes característiques d'aquests paisatges sonors composts per múltiples i coexistents esdeveniments sonors. En aquest cas, els descriptors NB-ACF són capaços de representar sons espectralment no superposats gràcies a l'anàlisi detallat (consistent en la parametrització de la funció d'autocorrelació amb cinc paràmetres perceptuals) realitzat independentment a cada banda espectral. Els NB-ACF superen als MFCC independentment de la tècnica d'aprenentatge màquina emprada, especialment quan es calculen amb un banc de filtres *Gammatone*.

Finalment, s'estudia el cas particular de la classificació de fonts de soroll ambiental que afecten la salut i la qualitat de vida de les persones. En treballs preliminars es va detectar la dificultat de distingir entre fonts de soroll de trànsit (cotxe, camió, moto). Amb l'objectiu de millorar la classificació d'aquestes fonts de soroll, es proposa un sistema de classificació jeràrquica que considera les diferents fases del pas d'un vehicle. Les fases de pas es refereix a les fases en què es pot dividir el pas d'un vehicle i que són perceptualment distingibles: l'aproximació, el pas i l'allunyament. L'esquema proposat, que utilitza models de mescla de Gaussians, proporciona una precisió en la classificació comparable amb una aproximació clàssica emprant models ocults de Markov (una tècnica d'aprenentatge màquina que contempla intrínsecament l'evolució temporal del senyal) però amb uns requisits computacionals notablement inferiors.

Table of Contents

Preface	i
Abstract	iii
Resumen.....	v
Resum.....	vii
List of Acronyms.....	xi
1. Introduction	1
1.1 Context	1
1.1.1 Framework	1
1.1.2 Research group	2
1.2 Applications.....	2
1.2.1 Surveillance systems	3
1.2.2 Well-being	3
1.2.3 Polices and regulations	4
1.2.4 Health.....	4
1.2.5 Hearing aids.....	5
1.2.6 Information technology	5
1.2.7 Multimedia content	5
1.3 Scope	6
1.3.1 Terminology	6
1.3.2 Recognition process	7
1.3.3 Specifications	8
1.4 Objectives.....	8
2. Background.....	11
2.1 Environmental sound characteristics.....	11
2.2 Pattern recognition approach.....	13
2.3 Outline of the research work performed.....	14
3. Milestone 1: Frequency-domain signal features.....	19
3.1. Introduction	19
3.2 Publications Milestone 1.....	21

3.2.1 Publication Acústica Évora 2012	21
3.2.2 Publication IEEE Transactions on Multimedia.....	33
4. Milestone 2: Time-frequency signal features.....	49
4.1 Introduction.....	49
4.1.1 Gammatone Wavelet features.....	49
4.1.2 Narrow-Band Autocorrelation features	51
4.2 Publications Milestone 2	55
4.2.1 Publication EUSIPCO-GTW	55
4.2.2 Publication EUSIPCO-NBACF	67
4.2.3 Publication in peer review.....	79
5. Milestone 3: Application to road vehicle pass-by classification	101
5.1 Introduction.....	101
5.2 Publications Milestone 3	105
5.2.1 Publication Tecniacustica	105
5.2.2 Publication Archives Acoustics.....	117
6. Discussion.....	139
6.1 Results extracted from the Milestones	139
6.1.1 Milestone 1.....	139
6.1.2 Milestone 2.....	140
6.1.3 Milestone 3.....	141
6.2 General comments.....	142
7. Conclusions and future lines	145
7.1 Conclusions.....	145
7.2 Future lines.....	146
References.....	149
Annex A: First studies.....	153
A.1 Probability density estimation for road noise source monitoring.....	155
A.2 MPEG-7 parameters for environmental sound recognition	165
A.3 Machine Learning techniques comparison for soundscape recognition.....	173
Annex B: Other applications of environmental sound classification	185
B.1 Audio-based geo-tagging	187
B.2 SVM and SOM for urban sound recognition	199
B.3 Water sound event detection	211

List of Acronyms

ACF	Autocorrelation Function
ASE	Audio Spectral Envelope
ASR	Automatic Speech Recognition
BB	Broadband
BB-ACF	Broadband Autocorrelation Function
DT	Decision Tree
DWC	Discrete Wavelet Coefficients
END	Environmental Noise Directive
ERB	Equal Rectangular Bandwidth
GMM	Gaussian Mixture Model
GT	Gammatone
GTCC	Gammatone Cepstral Coefficients
GTW	Gammatone Wavelet
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
ISO	International Organization for Standardization
KNN	K-Nearest Neighbours
LPC	Linear Prediction Coefficients
MFCC	Mel Frequency Cepstral Coefficients
MIR	Music Information Retrieval
MPEG	Moving Picture Experts Group
NB	Narrow Band
NB-ACF	Narrow Band Autocorrelation Function
NN	Neural Network
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
SBER	Sub-Band Energy Ratio
SF	Spectral Flatness
SOM	Self-Organising Map
SRO	Spectral Roll-Off
STE	Short-Term Energy
SVM	Support Vector Machine
ZCR	Zero Crossing Rate

1. Introduction

1.1 Context

1.1.1 Framework

In the framework of Human-machine or Human-Computer Interaction (HCI), it is desired to provide the Computer with mechanisms to receive the input data, process this input information and finally generate an appropriate response. This thesis deals with the analytic part, which aims at mimicking human abilities as first step to improve the interaction with the environment.

In this context, there are five senses used by humans to collect information about the world around them that should be taken into consideration by HCI research: vision, hearing, touch, smell and taste. In principle, the sense of vision seems to be the most important of the five [1]. To date, apart from technical solutions of low complexity, such as light sensors, pressure, speed, etc., the field of HCI studied in greater depth is the one related to vision. Commonly called Computer Vision, its objective is to transfer the ability to analyse visual information to a machine. To that effect, it embraces methods for acquiring, processing, analysing, and understanding images from the real world.

Next, in order of importance, we find the ear. In this domain, there are some types of sound signals that have been deeply studied in the past years. We are mainly talking about music and speech signals. Music is a structured sound, organized by humans to convey some aesthetic intent. Music Information Retrieval (MIR) is the name behind the music signal study, which gathers several tasks such as: genre classification, mood classification, artist identification, instrument recognition or music annotation [2]. Speech refers to sounds produced by the human vocal tract that have linguistic content. Nowadays, there are commercial Automatic Speech Recognition (ASR) systems achieving high reliability in controlled environments. It is

certainly a very useful technique because it gives the machine the ability to "understand" people (in a controlled dialogue).

Nevertheless, beyond speech and music, much more information about the surrounding environment can be captured from audio signals. Since ancient times, mammals have employed the sound information for surviving, for instance, to recognize a predator approaching. Nowadays, in the midst of a city, the humans are able to identify multiple sounds such as cars, people talking, church bells ringing, the barking of a dog, the siren of an ambulance, etc.

The recognition of non-speech and non-music sound events can be of high interest in manifold applications. Section 1.2 gives an overview of some of these possible applications.

1.1.2 Research group

The research carried out in this thesis is developed in the Research Group on Media Technologies (GTM) from La Salle-Universitat Ramon Llull. The GTM is a multidisciplinary research group focused in the innovation of new and multimodal multimedia technologies. It was born from the union of two original research groups: Multimedia Processing Group, with official reference 2005-SGR-00806, and the Audivisual and Multimedia Technologies Group, with official reference 2005-SGR-00682. As result of the union, the new group receives the recognition from Catalanian Government, with official reference 2009-SGR-293. The objective is achieving a greater impulse from the merging of the different activities, given their similarity: all of them deal with the technologies media processing. Specifically, GTM integrates several knowledge areas, such as:

- Multimodal processing
- Acoustics
- Digital television
- Multimedia
- Usability and user experience

These knowledge areas enable to consider the main aspects about media technologies: signal processing (including speech, audio, image and video), media content transmission, synthetic content generation by means of graphics and 3D animation technologies and, finally, evaluation of the user experience of the multimedia content supplied to the consumer. The current work is settled in the acoustics knowledge area, and specifically, in the sound signal processing field.

Dr. Francesc Alías has been supervising the research work carried out culminated with this thesis. He is currently Director of the Human-Computer Interaction area at La Salle-URL.

1.2 Applications

Environmental sound recognition might have manifold applications, becoming interesting in a wide range of domains, such as tele-surveillance, smart cities, health, information technology, etc. In the next lines, we give some examples of applications in each domain.

1.2.1 Surveillance systems

Security surveillance

Traditional image-based surveillance systems may be improved by adding audio information, which provides several advantages: cheaper sensors, fixing the image limitation due to the blind spots, preservation of the personal privacy (no image is stored from watched people), possibility of easily triggering an alarm warning the emergency or police services, etc. [3], [4]. These systems may be implanted for security purposes in indoor environments, such as private houses, parking and elevators; or in outdoor environments, such as streets and other public places, e.g., to control violent acts.

Elderly people surveillance

Detection of abnormal sounds that could represent a hazardous situation for the elderly or people needing special cares who are living home alone [5]. For instance, when recognizing the sound of a person falling down the stairs, an alarm could be automatically triggered warning a relative or the ambulance service. The preservation of the personal privacy is a valuable feature in these applications.

1.2.2 Well-being

Smart Cities

Smart audio sensors distributed all around the urban area can provide with plenty of useful information in real time about the current city situation, such as: traffic conditions, celebration of events (e.g., demonstrations or street performances), identification of noisy areas, etc. A central server would centralise all the relevant information, thus rapid measures could be taken to solve or control the observed problems.

Ambient Assisted Living

Ambient Assisted Living aims at designing smart homes that anticipate to the needs of its inhabitants while maintaining their safety and comfort [6]. Smart Homes are typically equipped with many sensors that capture different information about the environment. Specifically, the sound data recorded by microphones can deliver highly informative data. For instance, the identification of coins or keys dropping may alert the user and help him to find them. Another example is the recognition of water sounds, which can prevent water waste due to leakage in tubes and flushing toilets, or alert the user if he/she left the tap from the sink or the bath opened.

Soundscapes

Environmental sound recognition systems could be useful for urban planners and soundscape designers who aim at designing pleasant urban and rural acoustic environments. In that sense, a support tool that provides them with the information about the typical sound events is relevant information that the urban planner would acknowledge to appropriately design an acoustic environment that improves the quality of life of their inhabitants.

1.2.3 Polices and regulations

Environmental noise polices

The European Environmental Noise Directive (EU/2002/49) [7] requires mapping all major environmental noise sources within urban areas in order to inform the European citizens about their exposure to noise besides drawing up appropriate action plans addressed to reduce noise impact on people. An important specification is that environmental noise sources must be mapped independently. Therefore, maps in locations of coexisting noise sources could achieve higher accuracy by employing advanced monitoring systems able to identify the sources that originate the measured noise levels.

Noise impact studies

Environmental sound source recognition systems may be employed as support tool in the assessment of noise impact assessments attending to local regulations. Such systems would become especially useful in situations where the citizen is exposed to several noise sources and it is difficult to determine the contribution of each noise source to the overall noise exposure. Frequently, long measurements lasting from several days up to several months are required to obtain a significant sample of the acoustic situation [8]. In those cases, a system able to automatically recognise the noise sources is of special interest, since no technician is needed to ensure the origin of the registered noise levels. For the unaware reader, it should be pointed out that noise sources are not all treated in the same way by noise regulations, e.g. they are more restrictive with noise from industrial activities than with road traffic noise. In addition, penalisations for exceeding noise limits also differ.

1.2.4 Health

Noise health impact

The impact of noise on human beings has not been clearly determined yet. The understanding of noise health effect could be improved by means of epidemiological studies [9], [10]. These studies could be performed with the information provided by an environmental noise source detector embedded in portable devices, which would collect all the relevant information regarding the noise sources to which a person is exposed to in his/her daily-life. These devices would help to better understand the impact of daily-life noise on health, studying the correlations between noise exposure and other human health endpoints such as cardiovascular diseases, blood pressure, annoyance, etc.

Psychological studies

Likewise, environmental sound source detectors would be useful to assess the psychological human response to noise, studying the noise sources or specific characteristics of sounds that have a greater impact on humans and provoke different reactions to them, such as disturbance, annoyance or nuisance [11].

Support decision tool for medical purposes

The only way to diagnose some degenerative illness (such as Alzheimer) is observing the patient when performing several daily routine activities, such as preparing a coffee, washing hands, going to the toilet, etc. Thus, videos are recorded from the patient at his/her house during several days. Since it is not feasible that the doctor had to watch videos lasting several

days of routine activities, the main tasks of the monitoring system may be focused on detecting abnormal patterns by analysing the sound of the events, thus generating smart summaries with all the relevant information required by the doctor to make the diagnosis [12].

1.2.5 Hearing aids

Hearing aid systems

Enable an active control that could adapt the hearing reinforcement system to the particular characteristics of the surrounding environment. Interfering noise could be reduced and certain relevant sounds (e.g., human voice) emphasized [13].

Audio-to-Image transducers

Devices that display audio information (e.g., doorbell, phone ring, alarm signals, etc.) on a monitor. That application would be especially useful for hearing impaired people.

1.2.6 Information technology

Audio context recognition for portable devices

Portable devices like mobile phones or music players could automatically detect the surrounding acoustic scene and reacting accordingly without human intervention [3]. For instance, a mobile phone could detect that we are on a library and it could automatically switch off the volume in order to avoid bothering other library users.

Robotics

Guidance of robots and other automates by sound information. This application results especially useful in cases of absence of light, when the robot can get degraded or no visual information about the environment [14].

Automotive applications

Audio sensors in cars could identify certain events happening in a road, e.g. ambulance or police sirens, horns, etc. The driver could be warned in short time, even if he is listening to other audio devices that could prevent him from a quick reaction.

Speech recognition and synthesis systems

Systems devoted to recognise the background noise and cancel it, thus improving the Signal-to-Noise ratio and improving the spoken communication [15]. Moreover, explore how adverse environmental noise conditions influence into speech modification, so as to incorporate these changes to enhance natural and synthetic speech [16].

1.2.7 Multimedia content

Audio content identification in media broadcasts

Both in pre-recorded and live media broadcasts, information about audio content such as language, audio quality, etc. could be identified and transmitted together with the video signal to the final user.

Audio indexing

Large collections of audio databases could be automatically indexed and quickly accessed by retrieving special audio indexes such as sound sources or musical instruments contained [2], [17].

Multimedia geo-tagging

Tagging videos with the geo-coordinates of the place where they were filmed enables browsing and searching online multimedia repositories using geographical criteria. Frequently, there is no meta-data included about the filming location. In those cases, audio information might be used in order to determine the coordinates (geolocalization) where the video was filmed with a certain radius of tolerance [18].

Personal diary

A new tendency, especially extended among musicians and people working in the audio field, is the creation of personal diaries based on audio information. It represents an alternative method to register the different events and experiences lived [19]. Audio content indexing is especially relevant in this application.

1.3 Scope

In this section, we aim to determine the scope of the problem addressed in the current research work. First, some terminology issues are clarified. Next a typical recognition process is defined and finally, some specifications regarding the scope of the research are determined.

1.3.1 Terminology

As already commented in section 1.1, this work is focused on recognizing non-speech and non-music sounds. We will refer to them as *environmental sounds*, since all of them naturally occur in the environment. Even though the word *environmental* can refer to speech and music as well, we will exclude those in the rest of this document.

Second, the name of the addressed task is specified. For that purpose, the main actions related to the processing of environmental sounds are defined as:

- **Recognize:** to know someone or something because you have seen, heard or experienced them before. In this field, to notice a certain sound (which has been heard before) within a continuous audio stream.
- **Identify:** to recognize someone or something and say or prove who or what they are. In this field, recognize and label a certain sound within a continuous audio stream.
- **Detect:** to notice something that is partly hidden or not clear, or to discover something using a special method. In this field, to notice any sound event within a continuous audio stream.
- **Segment:** to divide something into different parts. In this field, divide an audio stream into different parts according to its content.
- **Classify:** to divide things into groups according to their type. In this field, label a certain isolated sound event according to its content.

According to the previous list of definitions, recognizing a sound means indicating whether it has already been heard or not. Identifying a sound requires to label exactly the sound under a certain predefined name or category. However, in the related literature, the term *recognition* is used more often¹, even though the task to fulfil also involves the identification of the sound. Therefore, hereafter both concepts will be indistinctly employed, as they are in the related literature.

1.3.2 Recognition process

Once noted the naming issue, let's define what a recognition (or identification) process consists of (see Figure 1). Firstly, the sound event is detected within a continuous audio stream recorded by a microphone. The detection allows the segmentation and isolation of the audio segments of interest. Secondly, the segmented sound event is automatically classified and labelled with a predefined sound name. To carry out this second part of the identification, the sound signal is parameterized with a set of features. Optionally, a signal feature selection or data compaction process may be included for efficiency purposes. The selected signal features feed a machine learning method, which gives the respective label among a collection of predefined sound names or categories already labelled in the past. Thus, the classification task assumes working with independent isolated audio streams.



Figure 1. Diagram block of a sound event recognition process

In the state of the art, there are works which are exclusively focused on the classification problem [5], [14], [15] whereas others have addressed the whole recognition problem [12], [20]-[25]. In the latter case, the need to also performing the detection prior to the classification is the reason behind the variety of the recognition approaches followed. In a meeting room context, Temko et al. proposed a hierarchical system was followed: first the continuous audio stream was segmented in “silence” and “non-silence” segments, and then the segments labelled as “non-silence” were classified into the different target sounds [20]. With the same goal, Zieger proposed merging the detection and classification in one step [21], by adding extra classes corresponding to the non-target sounds that could appear within the continuous audio stream. But these are not the only possible approaches. In a rather different application, the recognition of passing road traffic vehicles, Sobreira et al. triggered the passing vehicles using thresholds based on signal energy statistics [22], whereas Ntalampiras et al. eliminated the silence segments with a prior analysis of the signal amplitude [23]. In a medical tele-surveillance application, Istrate et al. [24] employed the Wavelet signal transform to detect the impulsive target sounds, and Guyot et al. used a spectra-based signal feature and some thresholds to detect water sound events [12].

¹ At the moment of writing this thesis, the keyword *Environmental Sound Recognition* yielded 537 entries in Google Scholar, by only 96 entries yielded by *Environmental Sound Identification*

Bearing in mind the wide scope of sounds considered in this thesis (*any* non-speech and non-music sound), facing the complete recognition problem would represent such a huge and extensive task, since *i)* the effort to collect databases of continuous audio streams containing a wide range of sounds would be enormous, and *ii)* the detection process will probably entail specific customizations for tackling the recognition of the different types of environmental sounds.

In contrast, in this thesis we will only address the classification problem. Firstly, because we aim at coming up with a quite generic approach for any environmental sound, thus avoiding the detection process adaptation for specific applications. And secondly, because collecting isolated audio samples from a large variety of environmental sounds is much more affordable than collecting continuous audio streams.

At any case, the obtained findings in the task of environmental sound classification may be integrated and applicable to the sound recognition problem, and would provide a positive effect also in that wider-scope task.

1.3.3 Specifications

In this work, only monaural sound signals will be considered, with an eye to implementing the sound recognition technique on any multimedia device disposing of a single microphone, such as mobile phones, tablets or sound level meters.

The case of mixed sound sources will be only considered in the case of soundscapes (see Sections 4.2.2 and 4.2.3), which may be composed of several coexisting environmental sound sources. It should be noted that, soundscapes are considered as audio scenes or acoustic environments in in this thesis, although their original definition is wider and may also integrate the concept of human-environment interaction.

Finally, it should be noted that behind the term *environmental sounds*, it may be included any non-speech or non-music sound, resulting into a quasi-infinite range of sounds (see Section 2.1). So far, building a device able to recognise *any* sound event at *any* given environment seems, in principle, utopic. Therefore, a realistic approach consists in focusing on a specific range of environmental sounds related to a certain system application. Along the research conducted, we will follow this statement, carrying out the experimental work with specific corpora containing a range of environmental sounds limited to the application at hand.

1.4 Objectives

The final aim of thesis is improving the current techniques for recognising and, specifically, classifying any type of environmental sound. That involves studying the different stages of the sound recognition process, including the signal feature extraction, dimensionality reduction and classification scheme (see section 2.2 for further details). There are myriad of works in the related literature that have stressed the importance of the signal feature extraction in order to achieve robust environmental sound recognition [2], [5], [14], [26]. No matter how sophisticated the classification algorithm is that it won't succeed in the recognition task if the input data is not able to capture the specific particularities of environmental sounds. Hence, parameterising sound signals by a set of signal features specifically designed to represent this

kind of sounds becomes essential, since they need to represent accurately their characteristics while being discriminative of the different sound classes (i.e., maximising the intra-class similarities while maximising the inter-class differences).

This is the primary reason to focus this research work on studying and proposing new features that could be useful to classify environmental sound signals. Both time and frequency signal analysis domains will be studied, since relevant information may be obtained from both [27]. Additionally, human auditory sensing and sound perception will be considered in order to propose signal parameterizations aligned with the information that humans perceive from a given environment. Although the primary scope is on feature extraction, the remaining stages of the sound classification process cannot be neglected, and will also be studied in this work.

In principle, it is desired that the studied technique could be applied to any type of environmental sound and, thus, to any of the derived applications (see section 1.2). However, certain applications are related to environmental sounds with some specific characteristics (i.e., impulsive behaviour in surveillance-related sounds or coexisting sound events in soundscapes). Thus, specific techniques might be designed for these specific cases.

In the next lines, the main objectives of this work are summarised:

- Analyse the environmental sound signals (from both time and frequency signal analysis domain), acknowledging their diverse characteristics.
- Consider the way how humans capture and process the sound signals, as well as human sound perception concepts such as loudness, pitch perception, timbre, etc.
- Study, propose and determine novel sound signal parameterisations that could improve the current state of the art signal features according to the observed sound signal characteristics (both in the time and frequency signal analysis domain).
- Select and adapt classical supervised machine learning algorithms to perform the classification of the sound events.
- Research into how the general technique might be adapted to tackle a very specific sound classification problem (e.g., environmental noise sources, soundscapes).

In order to objectively evaluate the proposed sound classification technique and relate it to previous approaches, the use of sound corpora is essential. The databases of environmental sounds are generally sparse, and in many cases field measurements will be carried out in order to complete the available sound data. The composition and origin of the employed corpus will be detailed in each publication.

The remainder of this document is organized as follows. Section 2 introduces the followed approach and outlines the performed research, which has three main contributions that are subsequently presented in Sections 3, 4 and 5. In Section 6 the obtained results are globally discussed. Finally, Section 7 draws up the conclusions and proposes new directions to continue this research in the future.

2. Background

2.1 Environmental sound characteristics

When it comes to design specific features for the analysis and recognition of environmental sound signals, it is especially relevant to know the characteristics of this kind of signals beforehand. This section gives an overview of the characteristics of such sound signals, while stressing their differences to speech and music signals.

Firstly, regarding the range of possible sounds, both in speech and in music we find a limited amount of sound units: phonemes and notes, respectively. On the contrary, the range of environmental sounds is infinite, since any occurring sound in the environment may be included in this category. Secondly, a certain periodicity can be observed both in speech and music signals when analysing these sound signals in the time domain (see Figure 2). Although with some exceptions (i.e., some natural sounds as bird chirps or cricket sounds), the periodicity in environmental sounds may not exist. Thirdly, the complexity of the spectrum of environmental sounds is notably larger than speech or music signals, as depicted in Figure 3. Moreover, the phonemes and musical notes are combined so as to obtain meaningful sequences that are actually transmitting a particular message. As opposed, the sequences on environmental sounds do not follow any rule or predefined grammar, although they may convey some kind of meaning. Unlike speech and music, also other important information is unknown, such as the duration of the sound events or the proportion between harmonic and inharmonic spectral structure (see Table 1), thus making the recognition of environmental sounds much more complicated.

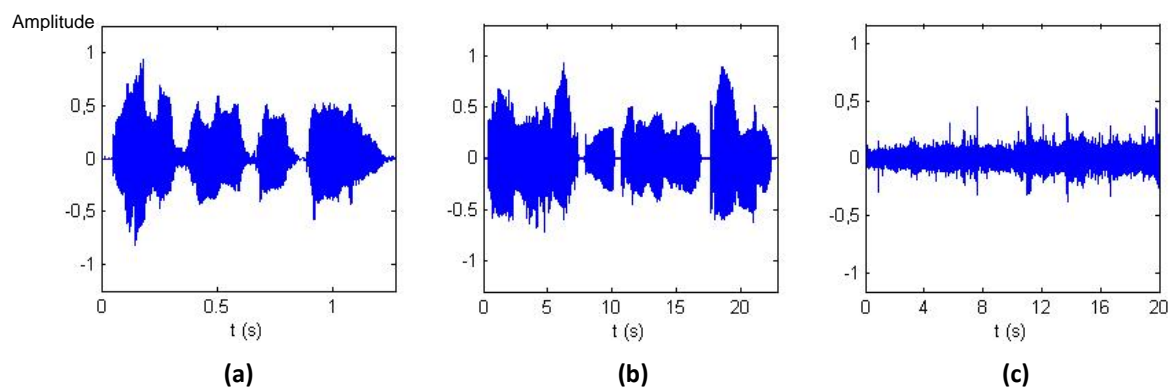


Figure 2. Time envelope of a: (a) speech signal; (b) music signal (clarinet); (c) environmental sound signal (traffic street).

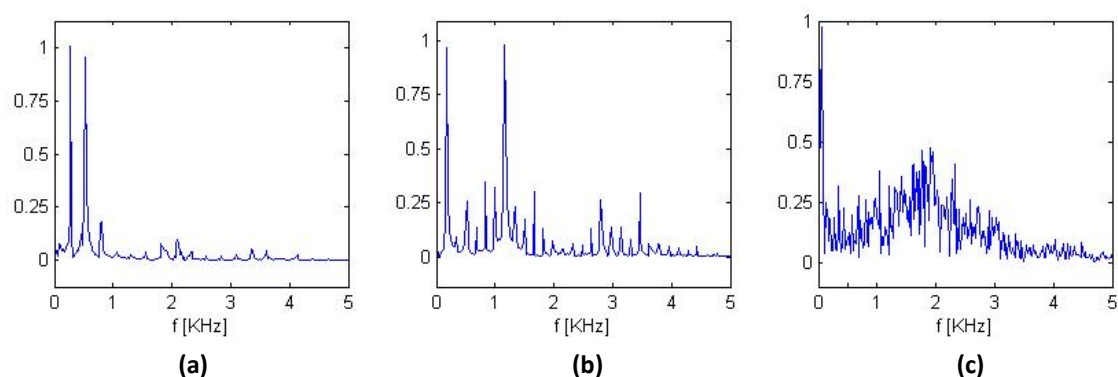


Figure 3. Normalized Fourier Transform of the: (a) speech signal; (b) music signal (clarinet); (c) environmental sound signal (traffic street) from Figure 2.

Features	Speech	Music	Environmental Sounds
Unit analysis	Phonemes	Notes	Events
Source	Human vocal tract	Instruments	Any occurring sound producing an event
Range of possible sources	Finite	Finite	Infinite
Temporal structure	Short (40-200ms). More steady than dynamic. Timing constrained but variable.	Long (600-1200ms). Mix of steady-state (strings, winds) and transient (percussion). Strong periodicity.	Long (500-3000ms). Unknown proportion of steady-state to dynamic. Variable periodicity.
Spectral structure	Largely harmonic (vowels, voiced consonants); Spectral energy tends to group as formants. Some inharmonic (stops, fricatives).	Largely harmonic. Some inharmonic (e.g., percussion).	Unknown proportion harmonic vs. inharmonic, although both exist.
Syntactic / Semantic structure	Symbolic, productive, can be combined in grammar.	Symbolic, productive, combined in grammar.	Not productive, unknown grammar, although meaning sequences may exist.

Table 1. List of features that characterise speech, music and environmental sounds [27].

Acoustic parameters such as fundamental frequency or more complex signal features such as the Mel Frequency Cepstral Coefficients are typically used to identify speech sounds. Music research also includes acoustic features to parameterize the timbre, time and spectral envelope characteristics. As opposed, in the literature it is difficult to find a set of validated signal features designed to specifically identify environmental sounds. So far, they are typically borrowed from speech and music research fields. However, these features designed for significantly different signals (see Figures 2 and 3) may not be optimal for dealing with environmental sounds. Thus, it arises the need of studying and designing signal features adapted to the complex characteristics of such sound signals. Indeed, this is the main goal of this research work.

2.2 Pattern recognition approach

In this section, the general methodology followed in this work is described. It is based on a typical sound pattern recognition approach, slightly adapted to the characteristics of environmental sounds (see Figure 4).

The sound signal is first windowed into short frames, usually of 10-50 ms [2]. This process has a dual purpose. On the one hand, the typically non-stationary audio signal can be assumed to be stationary for such a short signal frame, thus facilitating the spectro-temporal signal analysis. On the other hand, the efficiency of the feature extraction process is increased, since the system will deal with smaller chunks of data (thus, yielding a faster spectral analysis computation) [2]. Next, the framed sound signal is parameterised by a set of features, which may consider the time, frequency and perceptual signal characteristics. However, no automatic signal feature selection process is applied, since we are interested in understanding the explanation behind the results, and that is only possible when the feature sets are maintained in its original feature space. As noted in Section 1.3, the selection of appropriate signal features is a key issue for environmental sound recognition. Modelling the time evolution of those signals has been found to be of paramount importance when it comes to recognise environmental sounds [27]. To keep this time information, the features extracted from several subsequent signal frames are all merged into a single feature vector, which represents the sound signal aimed to be recognised.

It should be noted that, due to this feature merging process, the feature vectors acquire a very high dimensionality that may represent a hurdle to the machine learning algorithm, with the so-called “curse of dimensionality” problem [28]. In order to compact the feature vectors (process denoted as *Data Reduction* in Figure 4), several approaches may be considered: from extracting several statistics (such as mean values) [29] to more complex approaches like analysing the Principal Components of the feature vector [3]. There are other works

At this point, and based on the information provided by the compacted feature vector, a process is needed to perform the recognition of the sound signal. According to the definition of *recognition* (see section 1.3), this process is conducted by means of a machine learning technique that follows a supervised learning approach. Hence, regardless of the specific algorithm employed, a set of known sound data (hereafter denoted as labelled data) is needed for the training of the supervised learning approach. To that effect, multiple samples from

each sound class should be collected, building representative patterns for each type of sound class aimed to be recognised. These sets of sound patterns will compose the acoustic model of the sound class and will be employed to train the machine learning algorithm. This process is performed offline and, in general, it is conducted only once². Later, the machine learning technique is asked to identify unknown sound signals based on the information acquired. As opposed to the training stage, the testing process should be done *online* in applications that require real time response from the environmental recognition system.

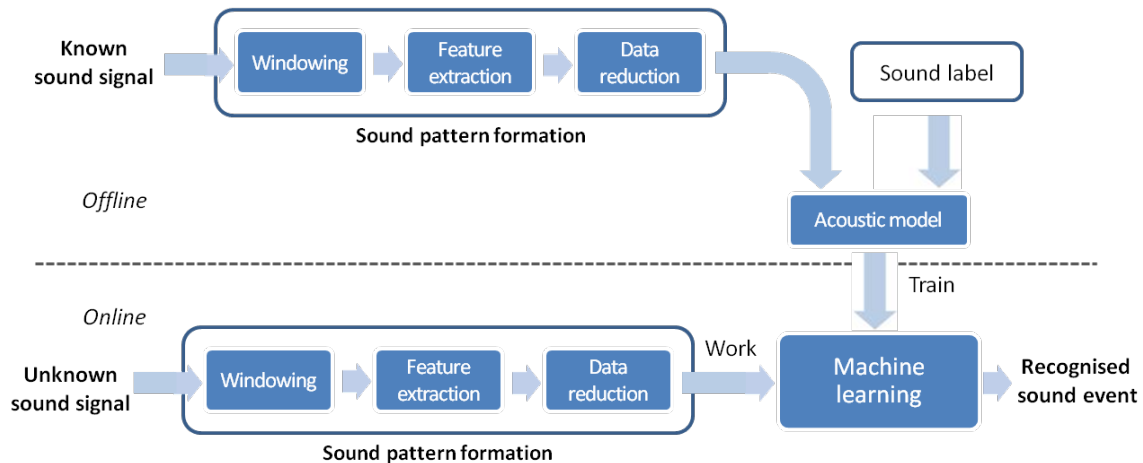


Figure 4. Block diagram of the process followed to recognise environmental sounds in this work.

2.3 Outline of the research work performed

As specified in Section 1.4, the research conducted in this thesis has been focused on determining features that optimally model the sound signals so as to allow a more robust sound event classification.

Traditional approaches on feature extraction have been mainly focused on modelling the spectral domain of the sound signal. Hence, in this work, the first idea was enhancing the spectral representation of the sound signals. This goal is faced by performing a spectral analysis inspired by the human auditory system (which models the spectral response of the cochlea) employing the so-called Gammatone Cepstral Coefficients.

In order to go a step further, it was sought to add the time-domain signal analysis, yielding to spectro-temporal sound signal features. Since the signal characteristics in the time domain notably vary depending on the environmental sounds nature, two solutions are proposed in this sense. On the one hand, a solution combining Wavelet analysis with the biologically-inspired Gammatone filters is proposed for short duration and impulsive environmental sounds. On the other hand, a narrow-band signal autocorrelation analysis is proposed to analyse coexisting sounds in auditory scenes.

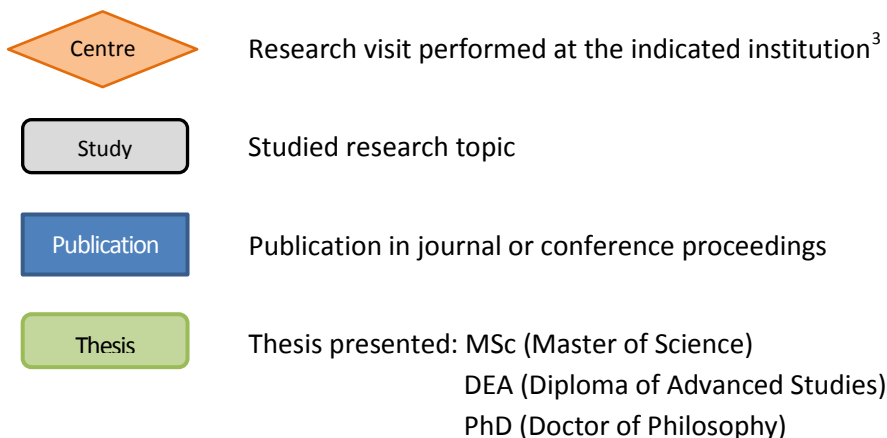
² Nevertheless, this process could be updated dynamically, by repeating the system training every time that new sound data is available.

Finally, rather than applying the proposed techniques on more or less general environmental sound classification problems, we wanted to discuss the adaption of the classification scheme to a particular application: the monitoring of urban noise sources in a sensor network (see Section 1.2.2). Specifically, a hierarchical classification scheme that takes into account the acoustic signature of road traffic vehicles is proposed. That case epitomizes the solutions that can be applied to retrieve sound signal information in specific applications.

Furthermore, other related contributions achieved during the research work are enclosed in the Appendix. The first part of the Appendix includes preliminary results presented in several conferences, including: *i)* recognition of road noise sources with autocorrelation parameters [30] and probability density estimation (outcome of the internship carried out at the EU Joint Research Centre); *ii)* environmental sound recognition with MPEG-7 low level descriptors; and *iii)* comparison of machine learning techniques for soundscape recognition. The second part shows several works that, despite not focusing on the signal feature extraction, they are interesting due to their application to specific problems, such as: *i)* combination of audio and video parameters for geo-tagging online videos; *ii)* combination of supervised and unsupervised machine learning for environmental sound event recognition (outcome of the research stay at the Acoustics Department of Ghent University); and *iii)* water sound event detection for assisting doctors in the diagnosis and follow-up of dementia illnesses (outcome of the research stay at Institut de Recherche en Informatique de Toulouse).

Figure 5 shows a block diagram representing the sequence of all the research work developed under the framework of this thesis.

Legend:



³ For simplicity, this symbol is also used to represent the courses followed at La Salle-Universitat Ramon Llull as part of the PhD. The courses introduced several topics which were useful during the development of the research, such as: Artificial Intelligence, Neural Networks, Multimedia Communications or Research Methodologies.

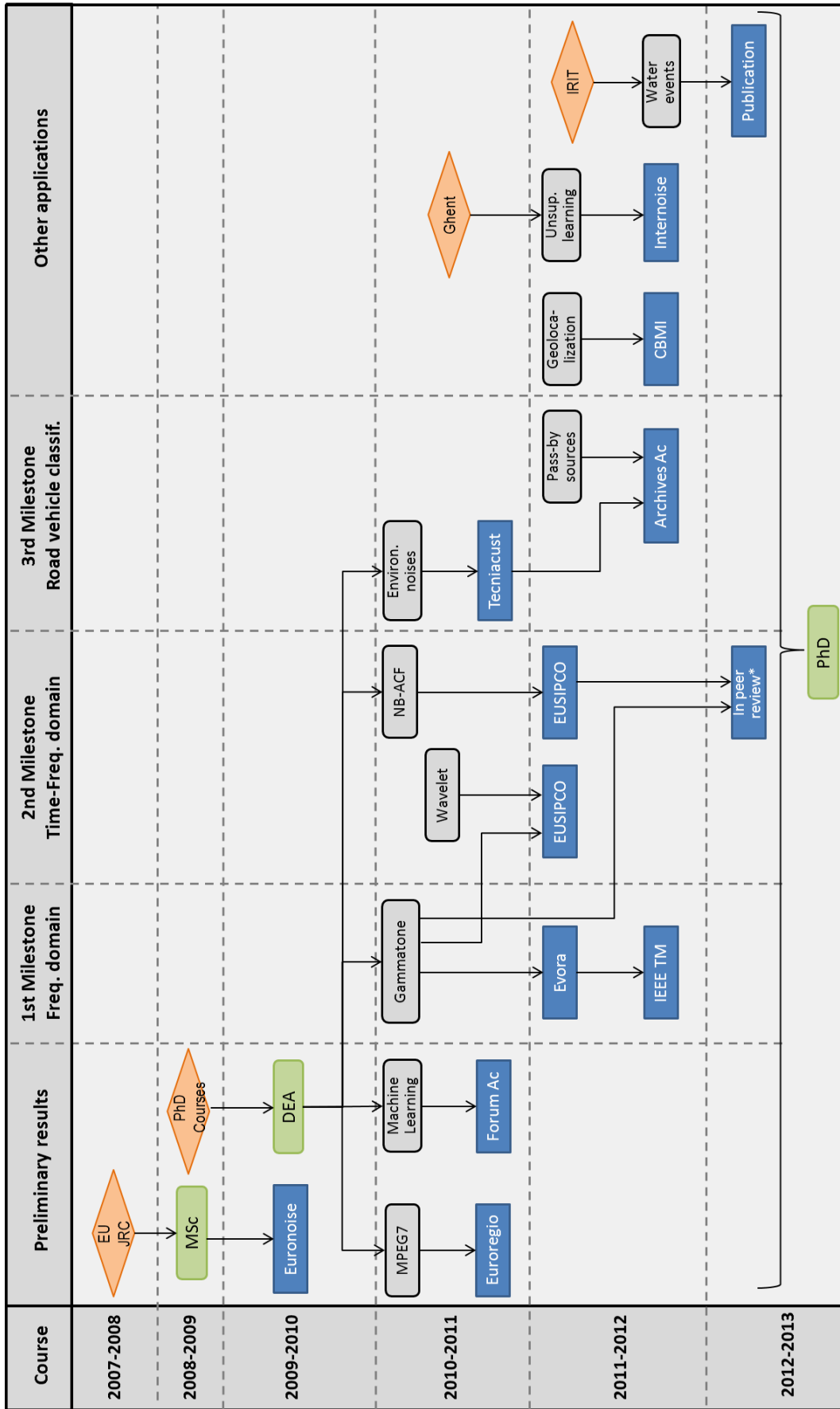


Figure 5. Flow chart of the research work performed.
* Paper in peer review, pending of final acceptance.

3. Milestone 1: Frequency-domain signal features

3.1. Introduction

The first Milestone attempts to enhance the modelling of the sound signal characteristics in the spectral domain.

MFCC, which are considered the standard de facto in audio recognition, make use of Mel filters. These filters take the name from the Mel scale frequency wrapping [31], generally presenting a triangular amplitude transfer function. This perceptual scale was obtained from human listening tests consisting on comparing pitched signals. This approach matches the characteristics of speech phonemes or music notes, but it may not fit well for environmental sounds (see Section 2.1).

The proposed solution consists on maintaining the effective computation scheme from MFCC but changing the Mel filter bank by a Gammatone filter bank, yielding the so-called Gammatone Cepstral Coefficients (GTCC). Gammatone (GT) filters were originally designed to model the human auditory spectral response, given its good approximation in terms of impulse response, magnitude response and filter bandwidth [32]. Another reason is that an n th-order GT filter has an efficient digital implementation, since it can be approximated by a set of n first-order GT filters placed in cascade [33].

The Gammatone filters are also characterized by the Equivalent Rectangular Bandwidth (ERB), a psychoacoustic measure of the auditory filter width at each point along the cochlea [32]. Specifically, ERB bands model the spectral integration derived from the channelling effectuated by the inner hair cells, which send signals of a certain bandwidth to the brain.

Gammatone filters with ERB bands have been previously employed for speech recognition [34], [35] and speaker identification [36]. The contribution of this thesis resides on the

adaptation of the GT filters to the characteristics of environmental sounds while integrating them in the well-known MFCC computation scheme. Hereafter, the main points of this first contribution are summarized:

- GT filter bank
 - The GT filter bank is adapted to the characteristics of environmental sound signals.
 - The chosen GT filter bank configuration employs 48 fourth order filters, with an extended bandwidth between the minimum audible (20Hz) and the Nyquist Frequency (11 KHz in the performed experiments), and follows the Glasberg&Moore ERB model.
- Experimental evaluation
 - Experimental work is conducted on 2 different corpora, comprising 15 general environmental sounds and 15 audio scenes, respectively.
 - Sound classification is performed with four different machine learning techniques: Decision Tree, K-Nearest Neighbour, Neural Network and Support Vector Machine.
- Results
 - When compared to other state of the art features, GTCC obtain the highest averaged classification rates (+2.5% with respect to MFCC and +3.2% with respect to MPEG-7).
 - Likewise, the adapted GTCC outperform previous versions from speech and speaker recognition.
- Analysis of results
 - A class-based analysis is performed on the two best performing signal features: GTCC and MFCC. Sounds containing particular components at low frequency bands (e.g., fountain, wind, certain birds) or speech-like sounds (e.g., bar, pedestrian street, crowd) achieve the highest improvements when employing GTCC.
 - The improvement achieved is due to a better representation of the signal spectral characteristics, (especially at low frequencies) thanks to the smoothed GT filter magnitude and, especially, to the higher ERB filter resolution at low frequency bands (see Figure 6).
 - Finally, experiments are repeated with an alternative cross validation scheme, where train and test sets are composed of recordings from different locations and employing different microphones. Also in those conditions GTCC outperform MFCC, hence showing greater generalization capabilities.

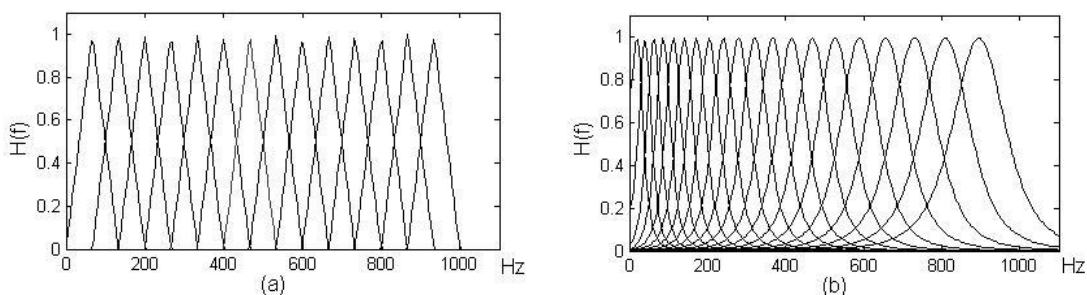


Figure 6. (a) Magnitude transfer functions of the MEL filters with a central frequency below 1 KHz; (b) Magnitude transfer functions of GT filters with a central frequency below 1 KHz.

3.2 Publications Milestone 1

Preliminary results were published in the conference Acústica Évora'12, and the complete results and conclusions commented below were then published in the journal IEEE Transactions on Multimedia. Both original articles are enclosed in the next section.

3.2.1 Publication Acústica Évora 2012

Title: “Análisis de la señal acústica mediante coeficientes cepstrales bio-inspirados y su aplicación al reconocimiento de paisajes sonoros”

Authors: Xavier Valero, Francesc Alías

Published in: Proceedings of Acústica Évora 2012, Évora (Portugal) [\[PDF\]](#)

Date of publication: October 2012.

3.2.2 Publication IEEE Transactions on Multimedia

Title: "Gammatone Cepstral Coefficients: Biologically-Inspired Features for Non-Speech Audio Classification"

Authors: Xavier Valero, Francesc Alías

Published in: IEEE Transactions on Multimedia, vol. 14, no. 6. [\[LINK\]](#)

Date of publication: December 2012.

4. Milestone 2: Time-frequency signal features

4.1 Introduction

The second Milestone seeks to model the environmental sound signals also taking into account the time domain information. However, in this Milestone the time-domain is not the exclusive domain of parameterization, since in preliminary tests we observed that signal features that only take into account the temporal characteristics of the signal (e.g., Short Time Energy or Zero Crossing Rate) are insufficient to achieve a good sound recognition. On the contrary, in this Milestone spectral information will also be included, coming up with the so-called spectro-temporal signal features. Two different contributions related to two target environmental sound types are proposed in this sense, which are next detailed.

4.1.1 Gammatone Wavelet features

The first milestone contribution is addressed to improve the classification of short duration and time-varying sounds. The environmental sounds holding those characteristics are frequently found in surveillance applications (e.g., glass breaking, gunshots, etc.) and tele-assistance (e.g., keys falling, door knocking, water tap dripping, etc.).

Wavelet techniques are a typical choice when dealing with the recognition of such type of sounds [24], [37], [38], given their potential to detect sudden changes in the sound signals.

The Wavelet analysis is strongly influenced by the Wavelet mother function used, i.e., the base of functions employed to represent the signal. Previous works on audio recognition employed

different Wavelet mother functions, such as Morlet [39], Coiflet [40] or, more typically, Daubechies [41].

Rather than following the same approach, this work proposes to employ Gammatone functions (which already showed good performance in the previous Milestone) as mother functions of the Wavelet analysis.

Two main characteristics define the Wavelet analysis [39]: *i)* the irregularity and asymmetry of the mother functions, so as to better capture abrupt signal changes; and *ii)* the variable length of the analysis windows to adapt to the signal frequencies: long windows track the low frequencies and short windows track the high frequencies.

Gammatone functions naturally hold these two characteristics: they are irregular and, by varying their central frequency, their duration is modified. However, in order to use them as mother functions, they should accomplish Wavelet admissibility conditions. That was done by adding an exponential term to the function so as to cancel the D.C. component, yielding the Gammatone Wavelet mother functions.

Finally, the scalar product of the sound signal with shifted Gammatone Wavelet mother functions is calculated, yielding the Gammatone Wavelet (GTW) signal features.

The proposed signal features are tested with a corpus of environmental surveillance-related sounds. Hereafter, the main points of the contribution are summarized.

- Framework
 - Wavelet analysis is chosen since it matches well the characteristics of surveillance-related sounds.
 - Gammatone functions are selected given the good performance shown in extracting the spectral characteristics of environmental sounds (see Section 3).
 - Both are combined yielding the proposed Gammatone-Wavelet features.
- Experimental evaluation
 - Experimental work consists on classifying a set of surveillance-related sounds into one of the 6 predefined categories making use of Support Vector Machines, as in [29].
 - Two versions of GTW features are computed, varying the number of Gammatone Wavelet mother functions employed: either seven (GTW-7) or sixteen (GTW-16). The baseline consists on the Discrete Wavelet Coefficients (DWC) computed with Daubechies mother function.
- Results
 - Both GTW-16 and GTW-7 outperform DWC in noiseless conditions (+4.6% and +3.7%, respectively).
 - In noisy conditions, the performance of the signal features varies with the SNR level. For $SNR > 0$ dB, GTW-16 is the best performing feature. For SNR between -5dB and -10dB, GTW-16 and DWC show an equivalent performance, whereas

for very adverse noise conditions (SNR<-10dB), GTW-7 seems to be the most robust signal feature.

- Finally, it is to note that combining the signal features yields the highest classification accuracies, thus suggesting that the baseline DWC and the proposed GTW features may be complementary.

4.1.2 Narrow-Band Autocorrelation features

The second milestone contribution in the spectro-temporal domain analysis tackles the classification of another type of environmental sound signals: those which are originated by several sound sources that coexist in a given location. *Soundscape*, *sound scene*, *audio scene* or *auditory scene* are typical names for referring to that concept.

In this field, it is important to come up with signal features that gather the characteristics of coexisting sound sources. The traditional MFCC (like other traditional spectral-based features) might fail in describing noise-like signals with strong temporal domain signatures (e.g., insects chirping while raining) [14], [26]. Thus, a more thorough and detailed analysis is needed.

The proposed signal features take their inspiration from two of the previous described techniques. First, the autocorrelation function analysis of a set of narrow-band signals, a technique previously used in speech segregation [42], [43]. This technique enables a detailed analysis of the signal, since one autocorrelation function is computed for each of the narrow-band signals at the output of a filter bank. Second, the autocorrelation function is parameterised by means of a set of perceptually-based parameters related to acoustic phenomena [30]: *i*) $\Phi(0)$, signal loudness; *ii*) τ_l , perceived pitch; *iii*) Φ_l , strength of the perceived pitch; and *iv*) τ_e , signal periodicity. These parameters, although computed on broadband signals, were used in previous works to describe environmental noise sources, such as road traffic sources [44], trains [45] or aircrafts [46]. On the contrary, the proposed technique computes these parameters for each autocorrelation function extracted from the different narrow-band signals. As a result, the so-called Narrow-Band Autocorrelation Function (NB-ACF) features are introduced. The NB-ACF ease the representation of coexisting sound sources given their complexity and higher degree of detail provided in the resulting feature vector.

Two publications have been written explaining and evaluating the proposed signal features.

The first one is an article presented at EUSIPCO 2012 conference, where the proposed NB-ACF features were presented and first evaluated. Hereafter, the main ideas of that piece of research are summarized.

- Focus
 - The NB-ACF signal parameterisation technique is presented.
 - The equations of the parameters extracted from each ACF ($\Phi_j(0)$, $\tau_{l,j}$, $\Phi_{l,j}$ and $\tau_{e,j}$) are defined.

- Experimental evaluation
 - Experiments are carried out by employing a corpus of 15 audio scenes. The parameterised sound samples are first compacted using PCA and then are classified using either KNN or SVM, as machine learning techniques.
 - The proposed features are compared to other time-domain analysis features (in that work, feature combination of STE and ZCR), frequency-domain features (MFCC) and time-frequency domain features (DWC).
- Results
 - The NB-ACF derived features achieve the best performance among all signal features, both with KNN and SVM classifiers.
 - With KNN classifier, the averaged classification rate using NB-ACF is 90% (2.6% higher than MFCC and 3.8% higher than DWC).
 - With the SVM classifier, NB-ACF features attain an averaged classification rate of 91% (2% higher than MFCC and 6.4 higher than DWC).
 - The time-domain features (STE+ZCR) attain a weak performance both with KNN and SVM (respectively, 47.5% and 41.4% of averaged classification accuracy).
 - The reduction of misclassifications when employing NB-ACF is particularly noticeable in audio scenes where changes of sound events over time are more pronounced (e.g., *office*, *library* and *classroom*).

The second article is an extension of the first one, providing a more detailed description of both the background and the computation of the proposed NB-ACF features. In addition, the technique is updated with the recent results from Milestone 1 (specifically, with the filter bank used based on Gammatone functions), and also the evaluation procedure is more robust, including an additional cross validation scheme. In the next lines, the main points of this research work are described.

- New contributions to NB-ACF computation
 - With respect to the previous article, the technique is improved by substituting the Mel filter bank employed to obtain the narrow-band signals by a GT filter bank with ERB bands (according to results obtained from Milestone 1)
 - In addition, a new parameter is added to parameterise ACF signals: the Autocorrelation Zero Crossing Rate (AZCR), which, according to previous works [47], is useful to discriminate between voiced and unvoiced signals.
 - The algorithms to compute the five ACF parameters ($\Phi_j(0)$, $\tau_{1,j}$, $\Phi_{1,j}$, $\tau_{e,j}$ and $AZCR_j$) are described in detail.
- Experimental evaluation
 - Whereas the corpus used is the same as in the previous work, two machine learning techniques are added in the experimentation: GMM and NN.

- Only the MFCC (the best performing feature in the former article) is included in the experimental comparison.
- Both versions of the proposed features, employing a Mel filter bank (NB-ACF-Mel) and a GT filter bank (NB-ACF-GT) are compared.
- The signal features derived from the Autocorrelation function of broadband signals (BB-ACF) are included as baseline.
- Two cross-validation schemes are followed to compose the train and test sets.
 - The first one randomly divides the data into two data sets (train and test), no matter the origin they have.
 - The second one divides the data by taking into account the location where they were recorded. In that way, it is ensured that sound samples recorded at the same place (even in different days) are only included in one of the sets. This experiment is called "evaluation in unknown locations" and allows evaluating the generalization capabilities of the recognition system.
- Results
 - The BB-ACF obtains the worst performance among the signal features compared, given their lower complexity than the NB-ACF.
 - Between the two NB-ACF versions, the one considering GT filter bank achieves higher classification accuracies, in agreement with the results obtained in Milestone 1.
 - The proposed NB-ACF-GT features yield an averaged classification rate 4.5% higher than MFCC. This improvement rate increases with respect to the first NB-ACF related article, validating the effectiveness of the introduced technique improvements (i.e., GT filter bank and AZCR parameter).
 - The improvement is even larger (+5.6%) when the features are tested in unknown locations, thus demonstrating the robustness of the proposed features.
- Results analysis
 - The results of the class-based analysis agree with those obtained in the previous publication: *office*, *library* and *classroom* are the audio scenes attaining the highest improvements when considering NB-ACF-GT features.
 - The improvement of the recognition results with respect to the standard MFCC are attributed to the better modelling of the audio scenes thanks to the complementarity of the spectro-temporal features derived from NB-ACF analysis.
 - For better understanding the complementarity of the NB-ACF features, specific examples of audio scenes modelled with the different ACF parameters are shown and discussed.

4.2 Publications Milestone 2

The three original articles related to Milestone 2 are enclosed in the next sections. The first one presents the GTW features and the remaining ones the NB-ACF features.

4.2.1 Publication EUSIPCO-GTW

Title: “Gammatone Wavelet features for Sound Classification in Surveillance Applications”

Authors: Xavier Valero, Francesc Alías

Published in: Proceedings of EUSIPCO'12 [\[PDF\]](#)

Date of publication: August 2012.

4.2.2 Publication EUSIPCO-NBACF

Title: "Classification of Audio Scenes Using Narrow-Band Autocorrelation Features"

Authors: Xavier Valero, Francesc Alías

Published in: Proceedings of EUSIPCO'12 [\[PDF\]](#)

Date of publication: August 2012.

4.2.3 Publication in peer review

Title: "Narrow-Band Autocorrelation Function Features for Automatic Acoustic Environment Recognition"

Authors: Xavier Valero, Francesc Alías

In peer review

5. Milestone 3: Application to road vehicle pass-by classification

5.1 Introduction

The third milestone is focused on studying a particular application of environmental sound classification: the monitoring of environmental noise. Behind this name we find noises originated by transport, industrial or recreational activities that affect citizens' quality of life, besides involving harmful health effects [9], [10].

This work lies within the Environmental Noise Directive (END) framework, an European Directive whose aim is to inform the public about their exposure to noise and to draw up appropriate action plans so as to prevent the harmful effects derived from their exposition to noise [7]. In compliance with the END, the member states of the European Union are required to report the noise levels by means of strategic noise maps from their main cities, transport infrastructures and industrial sites.

To produce such noise maps, measurements in complex urban environments need to be conducted, with the presence of noise from diverse origins, such as road traffic, railway traffic, aircrafts, industries, etc. In this sense, the implementation of environmental noise recognition systems may provide with an automatic transcription of the types of noise sources present on a certain location and their contribution to the measured overall noise level. As a result, precise strategic noise maps could be obtained and, consequently, the action plans designed to reduce or prevent high levels of environmental noise could be more efficiently addressed.

Following this thesis scope, the research work has been focused on the classification of the environmental noise sources. Two papers have been written to tackle this goal.

The first one is a conference paper that shows the preliminary results on this topic, mainly focused on the signal feature selection for the problem at hand. The main results and conclusions are next summarized.

- Scope
 - An environmental noise classification system based on Neural Networks is proposed.
 - Thirteen signal features are tested so as to select the optimal one for the problem at hand.
 - Time domain, frequency domain, time-frequency domain, linear prediction and Wavelet signal features are included.
- Experimental evaluation
 - A recording campaign was carried out in different urban locations so as to compose the corpus database.
 - This corpus is composed of audio samples from aircrafts (both landing and taking off operations), trains (including different types of trains), industry and road traffic (differentiating between light vehicles, heavy vehicles and motorbikes).
 - Principal Component Analysis is applied in order to compact the signal feature data and to extract the most significant information before applying the Neural Network.
 - Next, a Backpropagation Neural Network (NN) with one hidden layer performs the classification of the environmental noise samples.
- Results
 - A first group of four signal features (i.e., ZCR, Spectral Roll-Off, Spectral Centroid and STE) show very poor classification accuracies, with classification rates below 50%.
 - A second group of five features obtain average classification accuracies, with classification rates around 75%. In this group, we find Sub-Band Energy Ratio, Spectral Flatness, Linear Predictive Coefficients, Linear Predictive Cepstral Coefficients and Mel Frequency Discrete Wavelet Coefficients.
 - The best classification accuracies (around 90%) are obtained (in this order) by two frequency-domain features (i.e., MPEG-7 and MFCC), one linear prediction feature (Perceptual Linear Predictive) and one time-frequency domain feature (DWC).
- Analysis of results
 - A class-based analysis shows that road vehicle noise sources (i.e., light vehicles, heavy vehicles and motorbikes) are the environmental noise sources that yield the worst classification accuracies.
 - The confusion matrix indicates that the most frequent misclassifications are produced between heavy vehicles and motorbikes (17%) and between light and heavy vehicles (11.1%).

The second paper starts from the results obtained in that work. Its main aim is to improve the classification of road vehicle noise sources, which showed the worst classification accuracies in the first paper. For that purpose, the time evolution characteristics of the road vehicle's pass-by is taken into account. The paper argues that, due to both the Doppler physical effect and the multi-source composition of the road vehicles, the pass-by might be divided into three phases: *approaching*, *passing-by* and *receding*. A classification scheme is proposed to take into account the characteristics of such noise sources and their temporal evolution. On the one hand, it uses a hierarchical structure, so as to separately analyse the road vehicle noise sources from the rest. On the other hand, road vehicle pass-bys are identified and separated into the three aforementioned phases. Each phase is classified independently, and the final decision is taken by a simple majority vote. In case of draw, the class identified in the central pass-by is selected (see the paper for further details).

In addition, four new machine learning techniques besides NN are considered, including the Hidden Markov Models (HMM). Finally, listening tests are carried out in order to compare the system performance to the human recognition ability for the same task. The main ideas and findings are next summarized.

- Scope
 - A classification scheme is proposed to improve the classification of road vehicle noise sources.
 - It takes into account the time evolution of the road vehicles pass-by by differentiating three phases: *approaching*, *passing-by* and *receding*.
- Experimental evaluation
 - The corpus employed is the one employed in the former conference paper.
 - Thirteen signal features, and four machine learning techniques (i.e., DT, KNN, GMM and NN) are evaluated for the problem at hand, yielding 52 possible combinations for the classification system.
 - In addition, the combination of several successful signal features is also evaluated.
- Results
 - Signal feature and machine learning technique selection
 - The DT is the machine learning showing the worst performance.
 - The remaining ones yield comparable classification accuracies.
 - The impact of signal features on the results does not show significant changes with respect to the results obtained in the previous work.
 - MFCC (as signal features) and GMM (as machine learning technique) is the combination chosen for the rest of the experiments, considering both their good classification accuracy (89.5%) and their reduced computational cost.
 - Validation of vehicle pass-by phases consistency
 - A preliminary experiment consisting on classifying short samples that contain only one pass-by phase is conducted.

- Four of the most common misclassifications are produced between samples of different vehicles at the same pass-by phase, whereas only one is produced between different pass-by phases of the same vehicle.
- These results suggest that the pass-by phase is more discriminative than the type of vehicle itself for the classification, thus validating the proposed approach of dividing the vehicle pass-by in three independent phases.
- Performance of the proposed classification scheme
 - The averaged classification rate using the proposed classification scheme is 92.5%, which is 3% higher than the 89.5% yielded by the same MFCC+GMM combination with a traditional flat scheme.
 - Both the hierarchical structure and the pass-by differentiation cause a positive impact on the results.
 - The classification confusions produced between light and heavy vehicles are reduced to the half (about 8%), and also the confusions between motorbikes and heavy vehicles.
- Comparison to HMM
 - The proposed classification structure is compared to HMM, as a baseline, since it is a machine learning technique that inherently takes into account the time evolution of the features when trained.
 - Among the different HMM configurations tested, the left-to-right HMM with 3 states yields the highest averaged classification accuracy, which is in concordance with the proposed classification scheme.
 - When compared to HMM, the proposed classification scheme achieves an equivalent accuracy performance but with a significant lower computational cost.
- Listening tests
 - Listening tests are carried out to refer the system performance to human recognition ability.
 - Thirty subjects completed two different listening tasks.
 - Each task consisted on recognising 60 noise samples extracted from the same corpus used in the previous experiments.
 - In the first task, the noise samples contained all the three pass-by phases.
 - In the second task, the noise samples were shorter and contained only one out of the three phases.
 - An average non-trained human listener attains recognition accuracy significantly lower than the proposed system:
 - 10% lower for samples containing the whole vehicle pass-by
 - 25% lower for samples containing only one phase of the vehicle pass-by.
 - The results highlight:
 - The excellent recognition accuracy achieved by the proposed system.
 - The need of specific listening training for humans if they aim to achieve comparable recognition accuracy to the already trained automatic classification system.

5.2 Publications Milestone 3

The two publications within this Milestone are enclosed in the next sections.

5.2.1 Publication Tecniacustica

Title: “Monitorización Automática de Fuentes de Ruido Ambientales”

Authors: Xavier Valero, Francesc Alías

Published in: Proceedings of Tecniacustica’12 [\[PDF\]](#)

Date of publication: October 2011.

5.2.2 Publication Archives Acoustics

Title: "Hierarchical Classification of Environmental Noise Sources Considering the Acoustic Signature of Vehicle Pass-Bys

Authors: Xavier Valero, Francesc Alías

Published in: Archives of Acoustics [\[LINK\]](#)

Date of publication: December 2012.

6. Discussion

In this section, all the results obtained in this research work are evaluated and discussed globally. First, we refer directly to the three posed Milestones and the corresponding achieved results. Next, we comment several issues concerning environmental sound classification based on the acquired experience during the conducted research.

6.1 Results extracted from the Milestones

The objective of determining signal features for environmental sound classification has been approached from a double perspective, considering in both cases the way humans perceive the sound signals: firstly, studying just the spectral domain of the signal, and then, adding the time dimension. The results of both approaches are discussed in the following paragraphs.

6.1.1 Milestone 1

Regarding the spectral representation of sound signals, successful results have been achieved by incorporating the GT filter bank (that models the cochlear spectral analysis) to the computation of the Cepstral coefficients, yielding the GTCC. According to our investigations, the achieved improvement with regards to MFCC has two main causes. Firstly, the smoothed shape of the GT filter's magnitude transfer function increases the overlap between neighboring filters, thus minimizing the loss of signal spectral information with respect to the triangular Mel filters (which are aligned from its triangle base to the central frequency of the adjacent filter). And secondly, the filter bank resolution at low frequencies determined by the ERB frequency scale (used in GTCC computation) is much higher than the one yielded by the Mel scale (used in MFCC computation). This is the primary reason why GTCC gather more precisely the particular spectral components at low frequencies, which have been of particular relevance for classifying non-speech audio signals.

Therefore, the application of GTCC may be extended to the recognition of any kind of non-speech audio signal. As example, we show here the results of applying GTCC to the hierarchical scheme proposed to classify environmental noise sources (3rd Milestone, see Section 5.2.2), which could not included in the corresponding paper at that time. As detailed in Table 2, the GTCC outperform the MFCC in 6 out of the 6 cases (employing DT, KNN, NN, HMM, GMM and GMM with the proposed hierarchical scheme). Hence, it is proved again the convenience of using the proposed GTCC. Even though the accuracy increase is not very high (in this case, 2% as result of averaging the classification rates for all 6 cases), it always provides an advantage when compared to the traditional MFCC, while preserving a reduced computational cost.

Machine learning	GTCC	MFCC
DT	80.53	76.83
GMM	91.25	89.50
GMM (hierarchical scheme)	93.32	92.50
KNN	90.37	87.83
NN	92.22	89.49
HMM	93.52	92.96
Average	90.20	88.19

Table 2. Additional results from Milestone 3 employing GTCC features.

6.1.2 Milestone 2

As mentioned in the former section, the GTCC provide an improved classification when analysing *any* of the tested environmental sounds. The temporal information, even though is important as well, it should be analysed in a different way due to the different temporal characteristics of environmental sound signals. In this work, we have studied two contexts that present environmental sound signals with defined and particular temporal attributes: *i*) the surveillance-related sounds, which typically have a short duration and/or impulsive behaviour, and *ii*) the audio scenes, which are composed of multiple sound events that may coexist in time.

In the related audio-surveillance field, Wavelet analysis is typically employed. This choice is due to the temporal characteristics of the involved sounds (i.e., impulsiveness, short duration), which can be optimally analysed with time-frequency transforms. Hence, the basic idea is to convert the GT function (which according to the first Milestone it already provided a good spectral signal representation) into a time-frequency transform. The impulsive response of the GT function reminds the irregular and asymmetric response of a Wavelet mother function. Thus, the GT function has been integrated as a mother function of a time-frequency Wavelet analysis, yielding the GTW features. When compared to the DWC [29], the proposed features have yielded an average classification rate 4.5% higher. As other works from the state of the art on audio surveillance recognition, experiments have also been performed in noisy conditions [6], [29]. In this work, background noise from an urban area has been artificially added at different Signal-to-Noise Ratios (SNR). The results have revealed that the GTW slightly outperform the DWC for high SNR (0dB to 10dB), whereas an equivalent performance

is showed in the range of -5dB to -10dB SNR. However, for lower SNR (-15dB to -20dB), the proposed GTW features outperform the DWC, which suggest that GTW are especially robust in adverse noise conditions.

With regards to the audio scenes or *soundscapes*, they inherently present complex temporal characteristics as consequence of the simultaneity of the sound events they contain. This signal complexity is the main cause of the worse classification accuracy of these sound signals than single sound events, as noted in Section 3.2.2. Moreover, a class-based analysis (see Section A.3) have revealed that the soundscapes obtaining the lower classification accuracy were those presenting a wider variety of sound events coexisting in space and time. In order to improve the classification of such kind of soundscapes, a technique consisting in computing the autocorrelation function of band-filtered versions of the original signal has been developed. Furthermore, rather than employing the whole autocorrelation data (as in [42] and [43]), the autocorrelated signal has been analysed with a set of five parameters related to acoustic phenomena perceived by humans (i.e., loudness, dominant frequency, strength of the dominant frequency, reverberation and periodicity). The narrow-band analysis of the signal has provided two main advantages: *i)* identifying coexisting sounds with non-overlapped spectra; and *ii)* gathering very different aspects of the soundscape signals thanks to the five perceptual parameters (as opposed to MFCC or GTCC which only consider the energy of each spectral band). The results of the conducted experiments have demonstrated this theoretical advantage: the proposed features derived from the Narrow-Band Autocorrelation Function (NB-ACF) outperform in 5.6% the MFCC. The class-based analysis has revealed that the aforementioned soundscapes that obtain the worst classification accuracies with MFCC are also the ones that achieve the highest classification rate increase when employing the NB-ACF features (e.g., *office*: 9.8%, *library*: 4.1% and *classroom*: 3.1%). It should also be noted that from the two tested versions of the NB-ACF features, the one performing the spectral separation including a GT filter bank has yielded higher classification rates than the one using a Mel filter bank. This result reinforces the conclusions about the spectral domain signal analysis drawn up in Section 4.2.

6.1.3 Milestone 3

The third Milestone does not specifically focus on the signal feature extraction but highlights how the general technique for environmental sound signals may be adapted for specific applications or situations. In this case, human perception has also been taken into account: the sound perception of a road vehicle approaching is noticeably different to that when the vehicle is just passing-by or receding. This perceived sensation, which has a physical explanation (i.e., the Doppler Effect and the multi-source characteristic of road vehicle noises), has been used to design the classification scheme. Three independent machine learning techniques perform the classification of the approaching, passing-by and receding phases of the vehicle pass-by, respectively. The adapted classification scheme achieves a significant reduction (about 8%) in the confusions between light and heavy vehicles, which is attributed to the perceptible differences in the pass-by phases from both noise sources.

6.2 General comments

After discussing the results obtained from the three main Milestones achieved in this work, other general comments about environmental sound classification are next reviewed, based on the experience and the analysis of this kind of sounds during the conducted research.

All along this research work, a total of 6 machine learning techniques (i.e., DT, KNN, NN, GMM HMM and SVM) and 16 signal features have been tested. The selection of the machine learning technique and the signal feature has a very different impact on the classification results. Excluding the DT, the remaining learning algorithms have showed a quite similar performance. On the contrary, choosing one or another signal feature has a huge impact on the classification rates. Considering the same evaluation setup, the system might have yielded accuracy results from 50% up to 90%, depending on the signal feature employed. This fact totally agrees with the comments from the related literature [5], [14], [26], which suggest especially focusing on the signal feature extraction stage.

Another relevant issue refers to the testing protocol. Typically, different fold cross validation schemes are employed to split the available data into train and test sets, avoiding the same set of instances to be simultaneously used for train and test the classification system. However, the protocol on how distributing the data into different folds is not standardised. One possible approach is to compose a cross validation with as many folds as available sound data origins. That ensures that two instances recorded at the same location do not take part of the train and test sets simultaneously. An alternative approach consists in merging the sound data from all possible origins, thus randomly composing the different folds. In this case, the train and test data sets could include sound instances recorded at the same location. The latter approach (i.e., randomly splitting the data) represents the ideal situation where the recognition system is running on a device placed in a fixed location and can be trained beforehand with sound samples recorded at that specific location in a previous period of time. Examples of applications with fixed microphones are noise event recognition in fixed sports within an urban area or audio surveillance in home environments. On the contrary, the former approach (i.e., having as many folds as sound data origins) does not follow the assumption of having the recognition system in a fixed location. Thus, the system could not be trained beforehand with data from all the locations in which it will be working at. Context identification for portable devices or advanced hearing aid devices are two examples of applications that should follow this testing approach to validate the performance of the developed system in real operational scenarios. Rather than being incompatible, both testing approaches are valid, since they represent different application scenarios. Thus, both have been employed in this work.

Finally, in two of the contributions, some listening tests have been conducted in order to refer the classification accuracy achieved by the system to human sound identification ability (see Section 5.2.2 and Annex B.2). Although both considered the same type of environmental sound events (i.e., environmental noise sources that affect to human's health and life quality), they agree in pointing out that an average human listener is not able to perform as well as the classification system if he/she is not specifically and exhaustively trained for that purpose, as the system is. These results agree with those collected in [48], where also human listeners attained lower identification accuracy than the trained classification system. However, it

should be noted that if the listener receives that training, then he/she is able to outperform the classification system (see Section 5.2.2). Finally, the listening tests also reveal that the shorter is the sound event (in this case, the fragment of vehicle pass-by), the more complex is for the human listener to identify the sound. On the contrary, the classification system is able to keep quite similar accuracy levels. These results may suggest that the average human listener needs longer periods than the classification system to track the time and frequency envelope of the vehicle pass-by signal and identify the sound.

7. Conclusions and future lines

7.1 Conclusions

In this thesis, we have faced the classification of environmental sound signals, which may provide the machine with relevant information about the surrounding environment so as to improve human-computer interaction. Specifically, we have focused on studying and determining signal features adapted to the characteristics of such kind of signals, which differ from the more vastly studied and developed features for speech or music signals.

To tackle the problem, we have posed the study of the sound signals from two analysis domains: firstly, by only considering the spectral information, and afterwards, by considering both the spectral and temporal information. In both cases, we have attempted to come up with signal features that adapt to the human auditory response and the human perception of sound.

In the spectral domain, a better characterisation of the sound signals has been achieved thanks to employing the biologically-inspired Gammatone filters. These filters have been adapted to the characteristics of environmental sounds and have been integrated into the effective MFCC computation model. The proposed GTCC are able to better gather the spectral particularities of such kind of signals, especially at low frequency bands. This advantage results in a statistically significant improvement of the accuracy of the classification system. Although the increase is not very high (since the MFCC already work quite well), this improvement has been observed in practically all the conducted experiments (regardless of the corpus and machine learning technique employed), demonstrating the consistency of the proposed signal feature extraction technique.

Regarding the spectro-temporal domain, two main contributions have been achieved. On the one hand, the possibilities of Gammatone filters have been further exploited by combining them with the time-frequency Wavelet transform. The resulting features fit well the characteristics of surveillance-related environmental sounds, and have been able to improve

the classification performance of the Wavelet coefficients typically used in surveillance-related applications. Nevertheless, further should be investigated in order to understand the variability of the classification accuracy performance in function of the SNR level when compared to the baseline. On the other hand, the sound mixtures (typically found in soundscape signals) have been characterised with a signal parameterisation technique that combines three information aspects: the spectral domain, the temporal domain and the human perception of the acoustic waves. The perceptually motivated parameters extracted from the Autocorrelation function of narrow-band signals have shown an advantage respect to the traditional spectral-based features when dealing with complex soundscapes. The classification accuracy improvement achieved is notable, and significantly higher than the one obtained with GTCC features (see the first Milestone). The main drawback of the technique is its high associated computational cost, which currently makes its implementation on real-time applications unfeasible.

Moreover, a specific application of environmental sound recognition has been addressed: the classification of the environmental noise sources that affect to daily human's quality of life, which may produce harmful health effects at mid-long term. In preliminary works, road vehicle noise sources showed the largest difficulties to be correctly classified, given the high degree of similarity of their acoustic signatures. We have proposed a classification method that, as the human listener does, takes into account the approaching, passing-by and receding phases of the road vehicle pass-by. In consequence, the classification method is limited to cases of non-mixed vehicle pass-bys, which are only representative of certain non-busy roads and streets. Under these specific conditions, the proposed method (implemented with a hierarchical structure) is able to especially improve the classification of such road vehicle noise sources. In addition, the conducted listening tests have revealed that a regular human listener does not perform as well as the developed environmental noise classification system, thus stressing the high classification rates that it yields.

7.2 Future lines

The main future line triggered from the results obtained in this thesis is the deeper study of cases in which the environmental sound events appear overlapped in time and/or spectrum. In the current work, this issue has only been addressed with the NB-ACF parameterisation technique for the case of soundscapes. However, in future works the sound mixtures should be processed in different contexts, such as overlapping road vehicle pass-bys or simultaneous environmental sound events. A great room for further investigations is available in this direction, linking with the Computational Auditory Scene Analysis, a research field that specifically addresses the recognition of such complex scenes and aims at segregating the coexisting sound sources found in the environment [49].

Furthermore, rather than assuming already segmented sound signals, continuous streams of sound signals could be considered for specific applications, thus facing the event identification problem prior to the classification of the sound sources (see Section 1.3.2). By including this stage, the complete recognition process in a real scenario would be closed. Once the sound event identification stage is included, further evaluation metrics (such as precision, recall, F1

measure, ROC curves, etc.) could be computed, thus yielding a broader evaluation of the full recognition system.

Next, some future lines to the specific Milestones in which this work has been structured are provided. Regarding the first Milestone, the analysis with the recent dynamic compressive Gammachirp filters could be studied. These filters extend the Gammatone concept to model the variation of the filter shape that is observed with the changes of the stimulus level in simultaneous tone-in-noise masking [50]. Eventually, if these filters were suitable to represent environmental sound characteristics, they could be integrated to compose Cepstral features, similarly as done with Gammatone filters and GTCC.

With regard to the second Milestone, on the one hand, further studies should be conducted to better understand the performance of GTW features as function of the background noise conditions. Moreover, the GTW features could be used to address the problem of sound event detection (see Figure 1), prior to the classification one addressed in this work. Their potential to detect transients and signal changes in short lags of time would be especially useful to conduct this task. On the other hand, the computational cost required to compute the NB-ACF features should be reduced. For that purpose, the number of spectral bands in which the broadband signal is separated should be studied. The hypothetical reduction of the number of bands would dramatically decrease the computational time required for calculating NB-ACF features, which is highly important for implementing the technique in portable devices.

Regarding the third Milestone, the experiments conducted in this work have shown the consistency of considering independent vehicle pass-by phases (i.e., approaching, passing-by or receding) by themselves (see Section 5.2.2). This approach may be the basis for classifying the road vehicles in cases of nearly-simultaneous passing-bys (e.g., a motorbike receding while a car is approaching). Forthcoming research should go a step forward, considering as well the case of totally mixed noise sources (e.g., a truck and a car passing-by simultaneously). This step would dramatically enlarge the number of streets and roads where this recognition tool could be implemented.

To finalise, some other general recommendations to continue the research on the environmental sound recognition field are next discussed.

With regards to the machine learning block, it would be interesting to consider the combination of both supervised and unsupervised algorithms so as to take the best from each other. In this sense, a preliminary approach has already been performed by proposing a method for labelling (unsupervised) Self-Organising Maps by employing (supervised) SVM (see Annex B.2). The proposed labelling method automatizes the process, avoiding the manual labelling and reducing the time required to carry out this task. In the same vein, diverse signal features may be combined, and feature selection techniques applied to adapt and improve the system efficiency for specific applications.

Beyond the fundamental research on signal features and machine learning methods, if the recognition algorithms are aimed to be implemented in real applications, the computational cost of the whole recognition system should be carefully studied. Only in this way the requirement of working in real-time could be verified. In case it could be feasible, the range of

potential applications of the developed technique would grow considerably. Also, implementing the technique on real-world applications would allow verifying the algorithms performance in adverse conditions, such as with presence of non-stationary background noise.

Finally, it should be highlighted the need of creating a common environmental sound database that could be shared in the scientific community. This kind of databases exist since many years and are widely employed in the more developed speech scientific community [51]. Researchers on the field could benefit from disposing of standard corpora to conduct their experiments, thus easing the comparison of their studies with other previous approaches.

References

- [1] R.S. Ulrich, "Visual landscapes and psychological well-being", in *Landscape Research*, vol. 4, no. 1, pp. 17 – 23, 1979.
- [2] Z. Fu, G. Lu, K. Ming Ting, D. Zhang, "A survey of audio-based music classification and annotation", in *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303-319, April 2011.
- [3] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho and J. Huopaniemi, "Audio-based context recognition," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.
- [4] S. Ntalampiras, I. Potamitis, N. Fakotakis, "An adaptative framework for acoustic monitoring of potential hazards", in *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2009, January 2009.
- [5] A. Rabaoui, Z. Lachiri, N. Ellouze, "Towards an optimal feature set for robustness improvement of sounds classification in a HMM-based classifier adapted to real world background noise", in *Proc. 4th International Multi-Conference on Systems, Signals & Devices*, 2007.
- [6] M. Vacher, F. Portet, A. Fleury, N. Noury, "Challenges in the processing of audio channels for Ambient Assisted Living", in *Proc. IEEE International Conference on e-Health Networking Applications and Services* , pp. 330-337, 2010 .
- [7] EU Directive (2002), *Directive 2002/49/EC of the European parliament and the Council of 25 June 2002 relating to the assessment and management of environmental noise*, Official Journal of the European Communities, L 189/12 , July 2002.
- [8] ISO (2007), *ISO 1996-2:2007 Acoustics -- Description, measurement and assessment of environmental noise -- Part 2: Determination of environmental noise levels*.
- [9] W. Babisch, "Transportation noise and cardiovascular risk: Updated review and synthesis of epidemiological studies", in *Noise&Health*, vol. 8 no. 30, pp. 1-29, 2006.

-
- [10] F. Rasche, "Arousal and aircraft noise – Environmental disorders of sleep and health in terms of sleep medicine", in *Noise & Health*, vol. 6, no. 22, pp. 15-26, 2004.
- [11] L. S. Finegold, C. Stanley Harris, H.E. von Gierke, "Community annoyance and sleep disturbance: updated criteria for assessing the impacts of general transportation noise on people", in *Noise Control Engineering Journal*, vol. 42, no. 1, pp. 25-30, 1994.
- [12] P. Guyot, J. Piquier, R. André-Obrecht, "Water flow detection from a wearable device with a new feature, the spectral cover", in *Proc. International Workshop on Content-Based Multimedia Indexing*, pp. 1-6, 2012.
- [13] M. Büchler, S. Allegro, S. Launer, N. Dillier, "Sound classification in hearing aids Inspired by Auditory Scene Analysis", in *EURASIP Journal on Applied Signal Processing*, January 2005.
- [14] S. Chu, S. Narayanan, C.-C. Jay Kuo, "Environmental sound recognition with time-frequency audio features", in *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, no. 6, pp. 1142-1158, August 2009.
- [15] F. Beritelli, R. Gasso, "A pattern recognition system for environmental sound classification based on MFCCs and Neural Networks", in *Proc. 2008 IEEE Int. Congress on Signal Proc. And Communication Systems*, 2008.
- [16] Listening Talker Project - LISTA, [Available Online]: <http://listening-talker.org>
- [17] T. Zhang, C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification", in *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, 2001.
- [18] X. Sevillano, X. Valero, F. Alías, "Audio and video cues for geo-tagging online videos in the absence of metadata", 10th Workshop on Content-Based Multimedia Indexing (CBMI2012), Annecy (France), June 2012.
- [19] D.P.W. Ellis, K. Lee, "Minimal-impact audio-based personal archives", in *Proc. ACM workshop on Continuous archival and retrieval of personal experiences*, New York, USA, 2004.
- [20] A. Temko, C. Nadeu, J.-I. Biel, "Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07", CLEAR'07 Evaluation Campaign and Workshop, Baltimore, MD, USA, to appear in *Multimodal Technologies for Perception of Humans*, LNCS, Springer.
- [21] C. Zieger. An HMM based system for acoustic event detection. Second International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2007.
- [22] M. Sobreira Seoane, A. Rodriguez Molaes and J.L. Alba Castro, "Automatic classification of traffic noise", in *Proc. Acoustics'08*, (Paris, France), 6211-6266, 2008.
- [23] S. Ntalampiras, I. Potamitis, N. Fakotakis, "Automatic Recognition of Urban Environmental Sound Events", in *Proc. International Association for Pattern Recognition Workshop on Cognitive Information Processing*, Santorini, Greece, June 9-10, 2008.
-

-
- [24] D. Istrate, E. Castelli, M. Vacher, L. Besacier, J.F. Serignat, "Information extraction from sound for medical telemonitoring", in *IEEE Trans. Information Technology in Biomedicine*, vol. 10, no. 2, April 2006.
- [25] X. Zhuang, X. Zhou, MA. Hasegawa-Johnson, T.S. Huang, "Real-world acoustic event detection", in *Pattern Recognition Letters*, vol.31, pp.1543-1551, 2010
- [26] K. Umopathy, S. Krishnan, S. Jimaa, "Multigroup classification of audio signals using time-frequency parameters", in *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 308-315, April 2005.
- [27] B. Gygi, "Factors in the identification of environmental sounds", PhD. thesis, Indiana University, July 2001.
- [28] R.E. Bellman, "Dynamic Programming", *Courier Dover Publications*, ISBN 978-0-486-42809-3, 2003.
- [29] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, "Using one-class SVMs and Wavelets for audio surveillance", in *IEEE Trans. Information Forensics and Security*, vol. 3, no. 4, pp. 763-775, December 2008.
- [30] Y. Ando, "A theory of primary sensations and spatial sensations measuring environmental noise", in *Journal of Sound and Vibration*, vol. 241, no. 1, pp. 3-18, 2001.
- [31] S. S. Stevens, J. Volkman, E. Newman, "A scale for the measurement of the psychological magnitude pitch", in *Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185-190, 1937.
- [32] R. D. Patterson, J. Holdsworth, "A functional model of neural activity patterns and auditory images", in *W. A. Ainsworth (Ed.), Advances in Speech, Hearing and Language Processing, vol. 3 part B*. London: JAI Press, 1996, pp. 554-562.
- [33] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," Apple Technical Report #35, Apple Computer Library, Cupertino, CA 95014, 1993.
- [34] W. Abdullah, "Auditory based feature vectors for speech recognition systems", in *Advances in Communications and Software Technologies*. Greece: N.E. Mastorakis&V.V.Kluev, Editor. WSEAS Press, pp. 231-236, 2002.
- [35] Y. Shao, Z. Kin, D. Wang, "An auditory-based feature for robust speech recognition", in *Proc. ICASSP*, 2009.
- [36] Y. Shao, D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis", in *Proc. ICASSP*, 2008.
- [37] M. Cowling, R. Sitte, "Comparison of techniques for environmental sound recognition", in *Pattern Recognition Letters*, vol. 24, pp. 2895-2907, 2003.
- [38] N. Mclachlan, D. K. Kumar, J. Becker, "Wavelet classification of indoor environmental sound sources", in *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 4, no. 1, pp. 81-96, 2006.
-

-
- [39] C.S., Burrus, R.A. Gopinath, H. Guo, "Introduction to Wavelets and Wavelet Transforms: A Primer", New Jersey: Prentice Hall, 1998
- [40] B. Bradie, "Wavelet Packet-based Compression of Single Lead ECG", in *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 5, pp 493-501, May 1996.
- [41] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis", in *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961-1005, 1990.
- [42] M. Slaney, "The history and future of CASA", Chap. 13 of "Speech separation by humans and machines", Pierre Divenyi (Ed.), Kluwer Academic Publishers, pp. 199-211, 2005.
- [43] R.O. Duda, R.F. Lyon, M. Slaney, "Correlograms and the Separation of Sounds", in *Proc. Asilomar Conf. On Signals, Systems and Computers*, 1990.
- [44] K. Fujii, J. Atagi, Y. Ando, "Temporal and spatial factors of traffic noise and its annoyance", in *Journal of Temporal Design in Architecture and the Environment*, vol. 2 no. 1, pp. 23-41, 2002.
- [45] H. Sakai, T. Hotehama, Y. Ando, "Diagnostic system based on the human auditory-brain model for measuring environmental noise – An application to railway noise", in *Journal of Sound and Vibration*, vol. 250, no. 1, pp. 9-21, 2002.
- [46] K. Fujii, Y. Soeta, Y. Ando, "Acoustical properties of aircraft noise measured by temporal and spatial factors", in *Journal of Sound and Vibration*, vol. 241, no. 1, pp. 69-78, 2001.
- [47] H. Ghaemmaghami, B.J. Baker, R.J. Vogt, S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function". In *Proc. Interspeech 2010*, pp. 3118-3121, ISCA, (Makuhari, Chiba, Japan), 2010.
- [48] C. Couvreur, V. Fontaine, P. Gaunard and C.G. Mubikangiey, "Automatic classification of environmental noise events by Hidden Markov Models", in *Applied Acoustics*, vol. 54, no. 3, pp. 187-206, 1998.
- [49] D. Wang, G.J. Brown, "Computational Auditory Scene Analysis: Principles, Algorithms and Applications", New York: Wiley-IEEE Press, 2006.
- [50] T. Irino, R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank", in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2222-2232, 2006.
- [51] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus" Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, 1993.

Annex A: First studies

A.1 Probability density estimation for road noise source monitoring

Title: "Pattern recognition and separation of road noise sources by means of ACF, MFCC and probability density estimation"

Authors: Xavier Valero, Francesc Alías, Stylianos Kefalopoulos, Marco Paviotti.

Published in: Proceedings EURONOISE'09.

Date of publication: October 2009.

A.2 MPEG-7 parameters for environmental sound recognition

Title: "Applicability of MPEG-7 low level descriptors to environmental sound source recognition"

Authors: Xavier Valero, Francesc Alías.

Published in: Proceedings EUROREGIO'10.

Date of publication: September 2010.

A.3 Machine Learning techniques comparison for soundscape recognition

Title: "Comparison of Machine Learning Techniques for the Automatic Recognition of Soundscapes"

Authors: Xavier Valero, Pau Farré, Francesc Alías.

Published in: Proceedings FORUM ACUSTICUM'11..

Date of publication: July 2011.

Annex B: Other applications of environmental sound classification

B.1 Audio-based geo-tagging

Title: "Audio and video cues for geo-tagging online videos in the absence of metadata"

Authors: Xavier Sevillano, Xavier Valero, Francesc Alías.

Published in: Proceedings 10th Workshop on Content-Based Multimedia Indexing. [\[LINK\]](#)

Date of publication: June 2012.

B.2 SVM and SOM for urban sound recognition

Title: “Support Vector Machines and Self-Organizing Maps for the recognition of sound events in urban soundscapes”

Authors: Xavier Valero, Damiano Oldoni, Francesc Alías, Dick Botteldooren.

Published in: Proceedings INTERNOISE’12. [\[LINK\]](#)

Date of publication: August 2012.

B.3 Water sound event detection

Title: "Two-step detection of water sound events for the diagnostic and monitoring of dementia"

Author: Patrice Guyot, Xavier Valero, Julien Pinquier, Francesc Alías

In peer review

Date: December 2012.