Bayesian and Echoic Log-surprise for auditory saliency detection

by

Antonio Rodríguez Hidalgo

in partial fulfillment of the requirements for the degree of Doctor in
Multimedia and Communications

Universidad Carlos III de Madrid

Advisors:
Dr. Ascensión Gallardo Antolín
Dr. Carmen Peláez Moreno
Tutor:
Dr. Ascensión Gallardo Antolín

Leganés, November 2019

Some rights reserved. This thesis is distributed under a Creative Commons license Attribution-NonCommercial-No Derivatives.

ACKNOWLEDGEMENTS

Dedico este trabajo a mi familia y a todos los amigos y amigas que he conocido a lo largo de los años. Gracias por ser mi inspiración.

PUBLISHED AND SUBMITTED CONTENT

This work has some of its content available in the following papers:

1. [Rodríguez-Hidalgo et al., 2016] Rodríguez-Hidalgo, A., Gallardo-Antolín, A., and Peláez-Moreno, C. (2016). Towards aural saliency detection with logarithmic Bayesian Surprise under different spectro-temporal representations. *Proceedings of Iberspeech 2016*, pages 99–108.

   This item is partly included and extended in Chapter 5. Whenever material from this source is included in this thesis, it is singled out with typographic means and an explicit reference.

2. [Rodríguez-Hidalgo et al., 2018a] Rodríguez-Hidalgo, A., Peláez-Moreno, C., and Gallardo-Antolín, A. (2018). Echoic log-surprise: A multi-scale scheme for acoustic saliency detection. *Expert Systems with Applications*, 114:255 – 266. DOI: `https://doi.org/10.1016/j.eswa.2018.07.018`

   This item is partly included and extended in Chapters 6 and 8. Whenever material from this source is included in this thesis, it is singled out with typographic means and an explicit reference.

3. [Rodríguez-Hidalgo et al., 2018b] Rodríguez-Hidalgo, A., Peláez-Moreno, C., and Gallardo-Antolín, A. (2018). The robustness of echoic log-surprise auditory saliency detection. *IEEE Access*, 6:72083–72093. DOI: `https://doi.org/10.1109/ACCESS.2018.2882055`

   This item is partly included and extended in Chapters 6, 7 and 8. Whenever material from this source is included in this thesis, it is singled out with typographic means and an explicit reference.

4. [Rodríguez-Hidalgo et al., 2019] Rodríguez-Hidalgo, A., Peláez-Moreno, C., and Gallardo-Antolín, A. (2019). Auditory saliency detection based on information fusion by means of statistical divergences. *Manuscript submitted for publication.*

   This item is partly included and extended in Chapters 6, 7 and 8. Whenever material from this source is included in this thesis, it is singled out with typographic means and an explicit reference.

OTHER RESEARCH MERITS

The author of this thesis also collaborated in the development of the following research papers:

1. Rodríguez-Hidalgo, A., Peláez-Moreno, C., and Gallardo-Antolín, A. (2017). Towards multimodal saliency detection: An enhancement of audio-visual correlation estimation. *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, Oxford (UK), pages 438-443. DOI: https://doi.org/10.1109/ICCI-CC.2017.8109785.

ABSTRACT

Attention is defined as the mechanism that allows the brain to categorize and prioritize information acquired using our senses and act according to the environmental context and the available mental resources. The attention mechanism can be further subdivided into two types: top-down and bottom-up. Top-down attention is goal or task-driven and implies that a participant has some previous knowledge about the task that he or she is trying to solve. Alternatively, bottom-up attention only depends on the perceived features of the target object and its surroundings and is a very fast mechanism that is believed to be crucial for human survival.

Bottom-up attention is commonly known as saliency or salience, and can be defined as a property of the signals that are perceived by our senses that make them attentionally prominent for some reason.

This thesis is related with the concept of saliency detection using automatic algorithms for audio signals. In recent years progress in the area of visual saliency research has been remarkable, a topic where the goal consists of detecting which objects or content from a visual scene are prominent enough to capture the attention of a spectator. However, this progress has not been carried out to other alternative modalities. This is the case of auditory saliency, where there is still no consensus about how to measure the saliency of an event, and consequently there are no specific labeled datasets to compare new algorithms and proposals.

In this work two new auditory saliency detection algorithms are presented and evaluated. For their evaluation, we make use of Acoustic Event Detection/Classification datasets, whose labels include onset times among other aspects. We use such datasets and labeling since there is psychological evidence suggesting that human beings are quite sensitive to the spontaneous appearance of acoustic objects. We use three datasets: DCASE 2016 (Task 2), MIVIA road audio events and UPC-TALP, totalling 3400 labeled acoustic events. Regarding the algorithms that we employ for benchmarking, these comprise techniques for saliency detection designed by Kayser and Kalinli, a voice activity detector, an energy thresholding method and four music information retrieval onset detectors: NWPD, WPD, CD and SF.

We put forward two auditory saliency algorithms: Bayesian Log-surprise and Echoic Log-surprise. The former is an evolution of Bayesian Surprise, a methodology that by means of the Kullback-Leibler divergence computed between two consecutive temporal windows is capable of detecting anomalous or salient events. As the output Surprise signal has some drawbacks that should be overcome, we introduce some improvements that led to the approach that we named Bayesian Log-surprise. These include an amplitude compression stage and the addition of perceptual knowledge to pre-process the input signal.

The latter, named Echoic Log-surprise, fuses several Bayesian Log-surprise

signals computed considering different memory lengths that represent different temporal scales. The fusion process is performed using statistical divergences, resulting in saliency signals with certain advantages such as a significant reduction in the background noise level and a noticeable increase in the detection scores.

Moreover, since the original Echoic Log-surprise presents certain limitations, we propose a set of improvements: we test some alternative statistical divergences, we introduce a new fusion strategy and we change the thresholding mechanism used to determine if the final output signal is salient or not for a dynamic thresholding algorithm. Results show that the most significant modification in terms of performance is the latter, a proposal that reduces the dispersion observed in the scores produced by the system and enables online functioning.

Finally, our last analysis concerns the robustness of all the algorithms presented in this thesis against environmental noise. We use noises of different natures, from stationary noise to pre-recorded noises acquired in real environments such as cafeterias, train stations, etc. The results suggest that for different signal-to-noise ratios the most robust algorithm is Echoic Log-surprise, since its detection capabilities are the least influenced by noise.

RESUMEN

La atención es definida como el mecanismo que permite a nuestro cerebro categorizar y priorizar la información percibida mediante nuestros sentidos, a la par que ayuda a actuar en función del contexto y los recursos mentales disponibles. Este mecanismo puede dividirse en dos variantes: *top-down* y *bottom-up*. La atención *top-down* posee un objetivo que el sujeto pretende cumplir, e implica que el individuo posee cierto conocimiento previo sobre la tarea que trata de realizar. Por otra parte, la atención *bottom-up* depende exclusivamente de las características físicas percibidas a partir de un objeto y su entorno, y actúa a partir de dicha información de forma autónoma y rápida. Se teoriza que dicho mecanismo es crucial para la supervivencia de los individuos frente a amenazas repentinas.

La atención *bottom-up* es comúnmente denominada saliencia, y es definida como una propiedad de las señales que son percibidas por nuestros sentidos y que por algún motivo destacan sobre el resto de información adquirida.

Esta tesis está relacionada con la detección automática de la saliencia en señales acústicas mediante la utilización de algoritmos. En los últimos años el avance en la investigación de la saliencia visual ha sido notable, un tema en el cual la principal meta consiste en detectar qué objetos o contenido de una escena visual son lo bastante prominentes para captar la atención de un espectador. Sin embargo, estos avances no han sido trasladados a otras modalidades. Tal es el caso de la saliencia auditiva, donde aún no existe consenso sobre cómo medir la prominencia de un evento acústico, y en consecuencia no existen bases de datos especializadas que permitan comparar nuevos algoritmos y modelos.

En este trabajo evaluamos algunos algoritmos de detección de saliencia auditiva. Para ello, empleamos bases de datos para la detección y clasificación de eventos acústicos, cuyas etiquetas incluyen el tiempo de inicio (*onset*) de dichos eventos entre otras características. Nuestra hipótesis se basa en estudios psicológicos que sugieren que los seres humanos somos muy sensibles a la aparición de objetos acústicos. Empleamos tres bases de datos: DCASE 2016 (Task 2), MIVIA road audio events y UPC-TALP, las cuales suman en total 3400 eventos etiquetados. Respecto a los algoritmos utilizados en nuestro sistema de referencia (*benchmark*), incluimos los algoritmos de saliencia diseñados por *Kayser* y *Kalinli*, un detector de actividad vocal (VAD), un umbralizador energético y cuatro técnicas para la detección de onsets en música: NWPD, WPD, CD and SF.

Presentamos dos algoritmos de saliencia auditiva: *Bayesian Log-surprise* y *Echoic Log-surprise*. El primero es una evolución de *Bayesian Surprise*, una metodología que utiliza la divergencia de Kullback-Leibler para detectar eventos salientes o anomalías entre ventanas consecutivas de tiempo. Dado que la señal producida por *Bayesian Surprise* posee ciertos inconvenientes introducimos una serie de mejoras, entre las que destacan una etapa de com-

presión de la amplitud de la señal de salida y el pre-procesado de la señal de entrada mediante la utilización de conocimiento perceptual. Denominamos a esta metodología *Bayesian Log-surprise.*

Nuestro segundo algoritmo, denominado *Echoic Log-surprise*, combina la información de múltiples señales de saliencia producidas mediante *Bayesian Log-surprise* considerando distintas escalas temporales. El proceso de fusión se realiza mediante la utilización de divergencias estadísticas, y las señales de salida poseen un nivel de ruido menor a la par que un mayor rendimiento a la hora de detectar eventos salientes.

Además, proponemos una serie de mejoras para *Echoic Log-surprise* dado que observamos que presentaba ciertas limitaciones: añadimos nuevas divergencias estadísticas al sistema para realizar la fusión, diseñamos una nueva estrategia para llevar a cabo dicho proceso y modificamos el sistema de umbralizado que originalmente se utilizaba para determinar si un fragmento de señal era saliente o no. Inicialmente dicho mecanismo era estático, y proponemos actualizarlo de tal forma se comporte de forma dinámica. Esta última demuestra ser la mejora más significativa en términos de rendimiento, ya que reduce la dispersión observada en las puntuaciones de evaluación entre distintos ficheros de audio, a la par que permite que el algoritmo funcione *online*.

El último análisis que proponemos pretende estudiar la robustez de los algoritmos mencionados en esta tesis frente a ruido ambiental. Empleamos ruido de diversa índole, desde ruido blanco estacionario hasta señales pregrabadas en entornos reales tales y como cafeterías, estaciones de tren, etc. Los resultados sugieren que para distintos valores de relación señal/ruido el algoritmo más robusto es *Echoic Log-surprise*, dado que sus capacidades de detección son las menos afectadas por el ruido.

Contents

List of Figures

List of Tables

Glossary

B Circular buffer used for Welford's online algorithm during the computation of Bayesian Surprise

dth Depth of Echoic Log-surprise, defined as the number of Log-surprise blocks

f_c Analog frequency, defined in magnitudes of Hz

F_s Sampling frequency

k Frequency bin index from a spectro-temporal representation

L Length of the temporal window used to estimate the histograms for Echoic Log-surprise

L_{win} Length of the temporal window used to obtain a spectro-temporal representation

n_B Position index inside the circular buffer B

n_{fft} Number of frequency bins to compute the Short-Time Fourier Transform

n_{Mel} Number of filters defined according to the Mel scale

N Length of the circular buffer B for Bayesian Surprise and Log-surprise

N_1 For Echoic Log-surprise, it represents the memory of the first block of Bayesian Log-surprises, which is then used to compute the rest of N_z values depending on dth

P Precision

R Recall

T_{signal} Total duration of a signal, represented in frames

$x(m)$ Audio signal in the temporal domain

$X(k,n)$	Modulus of the spectro-temporal representation computed from $x(m)$
$X_c(k,n)$	Complex spectro-temporal representation computed from $x(m)$

Acronyms

AI	Anterior Insula
ASUN	Acoustic Saliency Using Natural statistics
AEC	Acoustic Event Classification
AEC/D	Acoustic Event Classification/Detection
AED	Acoustic Event Detection
ASM	Auditory sensory memory
Bhatta-N	Bhattacharyya distance computed for N distributions
CD	Complex Domain
CHIL	Computers In the Human Interaction Loop
CNN	Convolutional Neural Networks
dACC	dorsal Anterior Cingulate Cortex
DCASE	Detection and Classification of Acoustic Scenes and Events
DEMAND	Diverse Environments Multichannel Acoustic Noise Database
DFT	Discrete Fourier Transform
DKL	Kullback-Leibler divergence
EBR	Event-to-Background Ratio
EMD	Earth Mover's Distance
ERB	Equivalent Rectangular Bandwidth
FN	False Negative
FP	False Positive
GBVS	Graph Based Visual Saliency

HAS	Human Auditory System
HMM	Hidden Markov Model
Hz	Hertz
IEEE	Institute of Electrical and Electronics Engineers
JSD	Jensen-Shannon divergence
LRT	Likelihood Ratio Test
MATLAB	Computing environment that we used to program our experiments and algorithms
MEG	Magnetoencephalography
MFCC	Mel Frequency Cepstral Coefficients
MIR	Music Information Retrieval
MIT	Massachusetts Institute of Technology
MIVIA	MIVIA road audio events data set
ms	Millisecond
NWPD	Normalized Weighted Phase Deviation
PCA	Principal Component Analysis
PD	Phase Deviation
PNCC	Power-Normalized Cepstral Coefficients
ROC	Receiver Operating Characteristic curve
SF	Spectral Flux
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
stdev	Standard deviation
TN	True Negative
TP	True Positive
TVD	Total Variation distance
UPC-TALP	UPC-TALP database of isolated meeting-room acoustic events

VAD	Voice Activity Detector/Detection
ATOM	Visual Attention Topic Model
WPD	Weighted Phase Deviation

Chapter 1

Introduction

1.1 Motivation

Humans have developed a very rich perceptual system to be able to make sense of a complex environment. Indeed, we are able to select and combine several stimuli with different purposes, such as locating a source of food, listening to some enjoyable bird songs or panorama or even looking for potential threats. As a summary, we can say that perception plays a key role in the survival of human beings. Moreover, during the process of evolution, our brain has developed the necessary mechanisms to cope with the incoming sensory information, most of the times so vast and rich, that without them our brain would be unable to process and understand and therefore humans' capabilities to make decisions would be severely degraded.

Consequently, and in spite of the fact that our brain has a remarkable capacity to process data, the necessity to categorize and prioritize the incoming information is clear. The mechanism in charge of managing how this information is selected is commonly known as *attention*. As it will be explained in detail in Chapter 3, there are several ways to categorize attention. Generally speaking, there are two related types of attention that can be easily understood with two examples: first, when we focus our attention while trying to solve a particular task such as preparing a sandwich or finding a lost item and second, what captures our attention due to the peculiar characteristics of the stimulus, for example, a fire alarm or glass breaking event somewhere nearby. These two examples represent two well-known categories of attention: top-down and bottom-up, respectively, and more details about both of them are explained in Chapter 3. For the moment we will say that top-down is considered to be mostly task-driven, which means that a person tries to solve a certain task and his or her attentional resources are managed accordingly. On the other hand, bottom-up attention occurs when a certain external stimulus is perceived, which produces an automatic and fast response from our brain. The stimuli that have some properties or features that make them particularly prominent are said to be salient, and this level of salience depends strongly on the context in which they are perceived.

Understanding human attention is an active field of research, and many contributions have been proposed by specialists coming from areas of expertise such as neuroscience, psychology, computer science and engineering, among many others. The scope of this thesis is related to the progress made in the development of attentional models from the engineering perspective, whose main purpose is to emulate how a human participant would react when he or she perceives a specific stimuli, no matter whether visual, auditory, or other sensory modality.

In the last decade the body of work related to the particular case of visual attention has grown drastically. Specifically there have been several computational proposals designed to emulate visual attention and detect visual saliency. Particularly critical for their development was the proliferation of datasets obtained using eye-trackers, that allow to acquire attentional data (eye fixations) from human subjects for specific computer vision tasks. Such computational attention models evolved from *ad-hoc* algorithms biologically inspired in the human visual system, such as Itti's model [Itti et al., 1998] and Graph Based Visual Saliency (GBVS) model [Harel et al., 2006], to *machine learning* based algorithms, which after a process of training using adequate data and features are usually able to outperform the former. As a consequence, there has been a transition from unsupervised to supervised models, where labeled data enables the possibility to learn statistical models directly from data. This progress has been boosted with the resurgence of *artificial neural networks* and *deep learning* whose models are capable of crafting features automatically and, in combination with adequate classifiers and regressors, outperform the previous state-of-the-art models for visual attention [Kruthiventi et al., 2017; Kümmerer et al., 2017; Jia, 2018]. In exchange, the need of data to perform such computations has become imperative, since these models are very data hungry. More information about such an interesting area can be found in the MIT-Saliency benchmark [Bylinskii et al., 2019a].

There exist other modalities or combinations of them where saliency has been proposed for real-life applications, as it occurs with audio-visual saliency. At its essence, these algorithms combine signals from audio and video sources, extract features and compute saliency descriptors for each one of them, which are finally combined to solve a specific task. One example is video summarization, where an algorithm needs to extract relevant information from the frames of a movie, which might contain audio of different nature such as speech or music, as well as from the visual scene to determine which frames are more relevant. An example was proposed by [Evangelopoulos et al., 2013], which extracted not only auditory or visual saliency cues, but also processed the subtitles to measure relevance from textual information. Other authors proposed algorithms to predict eye-trajectories when participants were looking at conversations, as in [Coutrot and Guyader, 2015]. Alternatively, [Schauerte et al., 2011] implemented a model that allowed a robotic head to explore its surroundings to find what seemed perceptually relevant. The auditory part of their model was in charge of detecting salient sounds, as well as determining their location, whereas their visual stage located potential objects that might be of interest. The audio-visual system formed by these two stages helped to develop an exploration system, where a robotic head turned to locate the source of a sound when it heard it, no matter whether there were other visually prominent objects in the scene or not.

Even if there have been some works in the field of audio-visual salience, they tend to rely strongly on the visual modality and it seems that auditory information is used as complementary data that is helpful to improve the global performance of the systems. Despite the huge progress that has been made for visual attention models, it turns out that very few developments have been proposed for audio from a computational perspective. As it will be explained in Section 3.2, many experiments related to auditory attention have been performed with human participants, and our understanding about how acoustic signals are understood and how they affect to the management of attentional resources is better than at the beginning of the century. However, the algorithmic contributions proposed for this particular modality are far from being as mainstream and popular as their visual counterparts.

There are several reasons for such difference, but two of them seem to be particularly relevant: first of all, the availability of adequate datasets plays a crucial role, since their existence allows researchers to develop better models for more specific tasks. In the case of auditory perception, there is no such thing as an "ear-tracker", and to the extent of our knowledge there is no reliable way of measuring what human participants are attending to. As a consequence, there are no specific datasets for auditory attention, and researchers need to use creative experiments or proxies from similar areas of knowledge to measure the performance of their proposals. The second reason has to do with the previous state-of-the-art in auditory attention models. As it will be explained in Section 3.6, most of the models available nowadays are adapted from visual saliency algorithms. It does not mean that they are not reliable mechanisms to emulate attention, since it is theorized that sensory information is finally processed in a common area of the brain. However, some of the algorithms only change some of the early feature extractors of previous models, and keep using structures that were initially conceived for images.

From this exposition we can conclude that proposing a new attention model to detect auditory saliency can be challenging, due to the limitations of data and the lack of previous algorithms, but also an interesting line of research to propose new approaches, like the one presented in this dissertation.

1.2 Objectives

As it has been explained in the Motivation section, although some progress on the development of auditory attention algorithms exists, it is still far from that of visual attention and therefore, there is plenty of room to propose new models. The main objective of this thesis is the development of an auditory saliency detection algorithm that can improve the results of previous proposals. In order to cope with the difficulties that might arise during the development process, a thorough analysis of bibliographical resources is performed. The field of auditory saliency has been studied mainly in the area of psychology and neuroscience, where plenty of participant-based experiments have been performed. Nevertheless, algorithmic approaches to detect this modality of salience are scarce. Consequently, it is particularly relevant to understand and convey the discoveries performed in both areas of expertise for the model introduced in this work.

Considering that one of the limiting factors of this work is the lack of adequate data to develop and test our model, our second objective consists in finding an adequate proxy that allows us to verify the effectiveness not only of our proposal, but also those of the state-of-the-art. As it will be explained in Section 3.2, it turns out that humans are particularly sensitive to the onset of acoustic events, and not that much to their disappearance. As a consequence, our main assumption will be that the capability of saliency systems to detect sudden and unexpected onsets is directly related to the detection of saliency in human participants. Fortunately, it is common for event detection datasets to include not only their labels but also their onset and offset times. Thus, some of the available datasets for event detection will be reviewed and used when possible to design and test the performance of the system described in this work.

Additionally, in order to determine if the algorithm that we propose is a real contribution to the state-of-the-art, it will be compared with some of the available auditory saliency algorithms in the literature. We also include some well-known detection techniques that have been proposed for speech processing, since we consider that they might perform remarkably well for the onset detection task. Moreover, in the field of Music Information Retrieval it is common to evaluate onset detection algorithms, since they are quite useful for several tasks such as automatic annotation. Consequently, some authors have made contributions and crafted algorithms specifically designed to detect such changes in audio. Our intention is to use some of the most popular ones and evaluate their performance with the aforementioned datasets.

Finally, it should be noticed that the environment plays a critical role in the processing of audio signals. A presumably realistic model should be capable of working under scenarios where the environmental conditions are suboptimal, and even more, detrimental for its performance. Thus, it seems reasonable to verify how our system behaves when noise occurs, since it is a quite common disturbance in audio. We propose to contaminate our datasets with stationary and non-stationary noisy signals, and repeat the experiments in order to study the robustness against noise of the detection systems presented in this work, including the state-of-the-art systems and our approach.

1.3 Outline

In addition to the current chapter, oriented towards the introduction of the motivation and the objectives of this thesis, eight other chapters are included. The second chapter introduces the mechanisms that make humans capable of hearing the sounds that populate the acoustic environment. In fact, this chapter explains not only how the external ear works but also some efforts by previous researchers to emulate such mechanisms. In addition, we introduce certain methodologies that can be implemented with real-life signals that are useful for their processing and computational representation such as the spectrogram and the cochleogram.

The third chapter formally introduces the concept of attention, not only defining it but also describing the procedures by which psychologists have modeled it in the past in their effort to understand how the brain works. This chapter also defines salience, probably one of the most relevant definitions for this thesis, and how it has been modeled for visual and auditory signals in the state-of-the-art from a computational perspective. Some well-known algorithms are introduced for both modalities emphasizing those related to auditory saliency and including some alternative methodologies that were designed to detect the occurrence of acoustic events.

Chapter four explains how the models that are presented in this work were evaluated, as well as the benchmark data that was used during such process. It is followed by the fifth chapter, where we introduce Bayesian Surprise. Such methodology is in charge of the computation of saliency by means of Kullback-Leibler divergence, and is the core of the first model that we propose in this thesis: Bayesian Log-surprise. We devote this chapter to justify why we designed Bayesian Log-surprise, as well as the advantages that it introduces with respect to Bayesian Surprise and some of the limitations that we observed and that we think that should be appointed. We evaluated both techniques thoroughly considering the evaluation benchmark included in the fourth chapter.

Chapter six introduces our second model, an auditory saliency algorithm that uses and improves Bayesian Log-surprise for the detection of saliency in acoustic signals. We name this model Echoic Log-surprise, and we observe that it surpasses the rest of the techniques used in our benchmark, although we consider that it could be improved. We study this possibility in chapter seven, where we introduce new methodologies to compute Echoic Log-surprise, as well as an alternative thresholding algorithm that allows the system to perform online. Results suggest that the latter improvement is clearly significant for some of the datasets considered in the evaluation benchmark.

Although the experiments presented in the previous chapter are relevant, we consider that it is necessary to verify the performance of all the detection systems in adverse conditions. In chapter eight we contaminate audio files with both stationary and non-stationary noise signals using different Signal-to-Noise ratios (SNR). This should help to determine which one of the detection algorithms is more robust against the influence of noise.

Finally, in chapter nine we summarize the conclusions that were determined from the experiments of the previous chapters, in addition to some future lines of work.

We expect that this work allows both specialists and non-specialists in the area to pull out some knowledge from the field of computational saliency detection.

Chapter 2

Hearing

As defined in [Merriam-Webster.com, 2019], "*senses are specialized functions or mechanisms (such as sight, hearing, smell, taste, or touch) by which an animal receives and responds to external or internal stimuli*". Indeed, our senses allow us to perceive our surroundings, their evolution and help us, in the end, to acquire knowledge. Some of our senses are more critical than others for our survival, such is the case of vision and hearing. In particular, some authors theorized that hearing works as an *early warning system* that constantly acquires information from our surroundings and warns about potential threats (see [Mazza et al., 2007; Cervantes Constantino et al., 2012]). This chapter introduces how the Human Auditory System (HAS) works, considering some of the different elements that conform it as well as some of the tasks it performs. Generally speaking, it works as a transducer that converts environmental acoustical data into electric impulses that are compatible with our brain, and during this transformation stage it pre-processes and enhances the signal before it reaches auditory cortex. Besides the hearing process, we consider relevant to explain more in detail the concept of critical bands and lateral inhibition since they have an important role in the analysis performed in later sections of this thesis, in addition to some of the most well-known spectro-temporal representations developed in the recent decades that took significant inspiration in the HAS.

2.1 Human auditory system

The HAS is divided into three stages [Fastl and Zwicker, 2007]: outer, middle and inner ear. The outer ear functions as the interface with the surrounding world, and is in charge of directing sounds towards the ear drum. In an overly-simplified explanation, incoming sounds make the drum vibrates and this membrane transmits the acoustic wave to a chain of ossicles that finally reaches the inner ear. During this process there is a transformation of impedances [Fastl and Zwicker, 2007], since the incoming sound is prop-

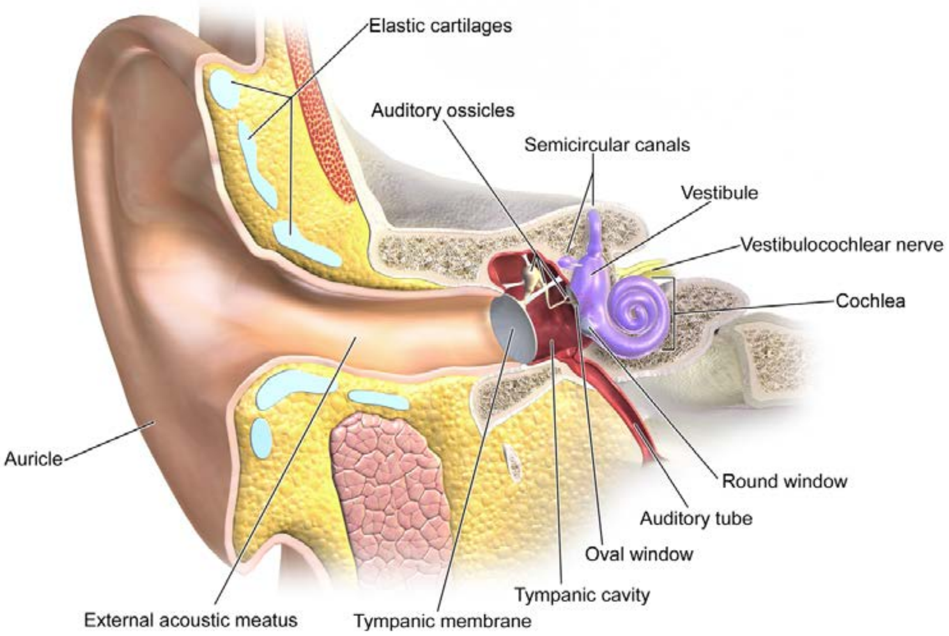


Figure 2.1: Anatomy of the ear including the external, medium and internal areas. Figure authored by [Blausen, 2014].

agated through the air whereas the inner ear sensory cells are surrounded by a liquid. The vibrations that come from the tympanic membrane and the chain of ossicles get into the inner ear through a membrane called oval window.

One of the components of the inner ear is the *cochlea* and, as its name suggests, has a spiraled shape. It is essentially formed by three conducts: the vestibular, the tympanic and the cochlear ducts. The first one is connected with the oval window and receives the vibrations from the middle ear ossicles through it. At the same time, there is a round window that moves in counter-phase with the oval window, allowing the liquid inside the ducts to flow and move. This round window is connected with the tympanic duct. Both the tympanic and vestibular ducts are separated by a third one, denoted as cochlear duct. This duct receives the vibrations of the liquid from the vestibular duct by means of another membrane, denoted as Reissner's membrane. This vibration is perceived by the organ of Corti, which is placed over the basilar membrane inside the cochlear duct. This organ is crowded by hair cells that amplify and convert the incoming vibrations into electro-chemical impulses that are directly sent to the brain.

The task performed by the basilar membrane is particularly interesting from the signal processing perspective. Essentially, when a sound propagates through the cochlea the membrane vibrates at different spatial locations

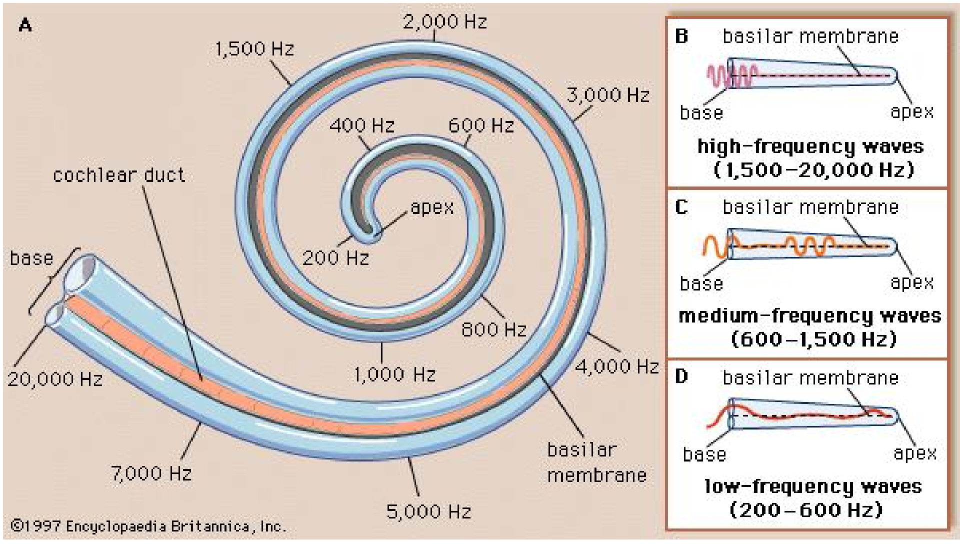


Figure 2.2: Representation of the cochlea and how it processes different frequency bands. Notice that frequency content would be gradually processed from higher to lower bands, beginning at the base of the helicoid. Figure from [Encyclopædia Britannica, 2019].

depending on the frequency content of the incoming signal, as shown in Figure 2.2. If the content consists of high frequency components, these produce maximum responses at the beginning of the cochlea, where its radius is bigger. The lower the frequency components, the closer to the apex the excitations perceived by hair cells take place, where the membrane is softer and more sensitive to vibrations. As a consequence, the behavior of the cochlea is commonly compared with a bank of bandpass filters placed along its surface. The spacing and the location of such equivalent filters is not linear, and it is usually denoted as tonotopic organization. Interestingly, this organization is propagated into the primary auditory cortex.

2.2 Critical bands

As we previously said, the cochlea works as a bank of filters that processes different frequencies of incoming audio individually. In fact, we could try to design a system to emulate such behavior if we managed to describe the physical properties of these filters. Some of the first authors that described this filterbank and modeled it as a set of critical bands are [Fletcher and Munson, 1933; Fletcher, 1940], who related them to the concept of *frequency masking*.

Frequency masking occurs when two tones with different central frequencies are presented to a test subject, but he/she is only able to perceive one. This behavior occurs because of two reasons: first of all, the dominant tone creates a mask that covers its surrounding frequencies with a certain bandwidth. Then, if a secondary tone lies within such bandwidth and its magnitude is not sufficiently big then it gets masked by the dominant tone. In this scenario, the participant becomes unaware about the secondary tone information. The bandwidth of the dominant tone mask is related to the central frequency value, considering that the bigger the central frequency the wider will be the bandwidth of the masker tone. In addition, filters would be spaced in frequency following a logarithmic pattern from low to high frequencies which implies that there is a larger number of low frequency filters in comparison with high frequency ones (see Figure 2.3, where some filterbanks with these properties are depicted). We can therefore conclude that the HAS employs more physiological resources to process low frequency contents that in turns explains why humans are more sensitive to these ranges of frequencies.

Inspired by the works of Fletcher et al. some other authors proposed different auditory filterbanks, such as the ones described below:

The Equivalent Rectangular Bandwidth (ERB) filters were proposed by [Glasberg and Moore, 1990] and are spaced according to:

$$ERB_N(f_c) = 21.4 \cdot \log_{10}(0.00437 \cdot f_c + 1), \tag{2.1}$$

where f_c is the value of the central frequency to be transformed into the so-called ERB-scale and is measured in Hz. Regarding the bandwidth of a particular filter originally positioned at the central frequency f_c, it can be linearly approximated with the next equation:

$$BW_N(f_c) = 24.7 \cdot (0.00437 \cdot f_c + 1). \tag{2.2}$$

As a consequence, if we define a set of filters with different central frequencies f_c we can get a filterbank whose filters are logarithmically spaced according to their ERB_N, and their bandwidth grows with their f_c. Figure 2.3 on the top plot shows an example computed using 20 filters whose central frequencies belong to $f_c \in [100, 11050]$ Hz, logarithmically spaced and with an increasing bandwidth, according to the previous equations.

The Gammatone filters [Slaney, 1993] are described in the temporal domain by the following impulse response:

$$\Gamma_{tone}(t) = \frac{a \cdot t^{n-1} \cdot \cos(2\pi f_c t + \phi)}{\mathrm{e}^{2\pi b t}}, \tag{2.3}$$

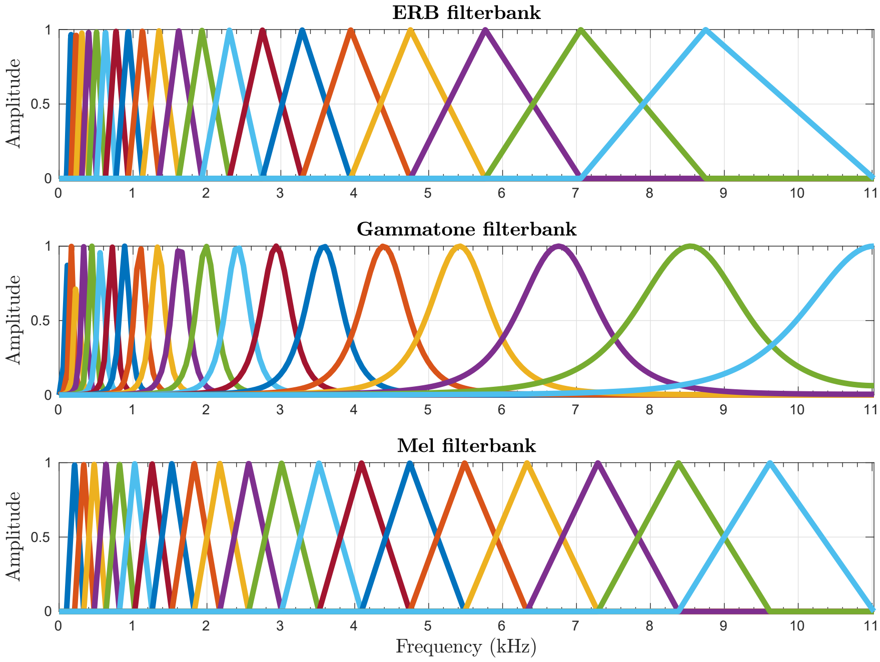


Figure 2.3: Examples of filterbanks that emulate the tonotopy present both in the cochlea and in the primary auditory cortex.

where the bandwidth of the filter is defined by b, the central frequency is f_c and n defines the order of the filter. Other parameters are the amplitude a, the carrier phase ϕ and the time t. In contrast with the ERB, we observe that these filters have more parameters to tune.

Alternatively, its transfer function would be:

$$\Gamma_{tone}(f) = a\Big[1 + j\frac{f - f_c}{b}\Big]^{-n} + a\Big[1 + j\frac{f + f_c}{b}\Big]^{-n}. \tag{2.4}$$

In the example depicted in Figure 2.3, the filterbank was designed using the Voicebox Toolbox [Brookes, 1997], with 20 filters of 4-th order spaced in frequency according to the ERB scale for frequencies in the range $f_c \in [100, 11050]$ Hz, although filters are not aligned with the previous ERB. Two differences contrast with the ERB example explained previously. First, the shape of the filters is different, since they were obtained as the combination of a tone with a gamma-like signal. Their Fourier transforms present the shape depicted in Figure 2.3, second plot. Secondly, the bandwidth of the filters also changes.

The Mel filterbank is obtained from the Mel scale [Stevens et al., 1937; Davis and Mermelstein, 1980; Xu et al., 2005] which similarly to the

ERB-scale spaces frequencies logarithmically according to:

$$Mel(f_c) = 2595 \log_{10}(1 + \frac{f_c}{700}) \tag{2.5}$$

Usually, the Mel filterbank is designed considering triangular filters as the ones depicted in Figure 2.3, where filters overlap at the point where their amplitude is halved. This bank can be efficiently obtained, and is able to generate a transformation matrix that can be directly used to project the frequency content of a Short-Time Fourier Transform vector into the Mel frequency space. In the example shown we considered the same setup parameters explained previously for Gammatone filters.

One of the properties of the filterbanks obtained according to any of the three methodologies explained above is not only that they resemble how human audition works, but also that they can be used to obtain biologically plausible features for signal processing algorithms. For instance, as we mentioned for the Mel filterbank, these techniques allow to project spectral information linearly spaced in frequency into an alternate frequency scale, where the effect of frequency masking and logarithmic spacing is considered. As it will be introduced in later Sections, we rely on these transformations to obtain more representative features from incoming audio signals.

2.3 Lateral inhibition

The neurons that conform our brain are subject to a constant flow of signals and information. In the particular case of sensory information, as it occurs with our visual or auditory systems, the processing is carried out by neurons located most of the times around the same area, and the excitation might propagate uncontrollably from one to another, even though only some of them are excited. In order to avoid such behavior these related cells are usually interconnected by inhibitory neurons in such a way that when a neuron gets excited by incoming information, an inhibition signal is sent to the neighboring neurons so their activity is attenuated [Goldstein, 2009]. Such behavior should indeed increase the *performance* of the activated neurons, which for the case of vision would imply, for example, an improved capacity to detect edges or colors.

For the perception of audio, [Shamma, 2001] suggested that inhibition would help to detect peaks and sudden changes in audio. For instance, [Okamoto et al., 2007] suggested that lateral inhibition could be asymmetric in the case of the HAS. In Section 2.1 we explained that low frequency sounds travel through the cochlea until its apex, where these frequencies are processed. However, the traveling effect might activate some neighboring frequencies in the cochlea, and as a consequence, the signal that is sent to the auditory cortex is not formed by discrete peaks representing the central frequencies, but instead is widespread in the neighboring spectra. According to [Okamoto et al., 2007], the lateral inhibition mechanism for the auditory cortex would be in charge of attenuating the spectral data around the dominating central frequencies considering notch areas around them. Frequencies appearing inside these regions would be strongly inhibited by auditory neurons, whereas those frequencies perceived outside of the notch would be activated as usual. As a consequence, this would ease the detection of central frequencies, suggesting that lateral inhibition would be in charge of helping us to detect spectral edges in incoming auditory data. In addition, the authors suggested that the lower would inhibit the higher frequencies more strongly than the opposite, which is in line with the idea that HAS prioritizes lower frequency contents.

Considering altogether that the auditory cortex is tonotopically distributed and the concept of neural inhibition, some of the saliency techniques mentioned in this work emulate these two mechanisms using a multi-scale structure and wavelet filters (see Sections 3.6.1 and 3.6.2 for auditory algorithms, and Section 3.5.1 for a visual algorithm).

2.4 Spectro-temporal representations

As it follows from the previous sections about the HAS, both the temporal and frequency domains are instrumental in the preprocessing of the acoustic signal to allow the subsequent *decoding* that the brain carries out. This poses important challenges for the computational representation of such signals. First, the spectral domain plays a crucial role, since it helps to organize their content into several frequency components in a similar fashion to what the HAS and specifically the cochlea does. Second, the temporal evolution of these frequency bins is essential for the detection of prominent acoustic events (but not only). In the following subsections we introduce how to represent an audio signal into the spectral domain, and some efficient representations that help to analyze the spectral evolution of such signals along time.

2.4.1 Short-Time Fourier Transform

Acoustic signals such as natural sounds or speech are non-stationary, which means that they usually comprise several components whose statistical properties change along the temporal axis. By means of the Short-Time Fourier Transform (STFT) these signals can be described as a linear combination of sinusoids with different amplitudes and phases [Oppenheim et al., 1999]. As an example, the shape adopted by the vocal tract during a conversation is what makes the components change over time and is the mechanism by which humans convey speech information. Fortunately, these changes can be considered slow varying for periods within a range of 10 ms to 40 ms. This is what is usually regarded as quasi-stationarity. This property is usually extrapolated to other acoustic signals, such as the ones acquired directly from nature.

As a consequence, when attempting to extract the frequency information from such signals it is necessary to consider temporal windows or chunks of data (possibly overlapping) with a proper length, so potential discontinuities and detrimental effects are minimized. This can be done thanks to the STFT, which extracts spectral content from a temporal window that has a limited duration. The STFT can be obtained as follows for a signal $x(m)$,

$$X_c(k,n) = \sum_{m=-\infty}^{\infty} x(m) \cdot w(m-n) \cdot \exp(-j\frac{2\pi k}{n_{fft}}m), \tag{2.6}$$

which shows that an incoming signal $x(m)$ can be transformed into a bi-dimensional complex signal $X_c(k,n)$, considering that L_{win} temporal samples are windowed by w and their frequency data is conveyed both in the temporal frame n and the frequency bin k. Moreover, unless stated otherwise, we work with the modulus of the STFT and use the following notation:

$$X(k,n) = |X_c(k,n)|. \tag{2.7}$$

Notice that in total n_{fft} frequency bins are computed, and the temporal axes of $x(m)$ and $X(k,n)$ represent different information. The temporal signal $x(m)$ represents the amplitude of the signal for a single temporal instant m, whereas $X(k,n)$ represents the magnitude for the n-th temporal frame that contains and summarizes several consecutive L_{win} temporal instants under the temporal index n, for a frequency bin k.

The window, represented as $w(m-n)$, defines how many points of the original signal $x(m)$ are used to compute the STFT, and as we explained its duration is defined by L_{win}. If we had a theoretical signal with an infinite number of temporal instants, we could use the window w to divide it into pieces of length L_{win} and compute the Fourier Transform of each one of them. As it is described by [Oppenheim et al., 1999], the longer the window the better the frequency resolution, and the STFT will be capable of showing really fine-grained frequency values. At the same time, the bigger the window the longer will be the temporal sequence, to the extent that it might become so big that the sequence is not stationary anymore within the window, as we required. Hence, there is a trade-off between temporal and frequency resolution. It is desirable that the statistics of consecutive windows barely change along time, and at the same time we want a window sufficiently long to have a good frequency resolution. Some common window length values that meet these requirements for the usual acoustic signals are 10 ms to 40 ms.

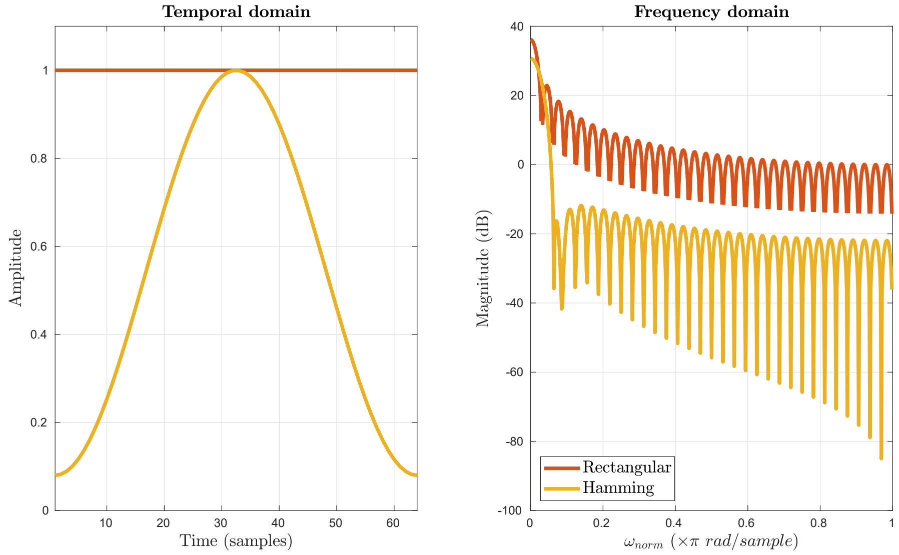


Figure 2.4: Representation of Rectangular and Hamming windows in both temporal and frequency domains.

In addition, the process of windowing can be understood as chunking the signal into pieces by applying a rectangular window. From the point of view of frequency analysis, the ideal window would be a rectangle with infinite temporal length, equivalent to a constant signal, i.e. no windowing, since its spectral response would be a Dirac delta function that would not introduce any artifacts in the spectra of the signal. However, the rectangular window has a limited duration and its spectral response is depicted in Figure 2.4, where it can be observed that it introduces artifacts due to its frequency contents around zero that affect to the outcoming STFT vector. Fortunately, there exist several window shapes whose spectral properties are more beneficial for the spectral content of the input signal, and they are usually characterized by two parameters according to [Oppenheim et al., 1999]: the width of their main lobe, which relates directly with the frequency resolution, and the ratio between main and side lobes, which relates inversely with the leakage effect. In this work we make use of the Hamming window, depicted in Figure 2.4, since its properties are appropriate for our requirements.

2.4.2 Spectrogram

A spectrogram is a spectro-temporal representation that can be computed using the previously explained STFT. It essentially computes the STFT, as explained in Section 2.4.1, repeatedly considering a sliding temporal window with a certain overlapping factor in order to maintain the continuity of the

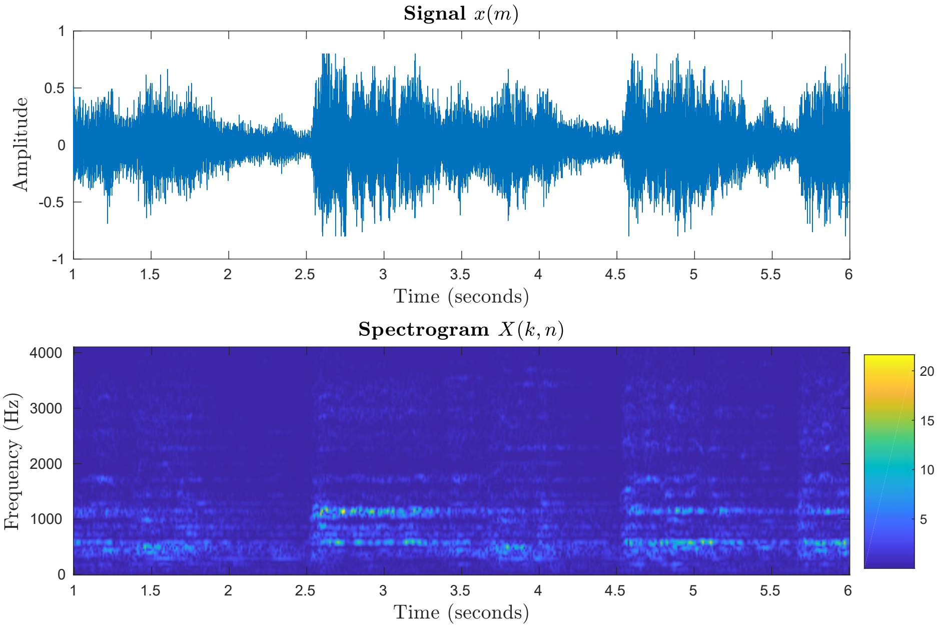


Figure 2.5: Example signal $x(m)$ and its spectrogram $X(k,n)$ computed using a window length of 20 ms and 50% overlapping factor, and sampling frequency F_s= 8 kHz.

signal. Its properties are determined by three parameters: the number of frequency points (n_{fft}), window size (L_{win}) and overlapping factor. As its name indicates, the window size (or equivalently, frame length) determines the length of the temporal window that is going to be used to compute the STFT. As we explained in Section 2.4.1 it needs to be relatively small in order to keep the quasi-stationary properties of the acoustic segment. For example, for a sampling frequency of F_s = 44.1 kHz the window length would be L_{win} = 882 samples long, which would be equivalent to 20 ms of audio from the original signal. On the other hand, the overlapping factor helps to avoid discontinuities in the temporal axis of the spectrogram by keeping a portion of the signal from the previous frame. As an example, using L_{win} = 882 with an overlapping factor of 50% would mean that last half of the information used to compute the n-th frame, 441 samples, would be kept for the $(n+1)$-th frame. Finally, n_{fft} controls the frequency resolution of the spectrogram and indicates how many frequency bins are to be computed. In order to improve the resolution of the output spectra the system can be configured so that $n_{fft} > L_{win}$. In this case, during the computation of the STFT the acoustic segment is padded with zeros until its length becomes $L_{win} = n_{fft}$, which according to the zero padding theorem [Smith, 2007] leads to the optimal interpolation in frequency.

An example of the outcome of this process is shown in Figure 2.5, where the top subfigure displays the temporal representation of a signal whereas its spectrogram is in the bottom. It should be noticed that even though both signals have been depicted with a temporal axis represented in seconds, in the case of the signal $x(m)$ time is measured in samples, whereas for the spectrogram time is measured in frames, which were obtained after computing the STFT using temporal windows as explained before.

This spectro-temporal representation has been widely used in the state-of-the-art of many areas of research related to audio and speech processing, including acoustic saliency algorithms. Because of its two dimensions, time and frequency, it produces an image-like representation of an acoustic signal that eases the process of adapting a visual saliency algorithm to work with audio. Such is the case of [Kayser et al., 2005; Tsuchida and Cottrell, 2012; Schauerte and Stiefelhagen, 2013], among others.

2.4.3 Cochleogram

As we explained in Section 2.2, human beings are not equally sensitive to all the frequencies in the auditory spectrum, and some authors proposed auditory-inspired filterbanks that model how our cochlea behaves by introducing a non-linear, logarithmic-like transformation of the frequency axis. Using these filters can be fruitful if they are combined with a spectro-temporal representation such as the spectrogram, because the knowledge about tonotopy that they convey is directly applicable to the spectra obtained along time once again imitating the cochlea behavior. This produces a new version of the spectrogram that includes well-known perceptual knowledge about how HAS works, and it is usually denoted as *cochleogram*.

It should be stated that most of the configuration parameters of the spectrogram and cochleogram are shared, with the exception of the number of perceptual filters used to obtain the latter. In this work we use the Mel bank of filters proposed by [Davis and Mermelstein, 1980; Xu et al., 2005], and we refer to the number of filters as n_{Mel}. One of the advantages of this filterbank is that it can be modeled as a transformation matrix that can be directly multiplied by the spectrogram. The resultant spectro-temporal representation would share the number of temporal frames, but would have n_{Mel} frequency components instead of n_{fft}. In addition, since the perceptual filters explained in Section 2.2 group frequency content into a prescribed number of bands, we need to require that $n_{fft} > n_{Mel}$, which implies that during the transformation process there is also a dimensionality reduction.

An example, similar to the one showed in Figure 2.5 and using the same dummy audio is depicted in Figure 2.6 for $n_{Mel} = 150$ bands. It shows how the low frequency contents have been expanded whereas high frequencies appear compressed. Such behavior is a direct consequence of the logarithmic spacing of the Mel filterbank, where most of the filters concentrate in the

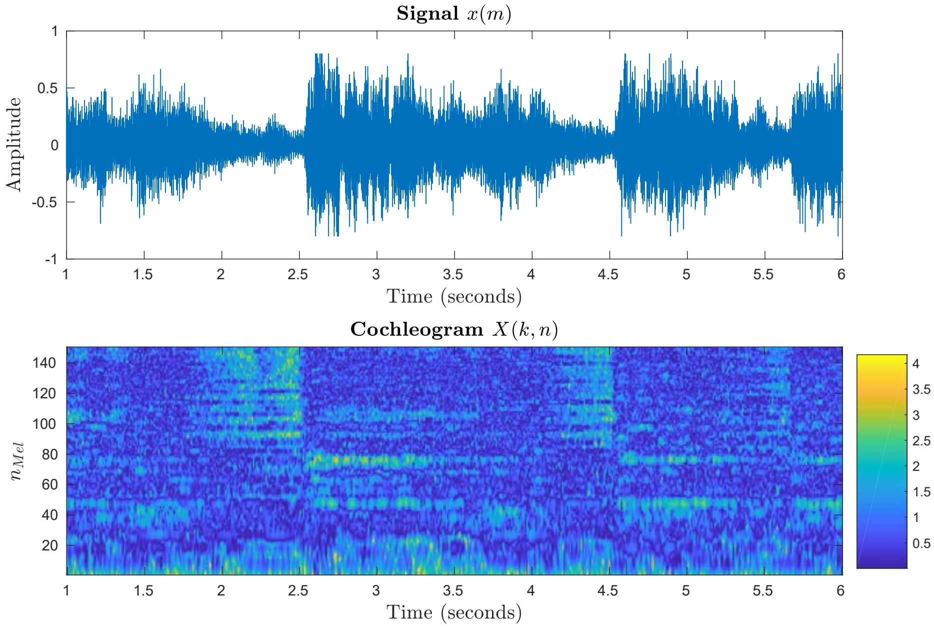


Figure 2.6: Example signal $x(m)$ and its cochleogram $X(k,n)$ computed using a window length of 20 ms and 50% overlapping factor, and sampling frequency F_s= 8 kHz.

lower frequency ranges giving as a result a finer resolution, whereas higher frequency ranges are represented with less filters.

It is common to use the cochleogram as a part of a more advanced feature extraction process, as it happens with the computation of the widely known Mel Frequency Cepstral Coefficients (MFCC) [Davis and Mermelstein, 1980; Xu et al., 2005], where the log-Mel energy bands are computed applying the logarithmic operator to the Mel cochleogram, and then projected using the discrete cosine transform. Another proposal based on cochleograms, although using gammatone filters instead of the Mel filterbank, are the Power-Normalized Cepstral Coefficients (PNCC) [Kim and Stern, 2010], one of the most recent features for speech processing that are not based on deep learning methodologies. See [Abka and Pardede, 2015] for an analysis of the robustness of some commonly used features for speech recognition.

Nowadays, however, with the trend of deep learning the log-Mel energy bands have become a popular feature for many recent proposals not only for speech processing [Seltzer et al., 2013; Drugman et al., 2015], but also in other fields of research such as audio classification and detection, where they have been used for a wide variety of applications such as scene classification and event detection [Mesaros et al., 2018; Serizel et al., 2018].

As a consequence, we observe that there are plenty of reasons in the state-of-the-art that justify the choice of the cochleogram as the spectro-temporal representation for this work.

Chapter 3

Attentional mechanisms

According to [Anderson, 2015], attention is related to perception in the sense that it explains how the human brain is able to determine what information is relevant in a specific context. Our brain is unable to focus on and understand all the incoming data acquired through our senses, and attention is the mechanism that helps to determine what cues should be prioritized. It is important to remember that humans face a vast flow of information from different sources and nature thanks to our senses.

Attention is commonly divided into two different categories [Pinto et al., 2013; Katsuki and Constantinidis, 2014; Menon, 2015]: bottom-up and top-down. Bottom-up attention has the property of being stimulus driven, which implies that specific cues are perceived in a such way that they automatically force a listener or observer to focus on them. As a consequence, one of the key points of bottom-up approaches consists of determining the physical properties and features of cues that make them so prominent or salient. Bottom-up attention is usually associated with the primal goal of survival, and therefore is strongly associated with spontaneous responses directed towards this primal task, as for example, when there is a fire alarm and everybody notices it because this specific sound contrasts with the environmental background.

The alternate class is top-down attention, which is task-driven since the subject has a specific task to solve and uses all the resources to do it. In addition, it can be observed that for top-down attention the subject has some prior knowledge about what to do or what is about to happen. An example of top-down attention could be a person playing pinball, whose main task would consist of scoring as many points as possible while avoiding that the ball went down, or a blind person that is crossing the street and uses sounds produced by traffic lights to get oriented and arrive safely to the other side of the road.

According to [Pinto et al., 2013], top-down and bottom-up attention cause the same reaction whether they are produced by different mechanisms or not: both help the brain to focus its neural resources on a certain object, task or concept. In their work they mention that the bottom-up mechanism might be a more primitive component of attention, since its existence has been proved for insects such as the fruit fly. During their research they worked with a group of participants that performed a visual search task and also a capture task. Their results suggest that top-down and bottom-up attention are independent of each other. This would imply that they are controlled by independent neural circuits. As they propose, a possible way to interpret this is that bottom-up attention is triggered really quickly as a reflex reaction, whereas in later stages of neural analysis the top-down reaction is produced. Measurements about the reaction times suggested that bottom-up attention required from 100 to 150 ms, whereas top-down attention usually needed around 100 ms more to be deployed.

This distinction between the two modalities of attention also exists in computational models. Bottom-up attention algorithms are usually used to detect and produce saliency maps from incoming multimedia signals, where saliency is identified as the feature of perceptual objects that makes them prominent in a context. We refer to these models as saliency detection algorithms. On the contrary, top-down attention algorithms try to detect prominence from the incoming perceptual data and use it to solve specific tasks.

In this chapter we introduced some attention algorithms that mostly belong to the bottom-up modality, since we are interested in modeling audio saliency. Nevertheless, it is important to understand the computational nature of such approaches and their dependence on the available datasets. We make two distinctions: supervised and unsupervised models. Supervised models are trained using data that includes both input signals and output labels, which are provided by human experts. On the contrary, unsupervised algorithms only use input signals and usually lack any labels. As an example, for visual saliency detection it is common to have massive datasets containing images or videos which are labeled using eye-fixations from human participants. As a consequence, there exist both supervised and unsupervised algorithms for visual saliency detection, and as it was introduced in the work of [Bylinskii et al., 2019a] there are several proposal nowadays. On the contrary, for auditory saliency there is no consensus yet about how to obtain reliable labels and to the extent of our knowledge there only exist unsupervised auditory saliency detectors.

More information about these concepts will be thoroughly explained in this chapter.

3.1 Attention models

As it was suggested in the previous section, the attentional capabilities of human brain are limited. We are able to perform various unrelated tasks at the same time with a certain degree of success, such as listening through our phones while we walk. However, other chores such as driving and having a phone conversation might not be adequate, since our attentional resources are more severely compromised, as [Ishida and Matsuura, 2001] pointed out in their work. Thus, these limitations are related to the difficulty of each task and how they interfere with each other. As an example, some experiments show that when a participant is asked to attend to a sequence in one ear and a different sequence is played in the opposite ear, some changes in the latter sequence go unnoticed (see [Cherry, 1953] in Section 3.2 of this work). This Section introduces the work of some authors that proposed global models of attention from the perspective of psychology, whose experimental hypothesis were tested using acoustic and visual data acquired from human participants.

Considering the fact that human brain has limited resources to understand its perceptual environment, [Broadbent, 1954] proposed a filter theory following the scheme depicted in Figure 3.1. This theory can be introduced considering two concurrent signals that are perceived by a person. Based on their physical properties, for example, pitch or intensity, the brain decides which one is the most relevant and puts all the attentional resources into processing it. This would be the attended signal, which gets into short-term memory (See Chapter 6) and later into working memory, being the latter a controversial term according to [Cowan, 2008] since depending on its definition it is considered to be synonymous of the former. We will follow the criterion of the aforementioned author, who states that information gets into short-term memory totally unprocessed, and then it quickly gets into working memory where attentional processing is performed. The unattended cue would be filtered out, and consequently removed from the processing pipeline. Since this theory states that this action is performed using certain features of the signal, the semantic content is not processed until the attended signal is selected and the other is discarded. Nevertheless this approach might have some issues, as [Moray, 1959] showed in his work where he demonstrated that test subjects were able to detect their names in the unattended channel, which inherently meant that unattended information was not removed. This implies that the filter model proposed by Broadbent could be overpassed by relevant information, as it turns out to happen with the names of the test subjects. Other authors also verified that the brain did not filter unattended information and participants were capable of processing it (See [Moray, 1959; Treisman, 1960; Underwood, 1974]).

An updated model was proposed in [Treisman, 1964a] as depicted in Figure 3.2, whose main difference with respect to Broadbent's approach was that unattended signals were not filtered but instead they were in a first in-

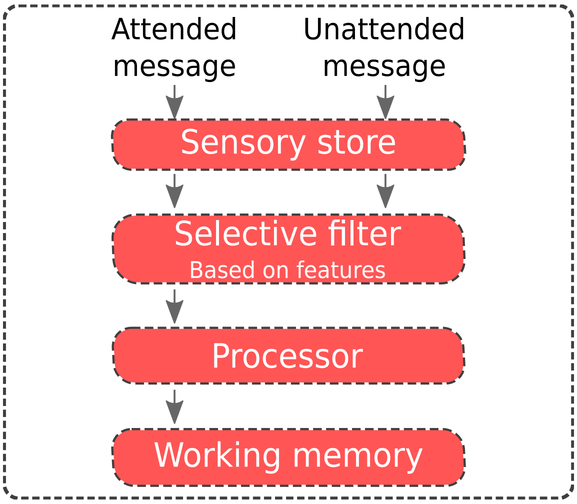


Figure 3.1: Broadbent's filter theory workflow.

stance attenuated considering their physical characteristics. Then, attended and attenuated signals got into an analyzer in charge of determining whether the incoming signals exceeded a threshold value depending on their meaning. This allowed some flexibility to extract the content from the unattended channel, although with bigger difficulty than that of the attended one. This approach gave an explanation about why people were able to recognize their own names directly from unattended signals, and suggested that the selection of the channel does not rely exclusively on the physical properties of the signals but also depends on their semantics. This model is usually referred to as Treisman's attenuation model.

Another model was proposed by [Deutsch and Deutsch, 1963] as an alternative to Broadbent's, since other works verified that signals are not filtered out but instead they are processed and then discarded so only the relevant information remains. As the Figure 3.3 shows, this proposal designed by [Deutsch and Deutsch, 1963] starts with the assumption that our brain is capable of giving a certain score of relevance to every incoming signal, which might be subjected to the task in progress and it is automatically computed according to the semantics of the incoming signals. Then, only the most relevant signal remains and the rest are ignored. In contrast with the models of Broadbent and Treisman, this scheme relies exclusively on the meaning of the incoming sensory data. On the contrary, one of the most criticized aspects of this model is the necessity to fully process the incoming information before selection is made [Treisman, 1964b].

The aforementioned proposals are usually referred to as bottleneck models, since they theorize that attention is somehow composed of a set of limiters that control what information gets into working memory, and somehow

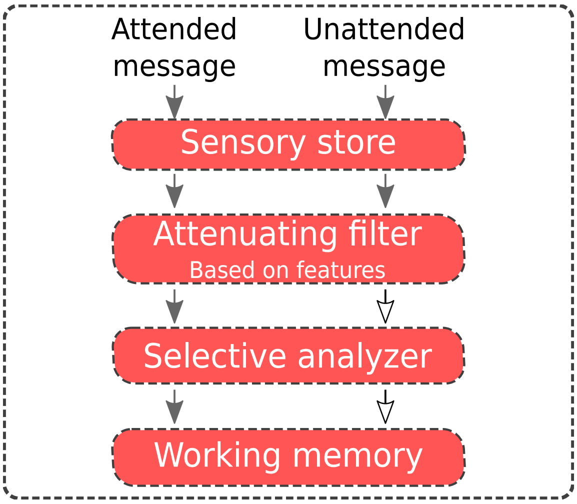


Figure 3.2: Treisman's attenuation model.

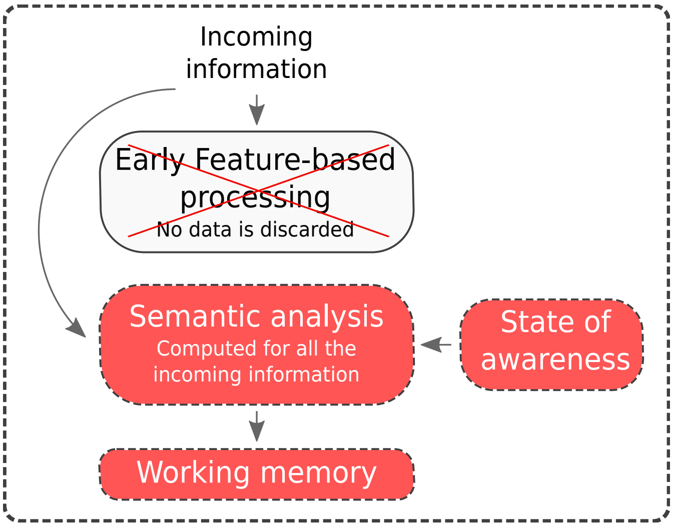


Figure 3.3: [Deutsch and Deutsch, 1963] attention model. Notice that according to this scheme incoming information does not get processed in the early stages of neural work.

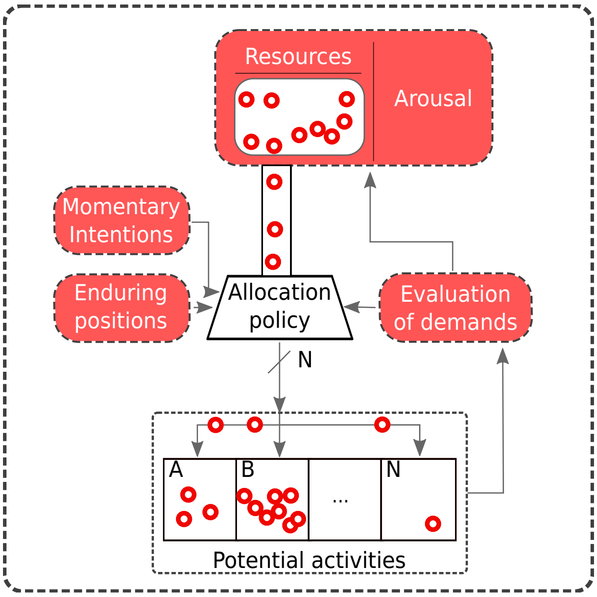


Figure 3.4: Kahneman attention model inspired by Figure 1.2 in [Kahneman, 1973].

reduce or even eliminate non-relevant data depending on the task and the context. [Kahneman, 1973] proposed an alternative model that, instead of a bottleneck, uses a processing unit that is in charge of evaluating the demand of mental resources, in addition to their allocation depending on their availability and the relevance of the tasks to be performed. As it is shown in Figure 3.4, the system is influenced by the degree of arousal of the brain, the effect of salient incoming signals that were involuntarily perceived, a re-evaluation of the demands when there is not enough capacity available for various concurrent tasks, and intentional behaviors such as performing specific tasks. Notice that the concepts of top-down and bottom-up attention are absorbed into this model, but it does not define the magnitude nor the capacity of the system, in the same way that it does not take into consideration that humans are able to improve their skills, and how this affects the global attentional capacity. Other authors, such as [Bruya and Tang, 2018], studied thoroughly the work of [Kahneman, 1973] and concluded that *effort* is a concept rather ambiguous and poorly defined in the original text. In fact, they analyzed other works which suggested that attention is not exclusively equivalent to effort in a literal way, since there are contexts where attention is used effortlessly and automatically by our brain.

As a consequence, we observe that there have been efforts to explain how human brain manages incoming information, considering that it is an organ with limited cognitive resources. Nevertheless, nowadays it seems that there is still no consensus about this topic.

3.2 Attention and change detection

A remarkable effect related with the perception of external stimuli is the detection of changes in the environment. It turns out that we are not as capable to detect such changes as we usually think, neither for appearing nor disappearing objects.

When the sense under analysis is vision, the inability to detect such changes in a scene is referred to as change blindness. An interesting work where the authors tried to understand this effect was conducted by [Rensink et al., 1997], who suggested that the most critical factor related to change blindness was attention, and stated that changes could only be detected when attentional resources were focused on the particular change that was about to occur. In any other scenario, changes would go unnoticed. In order to test this hypothesis they proposed a paradigm where participants had to watch a sequence of images and modified versions of themselves, which were separated by a blank image causing a flickering effect. The effect of this blank image was so intense that the time that participants took to detect changes in the images increased severely when it was displayed, and decreased quickly when it was not employed. They also tried to determine if this decrease in the performance of the participants happened because they did not have enough time to understand and memorize the images and their modified counterparts, but results showed that the time they had to store visual information was sufficient. In fact, they also tried to verify how external cues influenced their performance, and they observed that the flicker effect did not hinder the visibility of images. They concluded that changes could only be detected when attention was focused on the areas or objects that were going to suffer the change, and in this case their properties would probably get stored in visual short-time memory.

In addition to vision, some other senses show this blindness to changes, as it happens with olfaction [Sela and Sobel, 2010; Forster and Spence, 2018] and haptic perception [Gallace et al., 2007; Auvray et al., 2008]. Moreover, many researchers have studied this inability to detect changes in auditory scenes, a concept commonly known as change deafness.

An early experiment related to this concept was performed by [Cherry, 1953], who asked participants to listen to two different audio signals, one per ear. They were supposed to pay attention and repeat the speech presented in their right ear, which was a message in English, whereas they were told to ignore the speech perceived by their left one. The speech perceived in the left ear was originally presented in English, and the speaker switched to German at a certain unexpected moment. As it was expected, participants were capable of repeating the speech that they perceived through their right ear. However, they were unaware of the change in the language that was introduced in the left-ear message, which would imply that their attentional resources were totally oriented to solve the main task they were originally assigned: listen and repeat the audio perceived from the right-ear. Interestingly, their results also showed that one of the few things that participants were able to remember was if the speaker was male or female.

[Vitevitch, 2003] tried to verify the existence of change deafness. Similarly to the work of [Rensink et al., 1997] for change blindness in vision, the hypothesis to verify was that in order to detect a change in an acoustic object, such as its dialect or a change in the speaker, participants should be focusing on that particular aspect that is changing. To assess it the author proposed a experimental framework where two acoustic scenes were listened by the test subjects. In the first scene there was a speaker pronouncing words that participants should repeat to a microphone as quickly as possible. Afterwards, the scene two occurred where the same subject or a different one would read another sequence of words, which should be repeated by the participants as in the previous scene. Between the two scenes there might exist a silence gap with a duration of one minute for the first experiment, and no gap at all for the second one. As it has been explained before, the main task to solve consisted of repeating each word that was perceived. At the end of the trial, participants were asked if they noticed a change in the speaker at the beginning of the second scene. Results showed that some of the test subjects did not notice that a new speaker was pronouncing the words, and they performed the task with a performance similar to the case where the speaker remained the same. However, the participants that noticed the change also suffered from a decrease in their response time and also mistook some of the words. In addition, after removing the silence gap of one minute between scenes the results remained significantly similar, which suggested that the inability to detect the speaker change by some of the subjects was not due to a memory effect related to the duration of the gap. Instead, the author suggested that as it was expected the change deafness occurred because participants were not focusing their attention in the particular mission of detecting a new speaker, but instead used their attentional resources to solve the word repetition task.

More recent works have tried to explain this "deafness" by measuring the sensitivity of people to the appearance or disappearance of some controlled auditory stimuli. [Eramudugolla et al., 2005] proposed three experiments where the change detection capabilities of test subjects were tested considering two conditions: directed, where participants were asked to confirm if a particular acoustic object had disappeared, and non-directed attention, where they were asked to confirm which sound source had disappeared. In both conditions subjects were listening to an acoustic environment artificially formed of 4, 6 or 8 naturalistic audio signals distributed in different spatial locations. In their first experiment, participants were asked to detect the disappearance of one of the cues. Results showed that during the directed condition they had no trouble detecting if a source had disappeared, to the extent that their responses remained constant even for more crowded auditory scenes. Nevertheless, these idealistic results were not repeated for the non-directed attention condition, where their performance decreased severely when the scene became more complex acoustically speaking.

In order to assess if their results were biased by spatial information, they proposed another experiment where they presented all the sounds from the same spatial location. Their results depicted that for the directed condition the performance remained the same than in the first experiment, whereas for the non-directed one there was a clear decrease in performance. These could suggest that for the directed condition participants were more aware about the physical features that characterized each signal rather than their spatial location. During their third experiment participants listened to an acoustic scene where two audio signals exchanged their locations while the rest remained, and they were asked to report if there was any kind of location change for the non-directed condition and a specific object change for the directed one. Variations in the complexity of the scene showed a similar pattern to the results obtained previously for both directed an non-directed conditions. However, their magnitudes were clearly below previous experiments that focused on the detection of a disappearing sound. During their conclusions they suggested that there was evidence about a relationship between the auditory perception and attention, since the only way their participants were able to detect changes, whether they were disappearances or variations in their location, occurred when they were specifically asked to attend to a certain sound. In any other scenario, their performance was severely affected.

Another example is the work developed by [Pavani and Turatto, 2008], who criticized two aspects from the work of [Vitevitch, 2003]: on one hand, participants were not told to expect auditory changes. On the other hand, they considered a simplistic auditory environment, since the main task involved the repetition of word sequences. [Pavani and Turatto, 2008] tried to determine if change deafness occurred because of the acoustic transients that appeared between two scenes separated by a gap, and also if it was related to a limitation on auditory short-time memory. They proposed a set of experiments where participants listened to two consecutive scenes formed of animal sounds, some of which appeared or disappeared in the second scene. Two different conditions were presented:

- In the first condition, the first scene was formed of three acoustic objects, and a fourth acoustic object was added for the second scene.

- The second condition differed, and its first scene was formed of four acoustic objects, one of which was removed during the second scene.

At the same time, the two scenes of the experiments were separated by:

- Silent gap of a certain duration between the two scenes. This condition belonged to Experiment 1.
- Noisy gap of a certain duration between the two scenes. This condition belonged to Experiment 1.
- No gap, which means that both scenes were not separated. This condition belonged to Experiment 2.

Finally, for their third and fourth experiments they asked participants to detect and identify the appearing or disappearing objects. They observed that change deafness occurred no matter if there was a silent gap, a noisy gap or no gap at all between the scenes, which suggests that this deafness is not a consequence of acoustic transients. In fact, the results suggested that participants were more sensitive to detect changes for the trials where the first scene was formed by three elements and an additional one was included during the second scene, which implies that participants had to store a small amount of information using auditory short-time memory, and as a consequence they performed better. They did not discard that instead there might be a limitation in the number of objects that attention could handle.

However, [Cervantes Constantino et al., 2012] suggested that using natural sequences for the detection of change deafness introduced some negative aspects that should be taken into consideration. First of all, a signal coming from a natural source tends to be more complex spectrally speaking, since their content usually spreads around certain frequency bands and there is an unavoidable level of background noise. In addition, the only way to control and localize their frequency content is by using filters that attenuate irrelevant data. As a consequence, using several natural sounds to design a complex acoustic scene increases the chance of frequency overlapping of the individual signals, which might hinder the detection capabilities of the participants. Secondly, participants could unavoidably distinguish the source of the sound, let's say an animal, which might help them to remember the whole content of the scene. For instance, using these sounds might turn the change detection task into a problem related to working memory, and how capable of remembering labels the participants are. Instead, [Cervantes Constantino et al., 2012] developed acoustic scenes formed of a set of individual tones that occurred simultaneously. They measured the capability to detect changes when one of the tones that composed the scene appeared or disappeared. After testing the validity of their stimuli, they performed several tasks in addition to simply detecting if a sound appeared or disappeared, including a loudness analysis, the usage of distractors and modifying the latency of changes to measure adaptation to the stimuli. Participants showed a high degree of sensitiveness to the appearances, no matter whether the acoustic scene was simple or complex. Additionally, they were also capable of detecting to certain extent which tone had appeared recently on the scene. Nevertheless, disappearances were perceived differently by participants. The more complex the scene, the worse they performed, in addition to the fact that they had slower response times in comparison with the detection of appearances. They also showed difficulties detecting which tone had actually disappeared from the scene. Thanks to these experiments, the authors deduced that there is an asymmetry for the two tasks under analysis. For the detection of appearances they suggest that there is some degree of neural adaptation, and the brain is able to reduce its response times for the signals that it acquires from the environment, and the addition of a new tone induces a novel peak of activity that is rather easily detected. At the same time, the brain demonstrated to be sensitive to transients to onsets and offsets, although its responses were clearly biased towards the detection of the former. On the contrary, change deafness was observed for the disappearance of tones, just as it was stated previously by [Pavani and Turatto, 2008]. Finally, they stated that the observed effects during their analysis suggest the existence of an automatic mechanism that is apparently in charge of the detection of novelties in the environment, which could be potentially related to the task of survival.

[Barascud et al., 2014] tried to measure these change deafness using a sequence of repeating acoustic tones where an anomaly appeared, i.e. ABCABCABCB. They measured the reactions using both magnetoencephalography (MEG) and behavioral responses, which suggested that for sequences of short tones people were behaviorally unaware about the appearance of anomalies. However, MEG showed signs of activity that could be related to bottom-up processing, which invites to think that there might exist a "bottleneck" in attentional resources that prevented the detection of these anomalies.

From the previous works we conclude that there seems to be a relationship between the aforementioned change deafness and how our brain manages its attentional resources, to the extent that really complex scenes degrade severely our detection capabilities. In addition, results suggest that we are more sensitive to the appearance of acoustic events in our environment, and evidence suggests that the process of detection takes place autonomously.

Consequently, there seems to exist a relationship between change detection and auditory saliency. From this hypothesis, in Section 4.1 we introduce our proposal to evaluate acoustic saliency detection algorithms. Some of the datasets that we use include labels for the latencies of the acoustic events, i.e both the starting (onset) and ending (offset) time of these events are available. Thanks to the experimental results that we have explained in this Section, which agreed that our brain is more sensitive to the appearance of these events, we conclude that the onset times available in these datasets are more suitable to measure the performance of acoustic saliency detection algorithm.

3.3 Salience mechanism

From a physiological perspective, this ability to detect prominences in the vast flow of sensory information takes place in a neural area usually called Salience Network. According to [Menon, 2015], it is formed of the Anterior Insula (AI), the dorsal Anterior Cingulate Cortex (dACC) and some other areas of the brain that are in charge of perceiving corporal signals. Essentially, he proposed that through the AI our brain gets the information processed by our senses, such as vision or hearing, as well as some other body-related signals such as the heartbeat. The AI is in charge of detecting prominences in the stimuli that it has received from multiple sources and as a consequence Menon seems to suggest that this area is in charge of the detection of saliency. On the other hand, the dACC is in charge of producing motor responses and reactions for the signals processed by the AI, since it is connected to the spinal cord. Both elements are connected through a special type of neurons referred to as *von Economo neurons*, which thanks to their shape and structure allow to send quicker impulses than regular neurons allowing a fast communication between the AI and the dACC. As a consequence, the saliency network proposed by [Menon, 2015] has the capability of acquiring and processing sensory and internal information, but it is also capable of producing quick physical responses when there is a prominence in the environment.

3.4 Applications of saliency and attention

From a more technological point of view, in the recent decades there have been several proposals of algorithms that somehow exploited the concept of saliency. Some of them were described by [Li and Gao, 2014]:

Image and video retargeting, where the main goal consists of cropping some regions of an image or frame of video such that its global resolution is modified, keeping as much of the supposedly relevant content as possible and removing the rest. For this particular task, saliency is used as a descriptor that indicates what content of an image or frame can be actually relevant for a person.

Advertising, an application that arose due to the massification of online videos. Saliency can play multiple roles, from being used to append publicity in sections of video that are contentless and irrelevant for the spectator as well as for introducing relevant advertisements in salient frames that are related to an announced product. Consequently, there are many posible ways to take advantage of saliency in such applications.

Image retrieval, where a query image is used to obtain some other similar images. From the query, saliency is used to extract the most prominent objects that make up the scene. Then, a set of candidate images is obtained, with the condition that all of them should depict similar objects. The retrieved image is the candidate that shares more similarities with the query image. In this case, the base hypothesis is that an image can be adecquately represented by its depicted objects, which should indeed be salient respecting to the background.

Video summarization, where a set of frames is selected so that their content allows the spectator to get an overall idea about a whole video. There are multiple roles for saliency in summarization algorithms: in addition to alternative techniques of image and video processing such as face detection and camera motion estimation, visual attention models are commonly used to determine salient regions in frames considering both static and dynamic content, as well as auditory saliency detectors to discern the frames where spontaneous acoustic changes appear.

Compression, where saliency maps are once again used to determine the relevant areas of an image. This information shall be used directly to customize the parameters of a non-uniform quantizer for each region under analysis.

Syllable detection in speech, as [Kalinli and Narayanan, 2009] proposed in their work, where they designed an auditory saliency model that was capable of detecting prominent syllables and words in speech.

Onset detection, a computational task that is commonly employed in Acoustic Event Classification and Detection (AEC/D) and consists of measuring the latency when an acoustic event appears (onset) or disappears (offset). According to our previous works [Rodríguez-Hidalgo et al., 2018a; Rodríguez-Hidalgo et al., 2018b] acoustic saliency algorithms could be compatible with this task, since they detect events with certain degree of robustness against noise.

In this work we are particularly interested in saliency detection algorithms for audio, which generate saliency cues that try to resemble and highlight the content emulating the behavior of human attention mechanisms.

3.5 Attention algorithms for vision

In the recent times there have been several proposals for the computational detection of visual saliency. In fact, in the last decade the performance of these vision algorithms has taken huge steps thanks to the recent developments in machine learning and particularly in the field of deep learning. For more information, we recommend [Borji and Itti, 2013; Borji et al., 2019] and the proposals evaluated at MIT saliency benchmark [Bylinskii et al., 2019a] to see the most recent results.

However, the huge progress of the state-of-the-art is largely due to the collection of numerous labeled corpora that can be found online nowadays (See the datasets section at the MIT saliency benchmark, [Bylinskii et al., 2019a]). One of the common aspects of the previous datasets is that they rely on the usage of a specific device, popularly known as eye-tracker. As its name suggests, eye-trackers are particularly designed to accurately capture the eye's motion and are used by experts to design cognitive visual tasks were data is acquired from human test participants. Consequently, the quality of these datasets is critical for the performance of top-down attentional models for vision.

Nevertheless, in this work we pay special attention to bottom-up algorithms some of which later inspired their auditory counterparts. As we introduced at the beginning of Chapter 3, bottom-up attention is mostly feature-based and automatic. In addition, some of the classical models of visual saliency such as [Itti et al., 1998] and [Harel et al., 2006] are unsupervised, since labeled datasets with eye fixations were not so common as nowadays. These models are quite relevant for our work since their development helped to design auditory saliency algorithms, and some of them are directly inspired in their visual counterparts.

3.5.1 Itti saliency model

This model depicted in Figure 3.5 was originally proposed by [Itti et al., 1998]. The input image is decomposed into three different sets of features: intensity, four colors and four orientations. The colors are red, green, blue and yellow, whereas the orientations are computed using four different angles. Every feature is decomposed into several resolution levels forming a pyramid. This means that every feature will be available with different levels of detail. Then, a center-surround mechanism is used with consecutive scale levels, emulating how retinal ganglion cells behave and producing big responses to the activity perceived in a small area, whereas its surrounding is inhibited. This allows our visual system to distinguish discontinuities in objects rather easily, and as a consequence is useful for the detection of prominent objects. The center-surround mechanism tries to emulate such behavior by differentiating the content from finer scales against their coarser

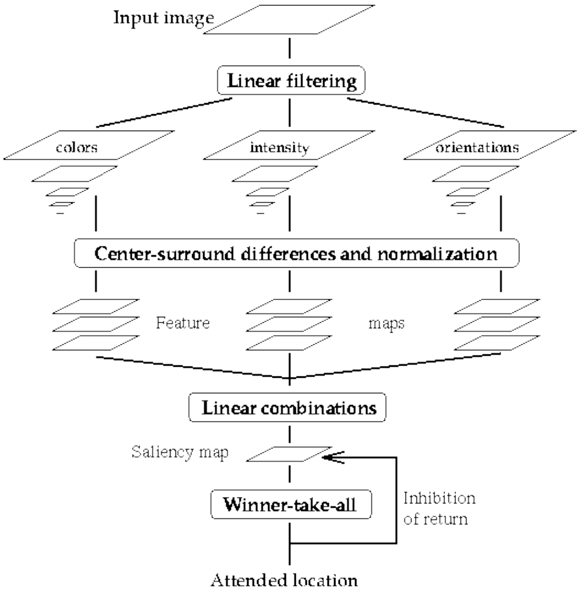


Figure 3.5: Itti saliency model. Figure from [Itti et al., 1998]. Copyright © 1998 IEEE.

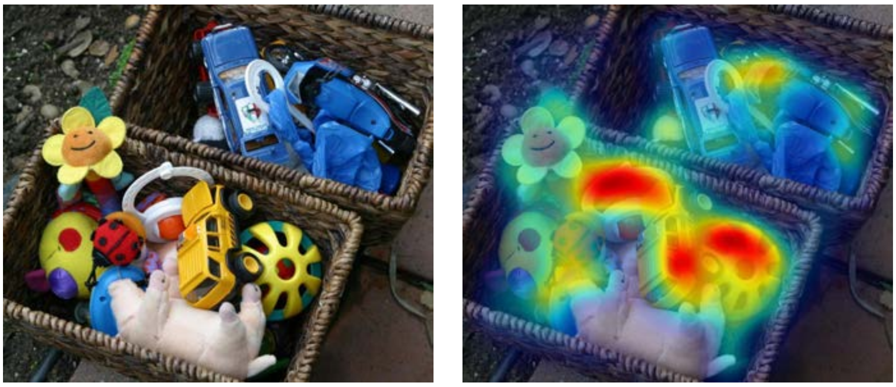
counterparts, and tries to mimic lateral inhibition (see Section 2.3).

These computations are performed independently for each of the aforementioned features. Afterwards, these multi-scale feature maps are normalized by means of a handcrafted operator that measures local maxima and their relative size in comparison with those of their surrounding. If the difference is sufficiently big, then the maxima is set to be prominent, whereas when there is no significant difference the local area is ignored. This process is repeated for every center-surround scale and the results are finally combined to provide a pool of three conspicuity maps, one for each possible feature. Then, these conspicuity maps are re-normalized and finally combined to constitute a global saliency map. Finally, a winner-takes-all neural net is implemented so a particular location is picked as the salient area candidate.

A visual saliency map obtained with the model of Itti would look like the plot in Figure 3.6, where the heat map represents the provided saliency map overlaying the original image.

3.5.2 Graph Based Visual Saliency (GBVS)

[Harel et al., 2006] observed that most of the times saliency models were divided into three different stages: feature extraction, generation of activation maps, and normalization. They proposed a Markovian approach for the last two stages.



(a) Input image.

(b) Itti's saliency map

Figure 3.6: Example of the outcome of Itti's saliency algorithm, represented as a heat map. The image is a sample that can be found in the implementation developed by [Harel, 2012].

For the computation of the activation maps they considered a specific feature obtained directly from an incoming image, such as color, intensity, etc. They proposed a dissimilarity metric that was computed for every pixel of the feature, which was then compared with their neighboring pixels. This dissimilarity constraint penalized pixels with similar neighbors, and also considered that the closer they were the smaller weight they should get. This procedure provided a weight for every individual pixel respecting to the rest of them, which should be normalized to sum one. As a consequence, the authors propose to consider this structure as a Markov Chain, where each feature pixel is considered to be a node that has a vector representing the transition probability to the neighboring areas. After conforming a transition matrix with these probability vectors the equilibrium state can be obtained for each pixel, and the most prominent areas should actually get a higher probability since they are the ones complying the dissimilarity constraint.

A similar procedure was performed for the normalization stage, where the dissimilarity function was modified and the rest of the process repeated. As a result, probabilities (or as the authors name them, masses) got concentrated into a small number of locations of the feature.

As the MIT benchmark shows [Bylinskii et al., 2019a], this system proved to be more effective detecting eye fixations than the algorithm previously proposed by [Itti et al., 1998].

3.5.3 Visual Attention Topic Model (ATOM)

This proposal denominated visual Attention Topic Model (ATOM) is based on topic modeling. The model that [Fernández-Torres et al., 2016; Fernández-

Torres et al., 2019] propose is generic and independent of the application scenario, and is supported by the assumption that, given a video frame, visual attention can be decomposed into several topics or sub-tasks which, in turn, are represented as combinations of either low-level visual descriptors, such as color and motion, or features derived from high-level concepts such as faces or text. This supervised model draws on the information provided by human fixations during its learning phase, with a two-fold objective. First, it allows to obtain comprehensive interpretations of visual attention in different contexts, learning both attracting and inhibiting sub-tasks. Second, it estimates visual attention in each spatial location as a logistic regression over the sub-tasks learned.

3.5.4 Models based on artificial neural networks

Nowadays, the state-of-the-art in visual saliency estimation has drifted towards supervised models based on neural networks, which most of the times make extensive use of Convolutional Neural Networks (CNN), a bidimensional structure that performs proficiently for the processing of images and videos. Three remarkable examples are [Kruthiventi et al., 2017; Kümmerer et al., 2017; Jia, 2018], which were the top performers for most of the scores proposed in the MIT saliency benchmark [Bylinskii et al., 2019a]. We recommend the book [Goodfellow et al., 2016] to acquire a more detailed understanding about neural nets and CNN, since the details about such techniques are out of the scope of our work.

The majority of the models proposed nowadays are trained using eye fixations datasets, whose labeling is made thanks to the aforementioned eye-trackers and using several human participants. In addition, due to the diversity of evaluation metrics available for this particular visual task, none of the previous techniques seems to be the top-performer in all of them. We recommend the work of [Bylinskii et al., 2019b] for a deeper analysis about the evaluation of such visual models. It can be concluded that it is a quite active area of research where new proposals are presented frequently and visually realistic saliency maps are produced.

3.6 Attention algorithms for audio

This Section is devoted to introduce some popular auditory saliency detection techniques. With saliency detection we refer to bottom-up attention models that are used to obtain prominent features from audio. It should be noticed that the majority of them are unsupervised models, since as we mentioned in Section 4.1 there are no datasets that can be used to train algorithms to model auditory saliency specifically.

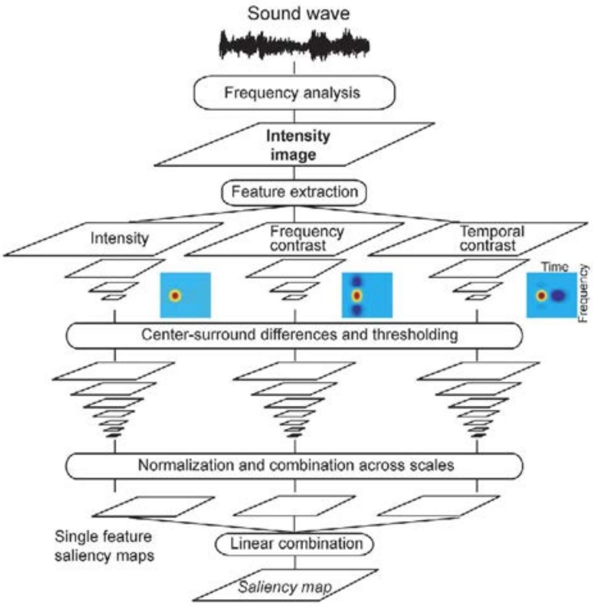


Figure 3.7: Structure of Kayser's saliency model for an audio signal, as in [Kayser et al., 2005]. Notice the similarities with the schematic proposed by [Itti et al., 1998] for vision.

3.6.1 Kayser saliency model

This model was proposed by [Kayser et al., 2005] as an auditory saliency detector following the steps in Itti's visual saliency algorithm [Itti et al., 1998]. As depicted in Figure 3.7, it functions using a spectro-temporal representation of the incoming signal as if it was an image. Then, in the feature extraction stage, three different features were computed: intensity, frequency contrast and temporal contrast. The intensity feature is computed using a Gaussian bidimensional filter, centered at the frequency and time of interest for each point of the spectrogram. The frequency contrast is computed using the combination of three bidimensional Gaussian filters. One of them would be placed in the frequency and time of interest and the other two, symmetrically located in the frequency axis around the first filter implementing a negative magnitude that would resemble inhibition mechanisms. With this design, we expect the point of interest to be compared, or contrasted, directly with the surrounding frequency information. Finally, the temporal contrast feature includes a Gaussian filter in the point of interest followed by a time-delayed inhibition area. Similarly with the frequency contrast, this mechanism should get potential differences with delayed temporal information.

Since this is a multi-scale approach, this feature extraction process is computed for different scales of the spectrogram, whose feature maps are later combined by means of a centre-surround mechanism. The centre-surround concept is directly inherited from human retinal ganglion cells, and determines the differences between the point of interest in an image and its surrounding information. In the case of audio, the center surround mechanism is used to resemble the inhibition effect explained for auditory neurons by [Schreiner et al., 2000]. Then, signals from different scales are normalised and combined, giving as a result three different feature maps that are finally combined linearly to provide an auditory saliency map. Since the authors of the original paper did not need a temporal saliency signal but rather an acoustic saliency map, we will obtain it by averaging this saliency map through all its frequency components.

The performance of this algorithm was originally assessed by means of subjective questionnaires for human subjects. First, test subjects received two signals, one per ear, and were asked which one they thought it was more interesting. On a second test authors tried to measure the amplitude level required to detect the appearance of a salient sound under noisy background audio. In addition, they performed some tests with animal subjects by means of measuring the pose of their heads when two different sounds were perceived in each of their ears.

This is one of the techniques chosen as part of the experimental benchmark of this work, and the original code used to compute the saliency map can be directly obtained from [Kayser, 2018].

3.6.2 Kalinli saliency model

This particular model was proposed by [Kalinli and Narayanan, 2007; Kalinli and Narayanan, 2009] and is a modification of Kayser saliency model. The multi-scale structure where several scales of a specific feature are computed and combined with the centre-surround mechanism remains the same, as it is depicted in Figure 3.8. However, the authors added two new features to the global scheme in addition to the three original ones: orientations and pitch distribution. Orientations are computed using bidimensional filters similar to the frequency contrast filter depicted in Figure 3.7, but rotated with angles $\theta = \{45°, 135°\}$, which according to the authors "mimic the dynamics of the auditory neuron responses to moving ripples" (see [Kalinli and Narayanan, 2009]). Additionally, pitch distribution is computed under the assumption that the brain obtains it by using the autocorrelation.

The scheme depicted in Figure 3.8 also shows that they extract gist features by means of PCA that they later on use to train a top-down saliency model for prominent syllable detection. However, since we want to model saliency and not a task-specific model such as the previous one we use their initial version of the algorithm, explained in [Kalinli and Narayanan, 2007]. The differences are not very noteworthy, since instead of the gist extractor they implemented a multi-scale normalization and combination structure

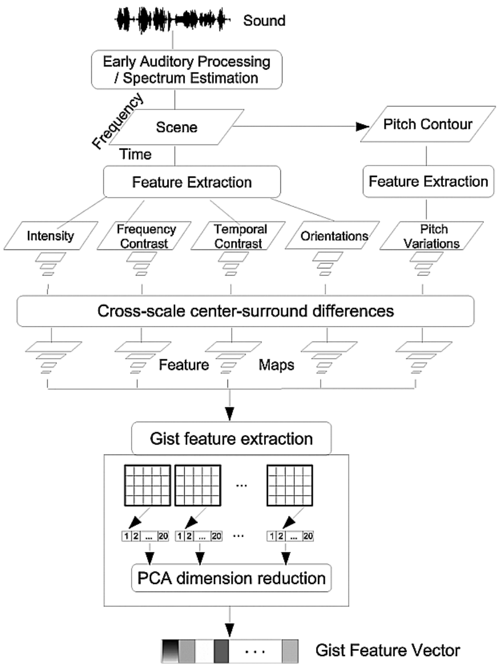


Figure 3.8: Kalinli saliency model. Figure from [Kalinli and Narayanan, 2009]. Copyright © 2009 IEEE.

such as the one designed by [Kayser et al., 2005] to finally produce a saliency map.

They tested the performance of their proposal measuring its capability to detect syllable and word stress directly from speech, using precision, recall and F-scores. This is another proposal that we included in our test benchmark, and the code used can be found at [Macaluso, 2018].

3.6.3 Acoustic Saliency Using Natural statistics (ASUN)

To the extent of our knowledge this is the first acoustic saliency detection algorithm to use a cochleogram, computed using the Gammatone filters described in Section 2.2. The schematic of this approach is depicted in Figure 3.9, where it can be observed that the cochleogram frequency bins are grouped and projected using Principal Component Analysis (PCA). The short-term and long-term blocks illustrate the computation of saliency as described by [Tsuchida and Cottrell, 2012]:

$$s(n) \propto -\log P(F_n|F_{n-1}, \cdots, F_{n-k}) - \log P(F_n|F_{n-k-1}, \cdots), \tag{3.1}$$

where F_n represents the feature used to compute saliency for a specific frame n and $s(n)$ is the output saliency signal. The first probability models audio statistics from a recent timespan using k previous frames of data. It is computed using a Gaussian mixture model using local information and reestimated after a quarter of second. The second probability models *lifetime*

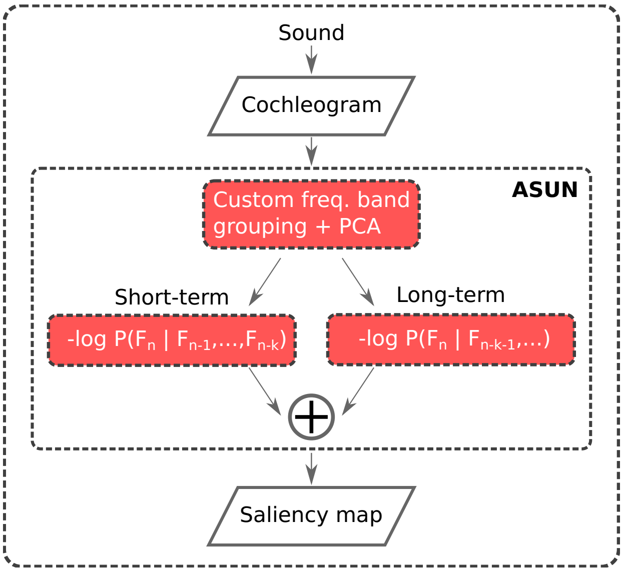

Sound
Cochleogram
ASUN
Custom freq. band grouping + PCA
Short-term
Long-term
-log P(F$_n$ | F$_{n-1}$,...,F$_{n-k}$)
-log P(F$_n$ | F$_{n-k-1}$,...)
Saliency map

Figure 3.9: ASUN saliency model, where after the cochleogram there is a stage that combines a customized grouping of frequency bands with PCA. Diagram inspired by Figure 1 from [Tsuchida and Cottrell, 2012].

information and is obtained fitting another Gaussian mixture model using four datasets of natural auditory events. Roughly speaking, this algorithm would compute the probability of being salient for a specific feature F_n locally for a specific memory length k, and also for a long-term register that is expected to model lifetime information.

The structure of this algorithm resembles a biological system, since it includes both long-term and short-term registers of memory. However, the influence of the long-term register might be excessive, since its estimation capabilities depend directly on the quality of the data used during the training stage, and consequently will perform better with signals of a similar nature as the auditory files used in the aforementioned stage. For this reason, we did not considered this model during our later analysis and discarded it in favour of totally unsupervised systems.

3.6.4 Kaya's saliency model

The model proposed by [Kaya and Elhilali, 2014] was based on the usage of the raw signal to extract features such as the envelope to obtain the intensity, harmonicity, spectrogram of the signal (splitted into low and high frequencies), bandwidth and temporal modulation. This feature set allowed to obtain information from the signal such as intensity, pitch and timbre, in addition to detailed spectral information. Globally, they obtained 6 different features.

Each of the features is processed by a Kalman filter set to predict the value of the current frame of data. When the prediction made by a filter for one of the features differs from its actual content an anomaly is said to occur. As a consequence, six signals representing anomalies are obtained from six Kalman filters, which need to be combined to produce a saliency score. They manage to do it thanks to a weighted sum, which by means of logistic regression produces a single output temporal signal representing the probability of occurrence of a salient event.

Interestingly, they evaluated the performance of their system by comparing its results with human participants by means of a correlation coefficient and the Receiver Operating Characteristic curve (ROC), which helped to determine the detection capabilities of the system.

3.7 Related methodologies for detection

Since in this thesis we introduce a new acoustic saliency detection algorithm it seems sensible to compare it with other saliency models from the state-of-the-art, like the ones presented in the previous section. However, there are alternative techniques that, in spite of being originally designed for other purposes, might fit the evaluation requirements we need to assess our proposal. Some of these techniques are introduced in this Section, two of which were particularly designed for the detection of speech segments in noisy sequences. The others were specifically designed to detect onsets in sequences of audio containing music. All of them fit the onset evaluation criteria that we proposed for this work (see Chapter 4).

3.7.1 Energy thresholding

As its name states, this technique consists of thresholding the energy directly from the acoustic signal. In our case, since we use spectro-temporal representations it is more interesting to compute energy directly from the spectrogram, where the input waveform has been transformed into the Fourier domain using the STFT as explained in Section 2.4.1. By means of Parseval's theorem (See [Huang et al., 2001]) it is possible to obtain the energy both from temporal and frequency domain. As a consequence, the energy for the whole signal can be computed as follows:

$$E_x = \sum_{m=-\infty}^{\infty} |x(m)|^2 = \frac{1}{E_w} \sum_{k=1}^{n_{fft}} \sum_{n=-\infty}^{\infty} |X_c(k,n)|^2, \tag{3.2}$$

where it is stated that the energy of the signal $x(m)$ in the temporal domain is equivalent to the energy summation of each one of the STFT vectors obtained for each window, divided by the energy E_w of such window. This means that the energy of the signal is conserved in both domains. The equivalence for a particular frame n would be defined according to:

$$E_{X(n)} = \frac{1}{E_w} \sum_{k=1}^{n_{fft}} |X_c(k,n)|^2, \tag{3.3}$$

Once the energy is computed for all the temporal frames n from the spectrogram a threshold is set to be the average of these energies. Any value that gets over this static threshold is considered to be a positive detection, which gives a binary signal that allows to easily measure the latency of both onsets and offsets. Of course, the simplicity of this technique is at the same time its main disadvantage. A straightforward analysis shows that its performance is severely biased by the length of the sequence under study and the length of the silences or background noise segments. For example, if an audio sequence is crowded by many segments of background noise, the threshold value would decrease to the extent that some portions of the background signal could be detected as positive activity. On the contrary, having little or no background noise at all might bias the threshold value towards the loudest acoustic events, which might degrade the detection of acoustic events that have lower acoustic intensity levels.

3.7.2 Voice Activity Detector (VAD)

As its name states the purpose of this kind of detector consists of locating speech in an auditory signal, including both the moment where the oral production begins and when it ends.

Voice Activity Detectors (VAD) have several applications. Some of them belong to the field of speech processing, such as the removal of silent frames from speech segments, avoiding the processing of useless data in tasks such as speech encoding and speaker recognition among many others. It is also widely used to reduce the energy consumption of telecommunication systems when there is no need, such as for silence areas where systems could avoid to send data to save energy or reduce their bit rate.

In this work a VAD implementation from the Voicebox Toolbox [Brookes, 1997] is used. According to the original paper proposed by [Sohn et al., 1999], this VAD is based on a Likelihood Ratio Test (LRT) to compare two hypothesis H_0 and H_1 representing non-speech and speech data, respectively, for each frequency bin k of the input spectrogram $X(k,n)$:

$$\Lambda(k,n) \triangleq \frac{p(X(k,n)|H_1)}{p(X(k,n)|H_0)}. \tag{3.4}$$

The authors model each one of the n_{fft} coefficients using the Gaussian distribution, and the LRT becomes:

$$\log \Lambda(n) = \frac{1}{n_{fft}} \sum_{k=1}^{n_{fft}} \log \Lambda(k,n) \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3.5}$$

Finally, as [Ramirez et al., 2007] state in their work, smoothing the outcome signal can help to improve the robustness of detection against the noisy background. The VAD proposed by [Sohn et al., 1999] introduces an HMM-based scheme, which uses information from previous temporal frames as well as the current one to modify the outcome of the LRT previously presented, assuming that consecutive frames of speech should be correlated.

3.7.3 Spectral flux (SF)

According to [Dixon, 2006] the spectral flux (SF) is defined as follows:

$$SF(n) = \sum_{k=1}^{n_{Mel}} H_w(|X_c(k,n)| - |X_c(k,n-1)|) \tag{3.6}$$

where $X_c(k,n)$ represents the complex spectro-temporal representation under analysis and $H_w(x) = \frac{x+|x|}{2}$.

This technique is generally used for onset detection in Musical Information Retrieval (MIR), and it subtracts the magnitudes of the spectral representations of two consecutive frames. Notice that this subtraction is performed independently for each one of the frequency bands k. This difference is finally rectified by the function $H_w(x)$, which after adding up the result for all the frequency components produces a temporal signal.

This technique belongs to the set of onset detection techniques, which are specifically conceived for this task. These algorithms usually include thresholding and peak-picking detection techniques that help to determine which of the detected peaks are actually representing signal onsets, and we include an example in Section 7.2. These methodologies are also used for Phase Deviation and Complex Domain, which are explained in the following subsections.

3.7.4 Phase Deviation (PD)

In order to define Weighted Phase Deviation (WPD) and Normalized Weighted Phase Deviation (NWPD) it is necessary to introduce Phase Deviation (PD) first, an algorithm designed to detect onsets directly from the information coded in the phase of a spectro-temporal representation. The incoming signal, normally a spectrogram, is defined as follows according to [Dixon, 2006; Böck et al., 2012]:

$$X_c(k,n) = |X_c(k,n)| \cdot \exp(j \cdot \psi(k,n)) = X(k,n) \cdot \exp(j \cdot \psi(k,n)), \tag{3.7}$$

where $\psi(k,n)$ represents the instantaneous phase of the signal and we define $X(k,n) = |X_c(k,n)|$. The first derivative of the phase represents the instantaneous frequency of the signal, and for this implementation is computed as the difference in phase between two consecutive temporal frames:

$$\psi'(k,n) = \psi(k,n) - \psi(k,n-1). \tag{3.8}$$

Then, changes in the instantaneous frequency of the signal are computed with an additional derivative:

$$\psi''(k,n) = \psi'(k,n) - \psi'(k,n-1). \tag{3.9}$$

Similarly to what occurs in image processing, computing the second derivative for each one of the frequency bands should highlight the temporal frames where great changes in the phase occur. Finally, in order to summarize the data from all the frequency bins the average of their instantaneous frequency derivative is computed, which produces the so-called Phase Deviation:

$$PD(n) = \frac{1}{n_{Mel}} \sum_{k=1}^{n_{Mel}} |\psi''(k,n)|. \tag{3.10}$$

The **Weighted Phase Deviation (WPD)** is computed as follows:

$$WPD(n) = \frac{1}{n_{Mel}} \sum_{k=1}^{n_{Mel}} |X_c(k,n) \cdot \psi''(k,n)|, \tag{3.11}$$

considering that each one of the frequency bands from the second derivative of the phase is multiplied by the original spectra, so the derivative components are directly weighted by the energy of the corresponding band of the original signal. Thanks to this modification, frequency bins that are not relevant in the spectra since they have a small magnitude reduce their influence in the Phase Deviation, which should indeed reduce the noise added by irrelevant bins (see [Bello et al., 2005]).

Finally, **Normalized Weighted Phase Deviation (NWPD)** is obtained similarly to WPD. However, rather than normalizing it by the number of frequency components n_{Mel} it is normalized by the summation of the magnitudes of frequency components:

$$NWPD(n) = \frac{\sum_{k=1}^{n_{Mel}} |X_c(k,n) \cdot \psi''(k,n)|}{\sum_{k=1}^{n_{Mel}} |X_c(k,n)|}. \tag{3.12}$$

As a consequence, it can be observed that WPD and NWPD seem to be proportional, although the former is divided by a constant value and the latter is normalized frame by frame considering the frame absolute magnitude. Consequently, their outcomes might be totally different.

3.7.5 Complex Domain (CD)

This technique detects onsets by means of the Euclidean distance of the complex spectro-temporal representation $X_c(k,n)$ and an approximation of itself computed using data from the previous frame, called $X_T(k,n)$. According to [Duxbury et al., 2003; Dixon, 2006; Böck et al., 2012], the procedure would be computed for each bin according to:

$$CD(n) = \sum_{k=1}^{n_{Mel}} ||X_c(k,n) - X_T(k,n)||_2. \tag{3.13}$$

The function $X_T(k,n)$ is predicted following the assumption stated by the original authors [Duxbury et al., 2003]: during steady segments of the signal, magnitude and phase should remain constant, or at least quite similar, from one frame to its consecutive one. This assumption should hold while no sudden onset occurs, where no spectral change could be expected. As a consequence, the magnitude of the estimation is defined as:

$$|X_T(k,n)|= |X_c(k,n-1)|, \tag{3.14}$$

where it was assumed that:

$$|X_c(k,n)| \backsimeq |X_c(k,n-1)|. \tag{3.15}$$

Respecting to the phase estimation and considering the same assumption, instantaneous phase should be similar in consecutive steady frames as depicted below:

$$\psi(k,n) - \psi(k,n-1) \backsimeq \psi(k,n-1) - \psi(k,n-2). \tag{3.16}$$

Since we expect to estimate the phase $\hat{\psi}(k,n)$, and working with Equation 3.16, it can be deduced that:

$$\hat{\psi}(k,n) \backsimeq \psi(k,n-1) + \psi(k,n-1) - \psi(k,n-2), \tag{3.17}$$

which thanks to Equation 3.8 turns into:

$$\hat{\psi}(k,n) \backsimeq \psi(k,n-1) + \psi'(k,n-1), \tag{3.18}$$

where the estimation of the phase in the temporal frame n is obtained as the summation of the phase information from the previous frame plus its instantaneous frequency.

Considering that the prediction is defined as:

$$X_T(k,n) = |X_c(k,n-1)| \cdot \exp(j\hat{\psi}(k,n)), \tag{3.19}$$

it holds that it can be computed as follows:

$$X_T(k,n) = |X_c(k,n-1)| \cdot \exp(j(\psi(k,n-1) + \psi'(k,n-1))). \tag{3.20}$$

With the estimated signal $X_T(k,n)$ and the original $X_c(k,n)$, Equation 3.13 would be used to obtain the Complex Domain signal.

Chapter 4

Evaluation

Assessing the performance of a system needs to ponder several factors. In Sections 3.6 and 3.7 we introduced the techniques that we consider adequate to conform a benchmark to test auditory saliency detection performance. It included both saliency algorithms and onset detection techniques, some of them designed for MIR. As we already mentioned, evaluating salience from a computational perspective is not straightforward. In this Chapter we intend to thoroughly explain the evaluation decisions we made during the course of our work, including the selection of tasks and datasets.

4.1 Evaluation proxy task

Section 1.1 introduced some of the challenges associated to auditory saliency, including the lack of ad hoc datasets and metrics to compare alternative models. Nevertheless, several authors have found creative ways to overcome this limitation.

As we explained in Section 3.6.1, [Kayser et al., 2005] used subjective questionnaires where the participants were assigned two tasks: first, they were asked to state which of two complex scenes was more salient and second, they had to detect the appearance of a salient acoustic event within a noisy background. They also performed an experiment where they validated the performance of their system measuring the head pose of some animals. [Kalinli and Narayanan, 2009] used speech datasets and their prosody labels as a proxy, considering that prominent words should have a certain prosody level. On the contrary [Tsuchida and Cottrell, 2012] proposed a behavioral study similar to the one proposed by [Kayser et al., 2005], where test subjects listened first to a white noise sequence and then two different acoustic sequences were presented, one per ear. Their task consisted of determining which sequence seemed to be more interesting.

[Schauerte and Stiefelhagen, 2013] used Acoustic Event Classification datasets where temporal marks were provided for each possible class. They

		Prediction	
		pos	**neg**
Ground truth	**pos**	True Positives (TP)	False Negatives (FN)
	neg	False Positives (FP)	True Negatives (TN)

Table 4.1: Confusion matrix that helps to understand the terms used to compare prediction results and ground-truth data.

assumed that in comparison with background sounds the existence of any acoustic event should be salient, and they measured the performance of their saliency system using the F-score. Alternatively, some recent works (see [Zhao et al., 2018]) showed that there might be a correlation between micro-saccadic movements and bottom-up auditory saliency, meaning that auditory saliency could be measured using eye-trackers.

For the results of this work we rely on the evaluation setup that we proposed in our previous works [Rodríguez-Hidalgo et al., 2016; Rodríguez-Hidalgo et al., 2018a; Rodríguez-Hidalgo et al., 2018b]. We use a proxy similar to the one of [Schauerte and Stiefelhagen, 2013], where AEC/D datasets were used to measure saliency.

Our hypothesis is that, as we explained in Section 3.2, humans are sensitive to the appearance of acoustic events in their environment, a behavior that seems to be related to bottom-up saliency since our reaction is quick and automatic. In fact, the temporal instants when these events appear are commonly denoted as onsets. Consequently, we proposed to use the onset marks from acoustic events to measure the global performance of our auditory saliency detector and the rest of the systems included in our benchmark.

4.2 Evaluation metrics

Similarly to the works of [Kalinli and Narayanan, 2009] and [Schauerte and Stiefelhagen, 2013], we evaluate our systems in terms of their F-score, that can be obtained according to the following equation:

$$F = 2 \cdot \frac{P \cdot R}{P + R}, \tag{4.1}$$

which represents the harmonic mean between precision P and recall R scores[1]. Following the notation of the confusion matrix illustrated in Table 4.1, precision can be computed according to:

$$P = \frac{TP}{TP + FP}, \tag{4.2}$$

where TP and FP refer to the number of true positive and false positive samples, respectively.

On the other hand, recall is obtained with the following equation:

$$R = \frac{TP}{TP + FN}, \tag{4.3}$$

where FN refers to the number of false negative samples.

However, we need to take into account that the manual annotation of the onsets (and also, the offsets) of an acoustic event is often a difficult task. In fact, depending on the nature of the audio event and the labeling criteria adopted, there might be an ambiguity in the determination of the exact position of the onset latency of the event. Consequently, we follow the proposal of [Mesaros et al., 2016a] for the DCASE-2016 challenge where a value is accepted as *true* if it falls within a neighboring area of the ground truth onset label. This means that if the computed events are detected close enough to the ground truth data, they are accepted as valid. The value proposed for the aforementioned challenge is a window of 200 ms, which we adopted for our analysis.

4.3 Datasets

This section is devoted to the description of the datasets that were used during the development of this work. Every dataset and condition studied in this work share some common aspects. First of all, although most of them usually are available with a sampling frequency of $F_s = 44.1$ kHz we resampled all the files to $F_s = 22$ kHz. This approach was used due to two different reasons: first of all, by halving the sampling frequency we reduce proportionally the size of the data files that are going to be processed, which affects directly to the computational time. Secondly, we share a common sampling frequency for all the datasets since some of them are not available at $F_s = 44.1$ kHz.

[1] Different weights for P and R are also possible but we do not have any a priori motivation to emphasise one or the other. Therefore we equally weight both metrics.

There are three datasets that were originally designed for the tasks of Acoustic Event Classification/Detection (AEC/D) and all of them share their labelling structure, being usually composed of the onset and offset times and followed by the label of the acoustic event under analysis. Finally, since this work does not include any binaural analysis we only considered monophonic audio, averaging the two available channels for stereophonic datasets and picking up the most representative microphone for those whose audio was acquired with microphone arrays.

In addition, since one of our objectives is to verify the robustness of the presented techniques under noisy circumstances, we also employed a dataset with environmental signal recordings that will be used to contaminate the previously mentioned AEC/D datasets.

4.3.1 DCASE

In the recent years some researchers whose works are focused on the processing of auditory signals have organized challenges where participants are encouraged to solve some specific tasks. That is the case of *IEEE AASP Challenge Detection and Classification of Acoustic Scenes and Events 2016* (DCASE) [Mesaros et al., 2016b], where the tasks to solve lie within the topic of AEC/D. In fact, to the date of this work there have been more recent challenges that included new tasks. Interestingly, the organizers provide both datasets and baseline models, so participants have the chance to start working on their preferred task rather quickly.

In this work we focus on the 2016 event, where four different tasks were proposed. The first task consisted on the classification of acoustic scenes, where participants had to develop systems capable of classifying an incoming audio as a particular scenario from a specific pool. The second and the third tasks were designed for the detection and classification of acoustic events, and both of them included onset/offset labels. The fourth task consisted of audio tagging, where the only labels for each audio represented the names of the sounds that were played.

We chose task 2 dataset and kept only the onset labels, since classes and offsets were useless for the evaluation proxy that we introduced in Section 4.1. We denoted this task *DCASE-T2*, and it comprises 72 auditory clips with different Event-to-Background ratios (EBR), going from -6 dB to 6 dB. The EBR is a score that defines how easy it is to distinguish an auditory event from the background noise, in a similar fashion to the Signal-to-Noise ratio (SNR). This particular subset was automatically generated by software, including audio files and around 2000 labels for onset, offset and classes. We divided it into two different subsets, a small one composed by 18 files that we used for the validation of the parameters of our saliency detector, and a bigger one with 54 files that we kept exclusively for evaluation. In addition to the aforementioned EBR, *DCASE-T2* has two auditory modalities: polyphonic, where several event classes overlap in time and can be perceived concurrently, and monophonic, that occurs when classes do not overlap and can be perceived independently. Regarding their recording characteristics we need to say that data for both subtasks were acquired with a sampling frequency of F_s = 44.1 kHz and a resolution of 24 bits.

4.3.2 UPC-TALP

Commonly known as *UPC-TALP database of isolated meeting-room acoustic events* (UPC-TALP), it is a part of the European project Computers In the Human Interaction Loop [Waibel and Stiefelhagen, 2009] (CHIL), whose main goal was to develop computer systems to assist people reducing the necessary interaction between machines and users, so the latter could focus on interacting with other people instead of having to handle a machine.

Some of the proposed tasks for the project were *automatic speech recognition*, which is clearly related to auditory information, *person tracking* considering auditory, visual and multimodal cues as a combination of the two previous ones or *person identification* using acoustic and visual signals, non-verbal communication detecting gestures, body and head pose, among many others. In fact, audio and video cues were acquired using smart rooms and multiple arrays of microphones and fixed cameras, which allowed to get information from different positions and multiple sources. Two scenarios were considered: lectures and meetings. During lectures, there was a presenter talking and answering questions from an audience. For meetings, conversations of three to five people occurred around a table, where they were all speaking and interacting while their signals were acquired constantly.

Particularly interesting was the CLEAR06 evaluation [Temko et al., 2007] that was developed for the task of AEC/D, including originally three subsets of audio data. We focused on the UPC-TALP dataset of such evaluation, from which we have 30 labeled files comprising around 1000 acoustic events, considering that they follow the criteria that we require: all onset/offset and class labels are included for each of the audio files. Similarly to the rest of the datasets, the sampling frequency of this dataset is F_s = 44.1 kHz, with a resolution of 16 bits, and the audio was acquired with an array of microphones. We selected the data available from microphone three, since it was the one originally used to label events. Some of the classes available in this dataset are: *steps*, *phone ring*, *laugh*, *cough*, *door slam*, etc.

4.3.3 MIVIA datasets

MIVIA is a research group from the University of Salermo that made public two audio datasets related to the topic of surveillance and dangerous scenarios. For instance, their *MIVIA Audio Events Dataset* [Foggia et al., 2015] is composed of audios from three different classes: *glass breaking*, *gun* *shots* and *screams*, which could easily be found in a wide range of vigilance scenarios. In fact, the authors provided the data in a way that allows to train robust systems under noisy conditions, since the dataset includes audio files with different SNR values, with a total length of 6000 labeled events. Nevertheless, we discarded it since the background noise consisted of periodic repetitions of audio instead of complete sequences of stationary or non-stationary noise.

They proposed another dataset, called *MIVIA road audio events data set*, exclusively for the task of road surveillance [Foggia et al., 2014; Foggia et al., 2016], considering two different classes that are commonly found in accidents related with vehicles: *car crash* and *tire skidding*. This dataset does not include noise corrupted files as the previous one, and is formed by 57 files. Originally, this dataset was split into 4 different folds, each one formed by 100 acoustic events with their respective onsets, offsets and class labels. However, since our goal is not to train a supervised system we simply consider that the 400 events belong to a common fold, that will be used for evaluation purposes of the techniques of this work. In addition, it should be mentioned that the sampling frequency of the audio files is $F_s = 32$ kHz. We used this dataset and referred to it as MIVIA.

4.3.4 DEMAND

One of the obstacles that most of the systems designed to work with audio need to overcome is environmental noise, and its intensity and frequency content will define how it degrades the detection of other signals in the environment. That is a typical scenario, since all the acoustic environments are plagued by background noise, not to mention spontaneous and unexpected sounds that appear and were not considered originally during the design of the system under analysis.

As a consequence, we propose to test the robustness against noise of our benchmark. For this reason we consider the *Diverse Environments Multichannel Acoustic Noise Database* (DEMAND) [Thiemann et al., 2013], which was originally created as a set of noise sources acquired using an array of 16 omni-directional microphones. The recorded noise signals were classified into six different categories: domestic, nature, office, public, street and transportation, each one of them formed by three different noise files with a duration of 300 s. Two of the categories are related to indoor scenarios, domestic and office, and they include noises acquired in a kitchen, from a washroom, a meeting room or even the sound of people working in an office, among other scenarios.

On the other hand, the four remaining categories are related to outdoor activities, including the sound of a sports field, the sound of a busy underground station, a traffic intersection or the sound of a bus, among many other examples.

	DCASE-T2	**MIVIA**	**UPC-TALP**
Event Average Duration (s)	1.0123	2.1175	2.2487
Silence Average Duration (s)	3.1728	6.0029	3.8495

Table 4.2: Weighted average duration of silence gaps and events for each one of the datasets, obtained using the histograms and the durations of each condition.

Consequently, there are several possible noise sources in this dataset that we can use to pollute the AEC/D datasets mentioned in other Sections of this work, in addition to other sources such as white Gaussian noise. Regarding the multiple channels of the array used to acquire the sounds, the channel number one is chosen for all the noise files. Finally, each of the audio files was originally sampled with $F_s = 48$ kHz.

4.3.5 Label analysis

It seems reasonable to study the differences existing among the three AEC/D datasets used in this work, regarding the frequency of occurrence of events and silences, in addition to their corresponding durations. Figure 4.1 shows six histograms that approximate the duration of the events for each dataset (red color) and the duration of the silences that separate such events (blue color). For each of the datasets a different set of bins was used, so their fitness to the underlying data was appropriate. In addition, Table 4.2 conveys the weighted duration of acoustic events and silences for each dataset, which was computed using the histogram obtained for each one of the configurations.

These durations need to be taken into account, and they will vary depending on the dataset. On average, we observe that the duration of acoustic events tends to concentrate around 1 to 2 s for the three datasets, although there are some longer events in MIVIA and UPC-TALP. This, of course, is a property that exclusively depends on the nature of the sounds selected to produce the dataset. However, the three datasets have in common that events are separated by silent gaps. In the case of DCASE-T2 and UPC-TALP they tend to have a duration of almost 0 to 6 s, both of them averaging around 3 to 4 s, whereas for MIVIA gaps have a fixed duration of $[5, 6, 7]$ s, implying that there exist certain degree of periodicity that makes this dataset noticeably artificial, since in real life scenarios there is no prior knowledge about how long silences last. Consequently, we observe that MIVIA differs clearly with respect to the other two datasets, and this might influence the performance of the algorithms.

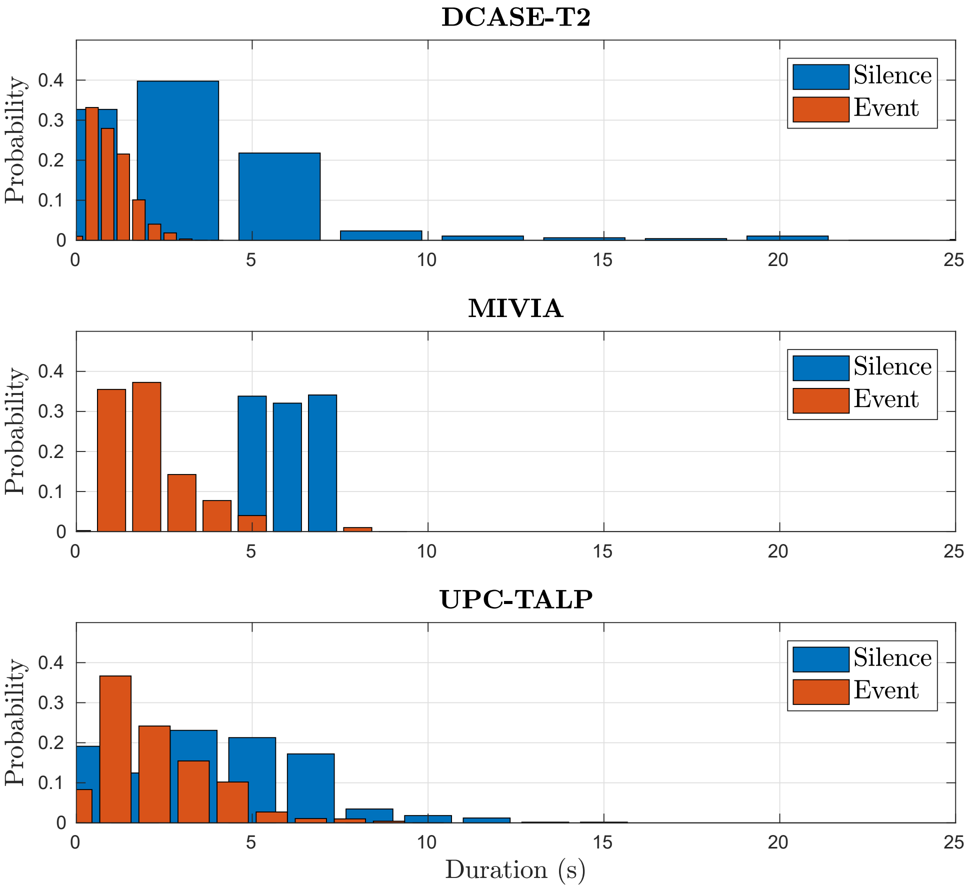


Figure 4.1: Histograms representing the duration of silence gaps and events for each of the datasets. Notice that we represent 10 bins for each of the configurations, and their width varies depending on the dataset and the duration of events and silent gaps.

4.4 Feature extraction

The two features used in this work are the spectrogram and the cochleogram, which were thorougly explained in Sections 2.4.2 and 2.4.3. For both methodologies we use a frame length of 20 ms with an overlapping factor of 50%. Moreover, for the particular case of the spectrogram the frequency bins are $n_{fft} = 1024$. On the contrary, the cochleogram is computed using $n_{fft} = 1024$ for the initial spectrogram and $n_{Mel} = 150$ for the Mel-filterbank, which finally provides our desired spectro-temporal representation. Notice that unless otherwise stated the default spectro-temporal representation used in this work is the magnitude of the cochleogram obtained using the Mel filterbank, represented as $X(k, n)$.

Chapter 5

Bayesian Log-surprise

As we explained in Section 3.6, some authors have developed acoustic saliency algorithms usually inspired in their visual counterparts. Some of them are direct adaptations, such as Kayser's saliency model [Kayser et al., 2005] and ASUN [Tsuchida and Cottrell, 2012], inspired in the works of [Itti et al., 1998; Zhang et al., 2008], respectively.

Another proposal is Bayesian Surprise, which to the extent of our knowledge was initially proposed by [Itti and Baldi, 2009], who developed it as a potential model for visual saliency detection. However, as the MIT visual saliency benchmark suggests [Bylinskii et al., 2019a], this technique was not as successful as other proposals like the well known GBVS proposed by [Harel et al., 2006] and more recent advances.

Nevertheless, some researchers reworked it to detect auditory saliency [Schauerte and Stiefelhagen, 2013], and their modifications allow the algorithm to compute a prominence signal for each one of the frequency components that comprise a spectro-temporal representation, which are finally combined to produce a temporal saliency signal.

Our proposal described in this chapter is inspired by this previous approach, since we understand that it resembles the mechanism that implements human attention, considering that it is a system that analyzes frequency content from different bands using a limited memory storage, and uses this information to detect novelty in the environment. In addition, it is a generic and unsupervised model that does not depend on the nature of the incoming signals, and can be used to detect onsets. However, an initial analysis we made suggested that the performance of Surprise could be improved by means of a logarithmic compression operator. This approach, that we named Bayesian Log-surprise, ultimately became the core of our own saliency technique: Echoic Log-surprise, that will be described in next chapters.

This chapter is devoted to explain the fundamentals of Bayesian Surprise: how it works, its underlying components, some of its disadvantages, etc.

We then study the solutions it produces for the particular task of onset detection. Moreover, we introduce our proposal called Bayesian Log-surprise and show several graphs describing how both approaches behave, and how they perform when they are used for the task of saliency detection.

5.1 Fundamentals of Bayesian Surprise

As [Itti and Baldi, 2009] described in their work, our survival depends directly on our ability to cope with surprising events, whatever their perceptual modalities are: acoustic, visual, somatosensory, etc. In fact, in order to show a noticeable degree of novelty an event must be unexpected, since expectation implies that there exists a previous knowledge about the appearance of the stimuli. Moreover, the context of the event plays an essential role, which can be better understood with the following example: a person walking on the street is more surprised when he/she perceives a car horn in the middle of a totally silent street at night that when he/she listens to it during rush hour in the middle of a big city. In both contexts, the candidate salient signal is the car horn. However, in the second scenario there are plenty of other interfering acoustic signals that mask it.

[Itti and Baldi, 2005] proposed a Bayesian approach to measure saliency, where they compared prior and posterior probabilities in order to measure the novelty of an event. The prior probability was defined as $P(M)$, where M would represent the background stimuli of a scene. For example, a crowded room, a lonely street, a mountainous scenario, etc. These are scenes where sensory information is abundant, considering that the incoming cues might be visual, acoustic or olfactory, among others. The aforementioned model M would be a summary representation of the scene as it is perceived by a particular subject.

On the other hand, the posterior probability was defined as $P(M|D)$, where the variable D is representing an incoming sensory signal. Consequently, the posterior probability would describe the context M once the data D is known. Needless to be said, both prior and posterior probabilities can be related by means of Bayes theorem:

$$P(M|D) = \frac{P(D|M)}{P(D)} P(M) \tag{5.1}$$

where $P(D|M)$ is the likelihood of a given stimulus D given the background M and $P(D)$ is the a priori probability of D. Consider that so far the information from a perceptual environment would be represented by both probabilities $P(M)$ and $P(M|D)$, which convey the available knowledge about the scene before and after incoming information is acquired, respectively. A potential way of measuring saliency would consist in determining how much the data D affects the variable M that represents the background scene, considering that for an anomalous event the difference observed among $P(M)$ and $P(M|D)$ should be bigger than for irrelevant data. According to [Itti and Baldi, 2005] this relevance could be measured by means of the Kullback-Leibler divergence as described below:

$$D_{KL}(P(M|D) \parallel P(M)) = \int_{\mathrm{M}} P(M|D) \log \frac{P(M|D)}{P(M)} dM, \tag{5.2}$$

which by definition is always non-negative:

$$D_{KL}(P(M|D) \parallel P(M)) \geq 0. \tag{5.3}$$

Additionally, it should be noted that the Kullback-Leibler divergence is not a distance, since it is not symmetric:

$$D_{KL}(P(M|D) \parallel P(M)) \neq D_{KL}(P(M) \parallel P(M|D)). \tag{5.4}$$

For the computation of Surprise from video, [Itti and Baldi, 2005] proposed to use their original multi-scale saliency map [Itti et al., 1998], that we described previously in Section 3.5.1, and compute Bayesian Surprise for each pixel from every feature map, considering a temporal Surprise that takes into account data for a single pixel individually along five different time scales, as well as a spatial Bayesian Surprise that manages information from the neigborhood of each pixel.

The concept of Surprise was later explored by [Schauerte and Stiefelhagen, 2013], who adapted Bayesian Surprise to detect saliency from acoustic cues. First, they chose an appropriate feature to represent auditory information, taking into account that these signals are typically monodimensional or might also be stereophonic. This contrasts with the visual features of the original Bayesian Surprise and many other visual saliency detection works, since they usually have two spatial dimensions and a third one representing time, for the particular case of video.

[Schauerte and Stiefelhagen, 2013] proposed the computation of the spectrogram for the acoustic signals. This approach is advantageous for two reasons: first, the spectrogram represents both spectral and temporal information. Considering that perceptual studies demonstrated that some frequency ranges are more relevant for human listeners (see Section 2.2), a spectro-temporal analysis could help to prioritize some bands depending on their perceptual relevance. Secondly, the spectrogram is a bidimensional signal similar to an image. The difference is that whereas an image has two spatial dimensions representing horizontal and vertical information, the spectrogram has one dimension representing frequency content and the other representing time. As a consequence, with proper modifications visual saliency algorithms could be adapted to work with acoustic saliency, as proposed by [Kayser et al., 2005; Tsuchida and Cottrell, 2012].

According to Equation 5.2, the first step to compute Bayesian surprise consists of obtaining the prior and posterior probabilities $P(M)$ and $P(M|D)$. [Schauerte and Stiefelhagen, 2013] proposed two distributions with Kullback-Leibler divergences in close-form: Gaussian and Gamma. We explore the Gaussian example in this Section, considering that Bayesian Surprise is computed independently for each one of the frequency bands k from the incoming spectrogram.

For a specific frequency bin k and frame n it can be computed as follows:

$$s_{surprise}(k,n) = \int P_{k,n}(\mathrm{x}) \log \frac{P_{k,n}(\mathrm{x})}{P_{k,n-1}(\mathrm{x})} d\mathrm{x} \tag{5.5}$$

where the notation used for the probabilities is simplified. Posterior and prior probabilities are represented as $P_{k,n}(\mathrm{x})$ and $P_{k,n-1}(\mathrm{x})$, respectively. This notation implies that the prior probability representing the background model has information from the previous frame, notated as $n-1$, whereas the posterior probability would represent the updated model considering the new incoming information from the frame n, both of them computed for a single frequency band k.

Since we propose to use a Gaussian model, the spectrogram $X(k,n)$ is used to determine the parameters to obtain the prior and posterior probabilities, which are necessary to compute Surprise according to the following Equation:

$$s_{surprise}(k,n) = \frac{1}{2}[\frac{(\mu_{k,n} - \mu_{k,n-1})^2}{\sigma^2_{k,n-1}} + 2\log\frac{\sigma_{k,n-1}}{\sigma_{k,n}} + \frac{\sigma^2_{k,n}}{\sigma^2_{k,n-1}} - 1], \tag{5.6}$$

where $\sigma^2_{k,n-1}$ and $\mu_{k,n-1}$ represent the variance and mean of the prior probability for the frequency k and the temporal frame $n-1$, and the same would happen with the posterior probability.

The mean and the variance for these Gaussian distributions are computed via Welford's online algorithm [Welford, 1962]. First, a circular buffer or sliding window B with N elements is defined. Then, for each temporal frame the mean is computed and updated as follows:

$$\mu_{k,n} = \mu_{k,n-1} + \frac{X(k,n) - \mu_{k,n-1}}{N}. \tag{5.7}$$

The algorithm would repeat the previous step for the n_{fft} temporal frames of the spectro-temporal representation $X(k,n)$. For the sake of simplicity we change the notation of the mean values, and instead of keeping their temporal indexes n and $n-1$, we consider the following expressions to be equivalent: $\mu_{k,prior} = \mu_{k,n-1}$ and $\mu_{k,post} = \mu_{k,n}$.

The computation of the variance uses the buffer more extensively and is more elaborated, as it is described in Algorithm 1. Notice that the pseudocode is not showing the optimized structure implemented by [Schauerte and

Stiefelhagen, 2013] in their original code (see [Schauerte, 2013]), but instead a more readable representation of the variance computation procedure. The whole process can be summarized into three steps: updating the buffer content, computing the updated variance value according to its equation, and finally updating the prior variables.

Algorithm 1 Online variance estimation using a circular buffer

```
$\sigma^2_{k,post} = 0$
for n do 1:$T_{signal}$
    $\mu^{'}_{k,prior} = X(k,n)$
    $A_0 = 0$
    Update buffer $B$
    for $n_B$ do 1:N
        $\mu^{'}_{k,post} = \mu^{'}_{k,prior} + \frac{B(n_B) - \mu^{'}_{k,prior}}{N}$
        $A_{n_B} = A_{n_B - 1} + (B(n_B) - \mu^{'}_{k,prior}) \cdot (B(n_B) - \mu^{'}_{k,post})$
        $\mu^{'}_{k,prior} = \mu^{'}_{k,post}$
    end for
    $\sigma^2_{k,prior} = \sigma^2_{k,post}$
    $\sigma^2_{k,post} = \frac{A_N}{N}$
end for
```

The auxiliary variable A_{n_B} helps to compute the final variance, which is obtained once per temporal frame and after using the N values stored in the buffer $B(n_B)$, where n_B is the index ranging $n_B \in \{1, \cdots, N\}$, $\mu^{'}_{k,prior}$ and $\mu^{'}_{k,post}$ represent auxiliary mean values and are only used to estimate the variance. Notice that they are different from the estimate $\mu_{k,prior}$ that was explained in Equation 5.7, since $\mu^{'}_{k,prior}$ updates using values from the buffer B and $\mu_{k,prior}$ updates after each temporal frame of $X(k,n)$.

It should be stated that T_{signal} represents the temporal duration of the spectro-temporal representation $X(k,n)$. Once that mean and variance values are computed for all the temporal frames, Kullback-Leibler divergence is used to obtain the Surprise level for the selected frequency band k thanks to Equation 5.6. Then, in order to get the final temporal saliency signal the Surprise signals for all the k frequencies are averaged:

$$S_{surprise}(n) = \frac{1}{n_{fft}} \sum_{k=1}^{n_{fft}} s_{surprise}(k,n). \tag{5.8}$$

Finally, a binary signal is obtained using a static threshold that is calculated as the average magnitude of the saliency signal.

A graphical example is depicted in Figure 5.1, where an incoming audio signal $x(m)$ is transformed into a spectrogram $X(k,n)$ with n_{fft} frequency

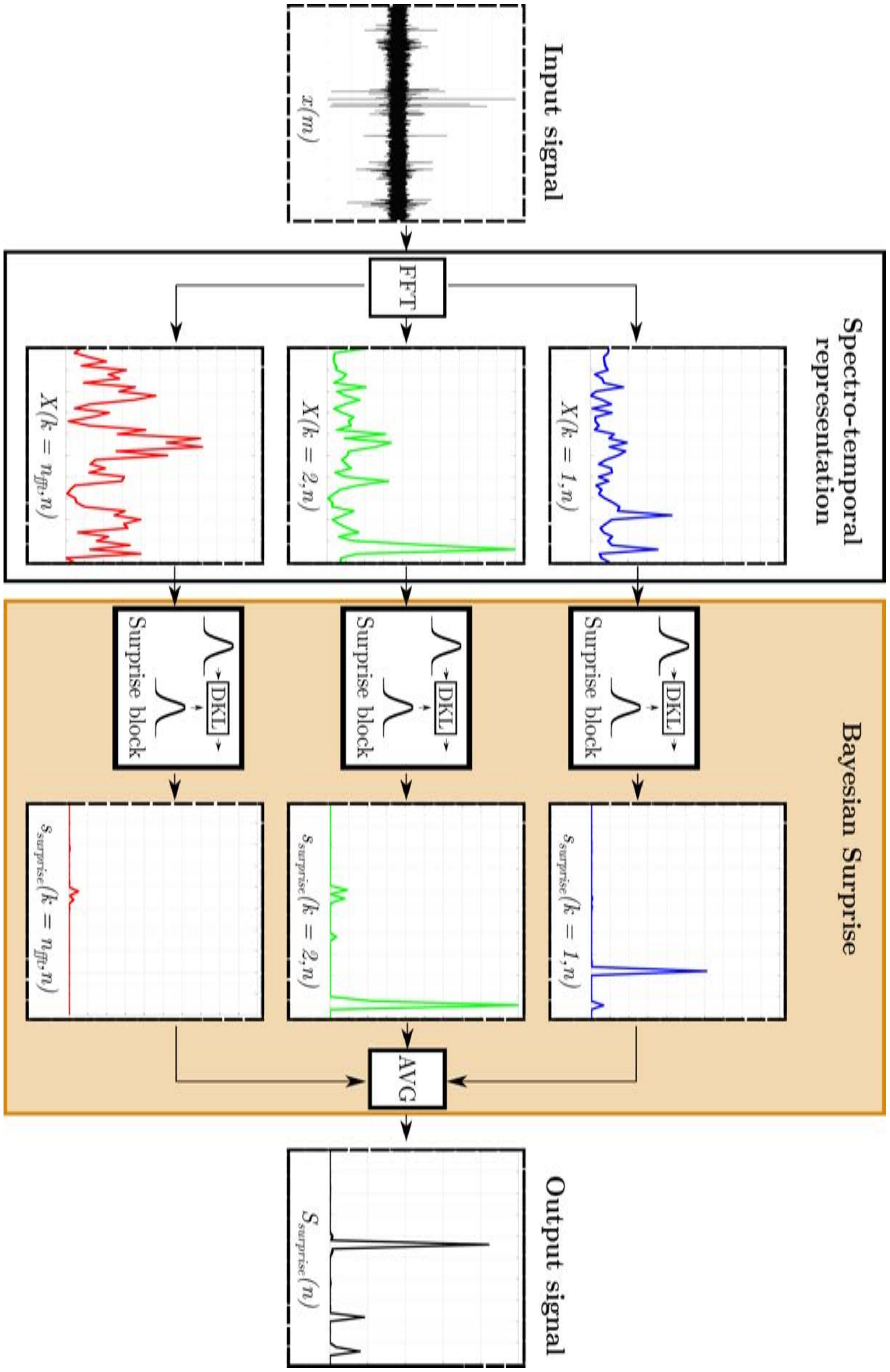


Figure 5.1: Schematic of Bayesian Surprise computation for audio signals, considering the Gaussian model.

bins. The example shows how Bayesian Surprise is obtained independently for three frequency components from the spectrogram, which are finally combined to produce a saliency signal $S_{surprise}(n)$. A more detailed explanation about the Surprise block is depicted in Figure 5.2, where two Gaussian distributions are fit from two consecutive frames of the spectrogram for a certain band, which are finally compared thanks to the Kullback-Leibler divergence.

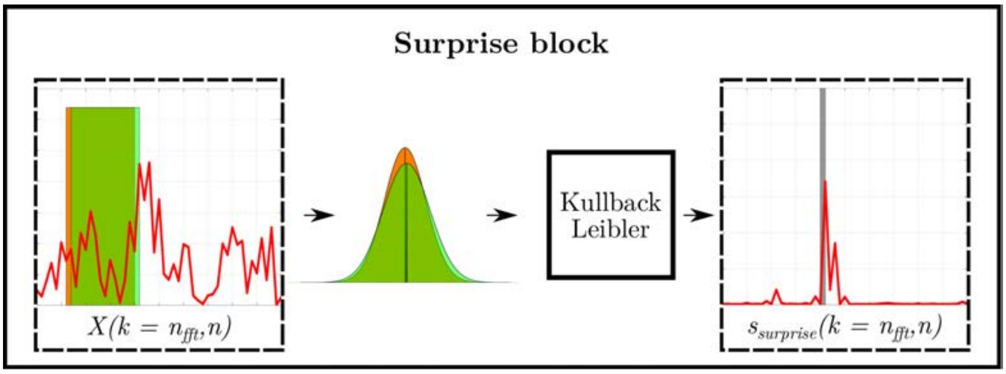


Figure 5.2: Surprise block representing an example of the computation of Bayesian Surprise for one of the frequency bins, in this case $k = n_{fft}$. Observe that it compares two consecutive temporal frames and computes the Kullback-Leibler divergence of their estimated Normal distributions. The result is the gray square depicted in the outcome signal $s_{surprise}(k = n_{fft}, n)$.

5.2 Bayesian Surprise for acoustic event detection

We explained in Section 5.1 the theoretical fundamentals behind Bayesian Surprise, paying special attention to the fact that the incoming audio signal is represented by a spectrogram or a cochleogram and each frequency band k is processed independently. Later we mentioned that in order to get a monodimensional temporal signal the average through all the frequency components k is performed, according to Equation 5.8.

However, before performing the aforementioned average a Surprise map can be depicted representing the saliency detected for each band, as illustrated in Figures 5.3 and 5.4. Both Figures show a thresholded saliency map on top and the input spectrogram on the bottom, which was computed using a frame length of 20 ms with an overlapping factor of 50%. The Surprise map on the top was obtained using $N = 50$, and it was thresholded by the average value of the map in order to ease visualization, since the activity produced by the algorithm is strongly conditioned by N. Both graphs are complemented by lateral subplots, which convey the magnified data for a temporal window of 1 second of duration, and should be useful to show a more detailed view of the corresponding Surprise map and the spectrogram.

For Figure 5.3 an example from MATLAB was used during the computations, with a sampling frequency of $F_s = 8192$ Hz and whose spectrogram shows important activity occurring during its whole duration. Thanks to the proposed threshold, the graph depicts that Bayesian Surprise seems to be able to detect activity at the beginning of some prominent peaks of the spectrogram depicted below.

On the other hand, Figure 5.4 shows a similar set of plots, but considering instead a file from one of our datasets (see *DCASE-T2* in Section 4.3.1). It contrasts with the Figure 5.3 in the sense that a finer frequency resolution can be used since $F_s = 22$ kHz. Additionally, acoustic activity seems to be less frequent than for the previous example, considering that the plot shows a wider timespan from 70 to 80 seconds and the spectra shows only subtle activity. In fact, a visual preview of the Bayesian Surprise map suggests that saliency is difficult to notice directly from the plot, and only thanks to the threshold we are able to depict such activity on the top of the Figure.

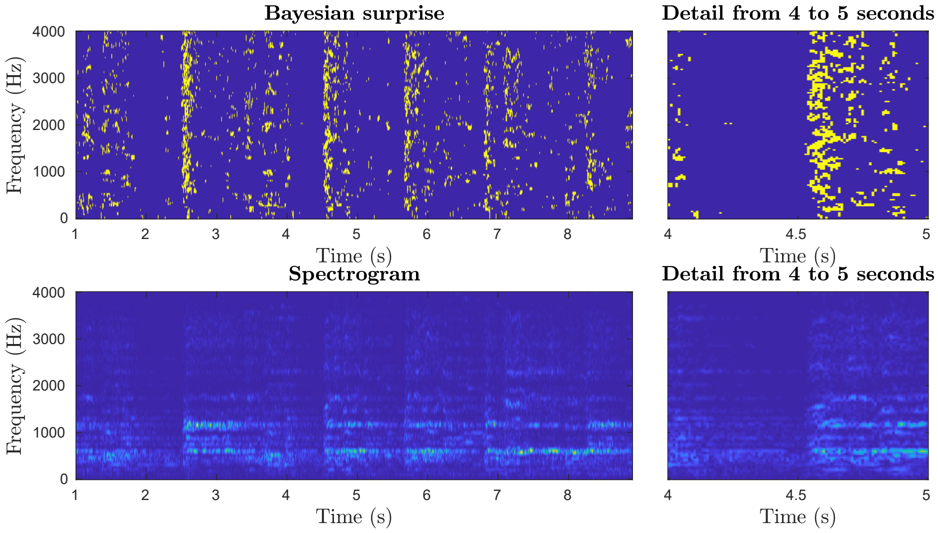


Figure 5.3: Spectrogram (bottom) and Bayesian Surprise map (top) for Handel's "Hallelujah Chorus" taken from MATLAB, and detailed graphs showing activity from 4th to 5th seconds. The Surprise map is represented as a binary image in order to ease its visualization.

Even if the acoustic events are barely observed from the Surprise map for every frequency, it does not necessarily mean that they are not detected. For instance a complementary analysis is depicted in Figure 5.5, which shows the results after computing the average through all the frequency bands according to Equation 5.8 for the same data file from *DCASE-T2*. Surprise curves were computed using three different frame lengths, namely 1 s, 100 ms and 20 ms, assuming that the smaller the frame length the finer the output temporal resolution, but also the noisier the outcoming saliency sig-

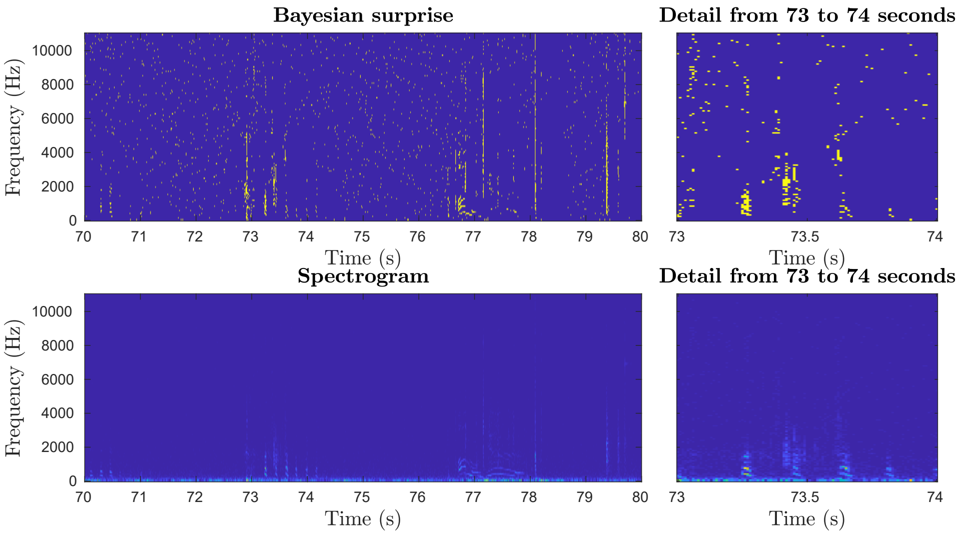


Figure 5.4: Spectrogram (bottom) and Bayesian Surprise map (top) using one of the audio files from the *DCASE-T2* dataset, and detailed graphs showing activity from 73rd to 74th seconds. Notice that the Bayesian Surprise signal is binary, since no activity was noticeable before applying an adequate threshold.

nal. We kept a constant memory value of $N = 50$ for the circular buffer for Bayesian Surprise. Figure 5.5 also shows the Energy of the original signal, computed directly from the spectrogram according to Parseval's theorem as in Equation 3.2.

Notice that both magnitudes, Energy and Surprise, are normalized for the sake of clarity to the range $[0, 1]$. A first insight suggests that using a frame length of 1 s should be discouraged, since for a buffer of $N = 50$ the first 50 frames are going to be computed using a circular buffer with some empty cells, which means that the first 25 s of Surprise are going to be poorly estimated. A solution for this approach consists of disregarding the initial N frames of the Surprise curve, so poor estimations are not represented. This problem is observed in the Surprise curve shown in Figure 5.5, whose value is nulled until around 25 s. For an audio sequence of short duration, a good amount of data available at its beginning would be wasted. In addition, this big frame length hinders the detection of quick changes that occur in fractions of a second, since the temporal resolution is so low that they cannot be observed in the spectrogram.

To palliate these problems the solution consists of reducing the frame length, as it is shown in the second and third plot of Figure 5.5. The first detail that can be observed is that both signals, Surprise and Energy, are sharper as a result of the increased temporal resolution. However, results also suggest that for Surprise the peaks with bigger magnitude tend to hinder the appearance of the smaller ones, to the extent that bigger ones grow huge, whereas the smaller ones become tiny in comparison. Two observable examples are the peaks with latencies of around 44 s and 60 s, both of them appearing in the three plots of Figure 5.5. It can be observed that their magnitude is huge in comparison with the rest of the detected peaks, so Bayesian Surprise would not have any issue detecting this activity no matter the frame length. On the contrary, peaks detected around 20 s and 30 s get a smaller magnitude along with the frame length, to the extent that the peak detected at 30 s seems to vanish when the frame length is 20 ms. This behavior suggests that the Bayesian Surprise outcome signals might suffer a compression problem that requires a solution.

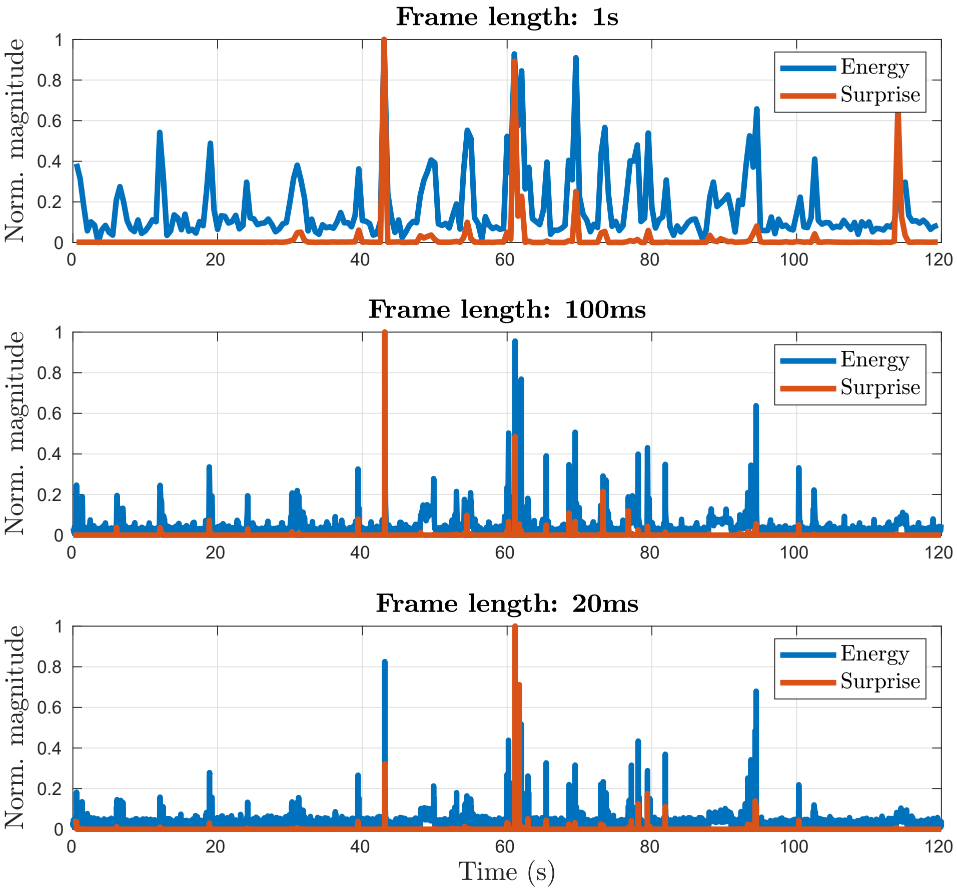


Figure 5.5: Energy and Bayesian Surprise curves obtained for an audio sample from *DCASE-T2* for several frame length values. Notice that the magnitudes of both curves are normalized to the same magnitude scale.

5.3 Improving detection: Bayesian Log-surprise

In the previous two Sections the concept of Bayesian Surprise was introduced, as well as how it can be implemented. In addition, the compression issue that occurs when using this technique has also been depicted, using three different frame lengths during the analysis.

The compression matter has to do with the difference of magnitude existing between small and big peaks, which makes the former to be masked by the latter. As a consequence, a potential solution consists in applying a technique that compresses signals into a certain dynamic range. This methodology is usually referred to as companding, and was proposed originally for communications and data transmission. In essence, this technique compresses the magnitude from an analog signal into a range of interest, so data fits properly the channel properties and can be transmitted more efficiently. In the receiver, the signal is expanded and then used depending on the nature of the original signal.

For this work it is interesting to understand how the compression stage works, since it might help to solve the issue mentioned above. The compression stage depends strongly on a compression curve, considering that there exist two widely known approaches [Proakis, 2001]: the μ-law, used in the United States of America, and the A-law, used in the rest of the world. Both techniques belong to the G.711 recommendation [ITU-T, 1988], and their compression curves are the ones depicted in Figure 5.6, which shows the similarities of both techniques.

The A-law is computed with the following function:

$$F_{A-law}(x) = \operatorname{sgn}(x) \begin{cases} \frac{A|x|}{1+\log(A)}, & \text{if } |x| < \frac{1}{A} \\ \frac{1+\log(A|x|)}{1+\log(A)}, & \text{if } \frac{1}{A} \leq |x| \leq 1, \end{cases} \tag{5.9}$$

where the parameter A controls the curvature, whereas for the μ-law the function is:

$$F_{\mu-law}(x) = \operatorname{sgn}(x) \frac{\log(1+\mu|x|)}{\log(1+\mu)} \quad \text{for } -1 \leq x \leq 1, \tag{5.10}$$

and the control parameter is μ. Both curves are used to compress a signal in the magnitude range of $[-1, 1]$.

The two functions use the logarithm to compress an input signal into a specific output range. During a preliminar experiment we decided to implement a similar compression stage using exclusively the logarithmic operator and a normalization system. We verified that our methodology produced similar curves to the ones of A-law and μ-law, so we decided to refine it and use it in our saliency system [Rodríguez-Hidalgo et al., 2016].

The first step of our proposal consists in applying the logarithmic operator directly over the Bayesian Surprise map. Then, the average through all

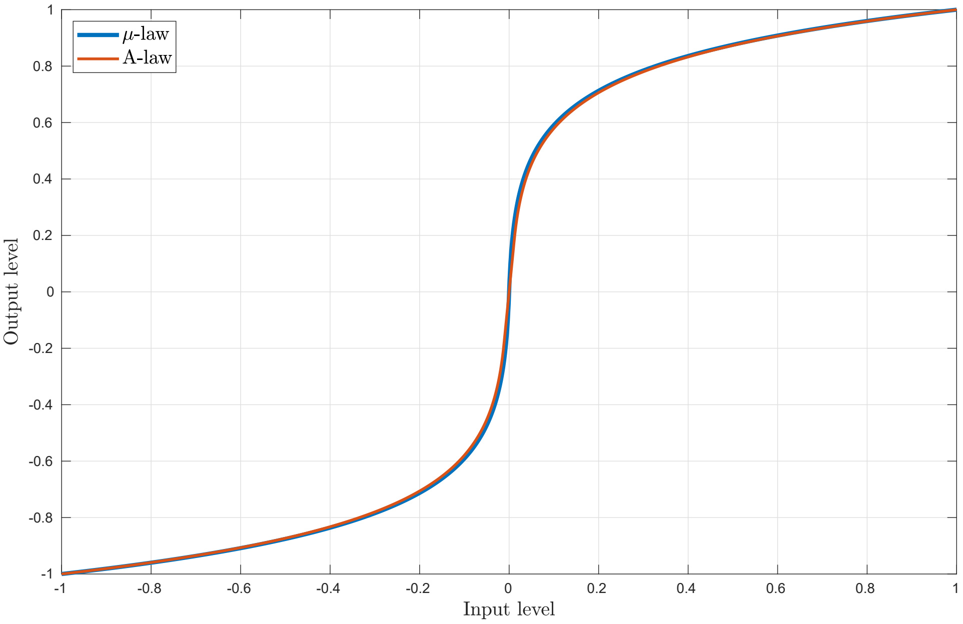


Figure 5.6: Curves described by the A-law and μ-law for the compression of signals according to Equations 5.9 and 5.10 using $A = 87.6$ and $\mu = 255$, respectively.

the frequency bands is obtained and normalized to $[0, 1]$:

$$s_{log-norm}(n) = \text{Norm} \sum_k \log s_{surprise}(k, n), \tag{5.11}$$

where $\text{Norm}(x)$ is the function that normalizes to the aforementioned range. The average or DC component $\bar{\mathbf{s}}_{DC}$ is subtracted from the previous equation and then the outcome signal is halfwave rectified using $H_w(x) = \frac{x+|x|}{2}$, which eliminates magnitudes that ended up below zero after the subtraction. Then, the signal is finally normalized:

$$s_{log-surprise}(n) = \text{Norm}\left[H_w\left(s_{log-norm}(n) - \bar{\mathbf{s}}_{DC}\right)\right]. \tag{5.12}$$

The reason to subtract $\bar{\mathbf{s}}_{DC}$ is related to the compression process. The logarithmic operator compresses both relevant information and noise, which implies that normalizing the outcome signal will introduce an average component that needs to be subtracted. This component is computed as follows:

$$\bar{\mathbf{s}}_{DC} = \frac{1}{T_{signal}} \sum_n s_{log-norm}(n). \tag{5.13}$$

Figure 5.7 depicts how the three compression techniques perform using once again an audio file from *DCASE*-*T2*, showing that they produce similar saliency signals. These techniques perform as expected, since they have a logarithmic behavior by definition, considering slight differences controlled by the parameters A and μ in the case of the companding techniques. As a consequence, we deduce that it is adequate to keep the logarithmic operator that we proposed in [Rodríguez-Hidalgo et al., 2016]. In that work we referred to this algorithm as **Bayesian Log-surprise**.

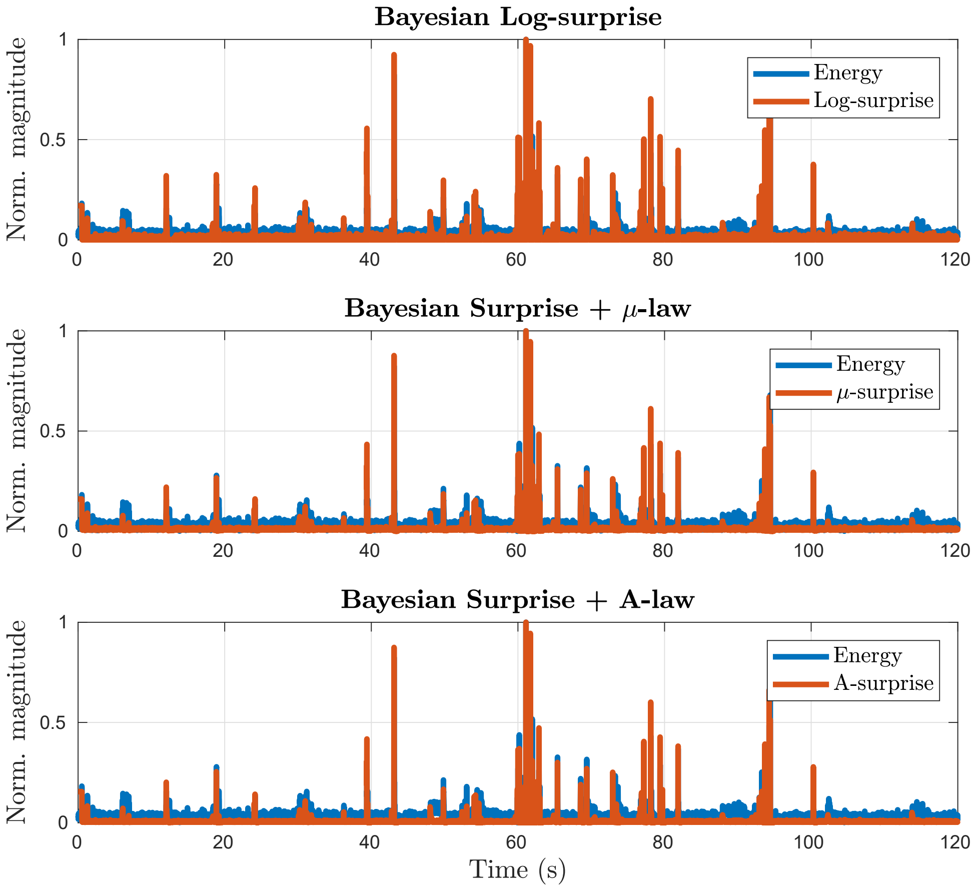


Figure 5.7: Comparison of the effect of three compression techniques for an audio example from DCASE-T2, considering a frame length of 20 ms and overlapping of 50% for the spectrogram.

Additionally, we noticed that the original proposal for Bayesian Surprise makes use of the spectrogram as its input image, which does not take into consideration how humans perceive sounds. That is, as it was introduced in Section 2.2 humans are more sensitive to changes and sounds concentrated in the lower frequencies of the spectra, an effect that is not so noticeable for the higher bands of the auditory spectrum. This issue was addressed by ASUN, as we explained in Section 3.6.3, where the authors used a gammatone filterbank to compute a cochleogram that they modified grouping consecutive frequency bands with certain degree of overlapping. In addition, they reduced the dimensionality of the resultant spectro-temporal representation using Principal Component Analysis to keep only the most relevant information.

Inspired by their work, we considered that Bayesian Log-surprise should include this prior knowledge about perception. Therefore, we implement a stage where these perceptual priors are included by means of a Mel-filterbank matrix, whose theoretical details were introduced in Sections 2.2 and 2.4.3. Hence, in the following Sections $X(k, n)$ will represent the Mel cochleogram where $k \in [1, n_{Mel}]$. The global scheme of Bayesian Log-surprise is depicted in Figure 5.8, where the example shown depicts each of the signals obtained stage by stage of the algorithm from the initial temporal signal $x(m)$ until the output Log-surprise signal $s_{Log-surprise}(n)$.

Similarly to the analysis that we performed in Section 5.2, Figure 5.9 depicts an example from DCASE-T2 considering Bayesian Log-surprise against energy for three different frame lengths. In contrast with the results depicted in Figure 5.5 it can be observed that peaks for the saliency signal have bigger magnitudes than previously for Surprise, which justifies the addition of a compression technique.

In addition, for these updated results we used the Mel cochleogram, in contrast with the spectrogram that was obtained to compute the curves from Figure 5.7. We observe that after using the cochleogram the curves become sharper, since the plot shows more peaks over the threshold of 0.5 than before. Considering that both of them use the logarithmic operator as their compression algorithm, the effect depicted in Figure 5.9 shows that introducing the Mel-filterbank in the Log-surprise scheme affects positively the magnitudes of the peaks, since it increases their prominence from the background noise.

Another aspect that should be taken into account is the length of the circular buffer used during the computation of Surprise and Log-surprise. During our previous analysis, the length of the buffer B was held static to the value $N = 50$, since our intention was to show the effect of the frame length over the effectiveness of Bayesian Surprise and Log-surprise. Regardless of the objective performance that the systems above might reach depending on the size of the buffer, a matter that will be analyzed in Section 5.4, Figure 5.10 depicts the curves obtained using the same example than previous Figures for Bayesian Log-surprise considering three different buffer sizes, $N \in \{16, 64, 512\}$. This plot shows a more detailed view of three acoustic events detected from 80 s to 100 s. In addition, the yellow background shows the ground truth label as provided in the DCASE-T2 dataset, representing the whole length of each acoustic event.

According to the graphs, one of the most notorious effects associated with the increase of N is the reduction of the background noise level that is observed when no acoustic event is active. This behavior is related to how Bayesian Surprise and Log-surprise work. Essentially, a Gaussian distribu-

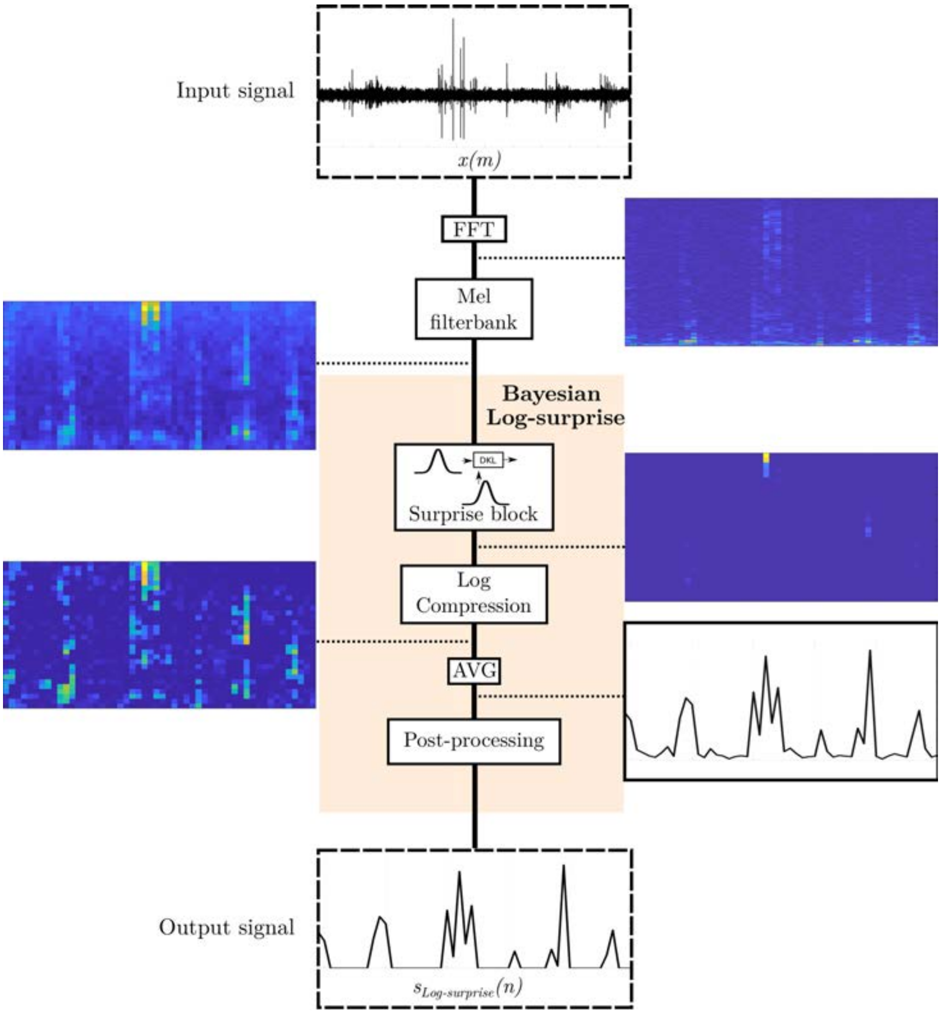


Figure 5.8: Global scheme of Bayesian log-surprise with examples of the outcomes of each of the stages. The AVG stage computes the average of the outcomes of each of the bands. The post-processing includes $\bar{\mathbf{s}}_{DC}$ subtraction, rectification and normalization.

tion is fit for each frequency band, and the longer the buffer the better the estimation of the distribution followed by the signal under consideration is. If the temporal area under analysis is mostly formed by background noise, its estimation will become more precise. If it is mostly formed by noise but there is a sudden transition to an event, the distribution should suffer a noticeable change both in its mean and variance. If the buffer is analyzing a long event with a transitory-like behavior, the event will become less and less salient when time passes depending on its stationarity. In view of

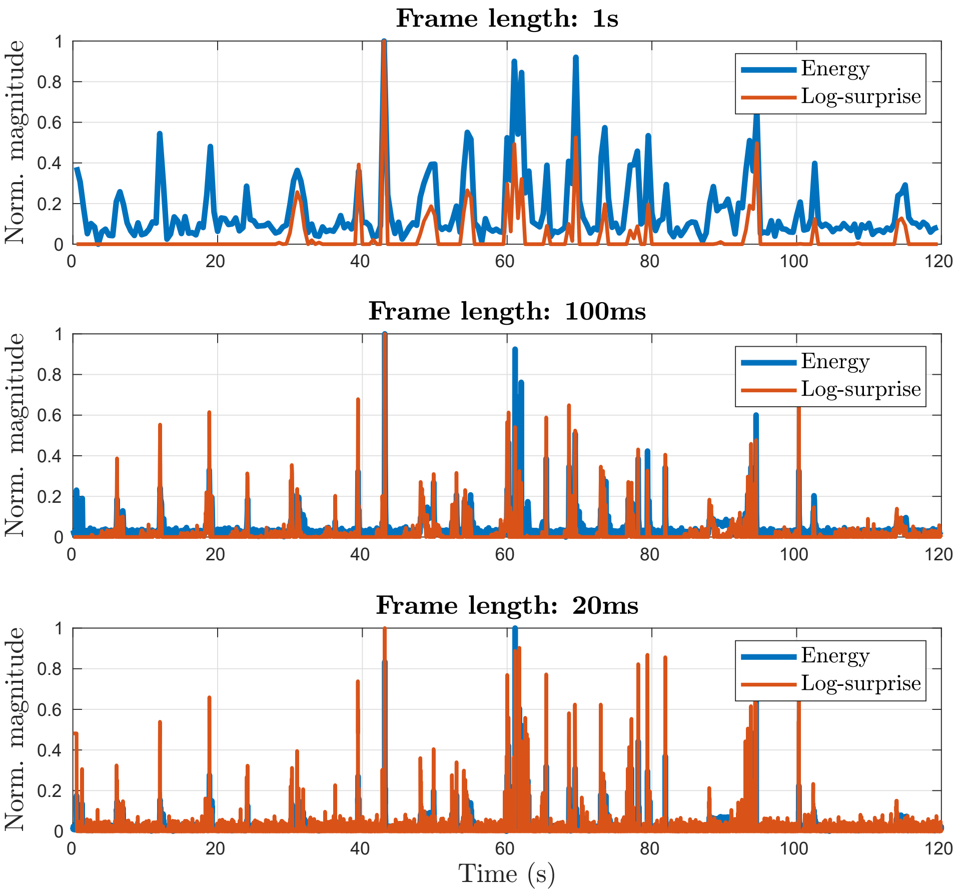


Figure 5.9: Analysis of Bayesian Log-surprise against the energy of the signal considering three different frame lengths for the cochleogram, computed for an example of DCASE-T2 with an overlapping factor of 50% and considering $n_{Mel} = 30$.

these observations, we decided to quantitatively evaluate the circular buffer length.

5.4 Results for Acoustic Event Detection

So far, this chapter has described the workflow of Bayesian Surprise, as well as some of its disadvantages. As a potential improvement we introduced a new model that we named Bayesian Log-surprise, which solves the compression problem by means of a logarithmic operator and boosts the magnitude of the event responses thanks to an alternative spectro-temporal representation, a Mel cochleogram. Nevertheless, it is relevant to assess the performance of both techniques considering not only the qualitative aspects obtained by visual inspection as we introduced in Sections 5.2 and 5.3, but

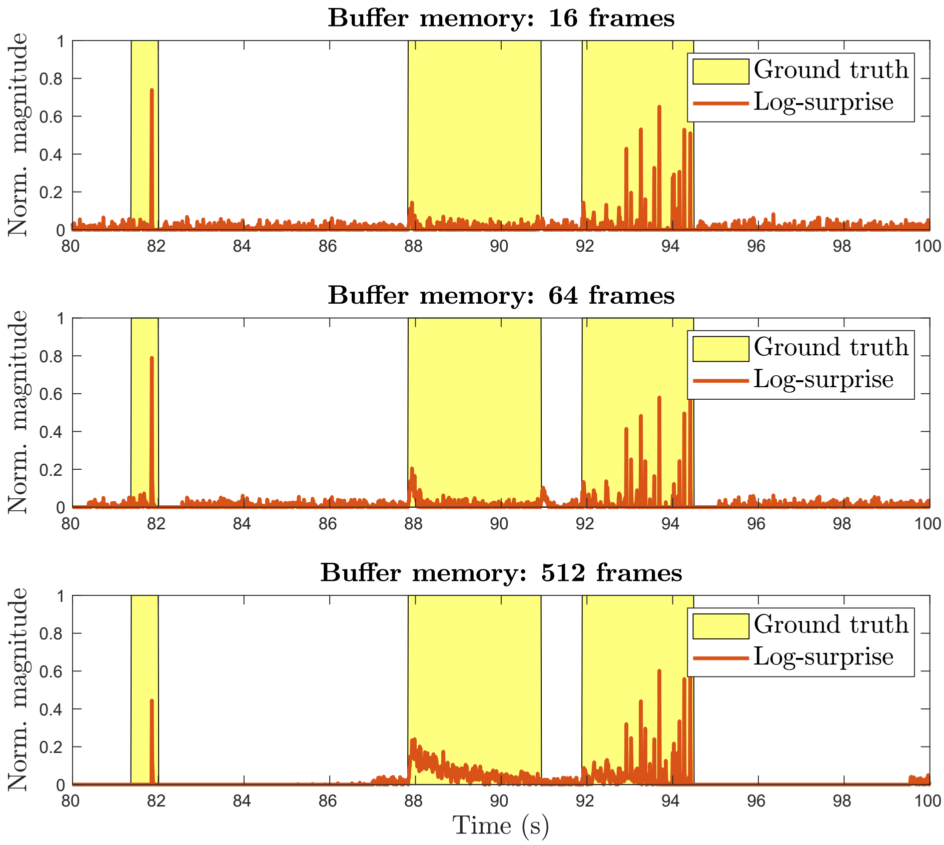


Figure 5.10: Analysis of the effect of the circular buffer length B for Bayesian Log-surprise, using a cochleogram computed with an overlapping factor of 50% and a frame length of 20 ms.

also carrying out an analysis that includes objective metrics.

As we mentioned in Section 4.1, in this work the performance of the saliency algorithms is measured according to the criterion proposed in DCASE-2016 for the task of AEC/D, where the F-score is computed considering that an acoustic event occurs if it is positively detected in a latency within a pre-established margin of error of 200ms. The F-score is computed for every audio file independently and is finally averaged to get a global F-score.

The first experiment designed for this Section evaluates the evolution of the performance of Bayesian Surprise and Log-surprise, since the latter is a modified version of the former. There are two different effects that need to be assessed in this analysis: first of all, under which circumstances do any of the two techniques perform better than the other. Secondly, since the performance of both techniques is strongly mediated by the size of the circular buffer B it is necessary to study how both evolve when this size changes. As a consequence, the parameter under analysis will be the length

N.

In this work, we define N to take values following $N = 2^p$ for $p \in \{1, ..., 10\}$, although bigger values of p might be considered depending on the length of the audio files. Thanks to this methodology we are able to measure the performance of both Surprise techniques considering a diverse range of memory values.

The results are depicted in Figure 5.11, where four bar diagrams show the F-scores computed for each of the datasets considered during the analysis. Note that these scores are obtained as the average from the individual F-score of each of the files from every dataset, whereas the last bar diagram includes the results after concatenating the scores of all the files from the three individual datasets. One of the first aspects that can be drawn from the graphs is that the Bayesian Log-surprise scores are bigger for most buffer sizes N in comparison with Bayesian Surprise, with the exception of $N = 128$, 256 and 512 in MIVIA where both techniques perform similarly. In addition, neither of the techniques seems to be capable of operating when the value of N is too small, which might suggest that memory needs to be sufficiently big to compute saliency and normally bigger than 8 frames. On the contrary, there is no clear conclusion on the value of N. In the case of Log-surprise we observe a common behavior for DCASE-T2 and UPC-TALP, where the maximum values of their average F-scores were obtained for $N = 64$ and $N = 128$, equivalent to 0.65 s and 1.29 s respectively, whereas for MIVIA there is a bigger variability in the results, which show that a maximum existed for $N = 16$ and $N = 256$, two really disparate values. In fact, for this dataset Log-surprise seems to start working with a reasonably good performance from $N = 8$, and then oscillates from such value until the biggest N considered. Since these variations were not too different for any value of N, it seems reasonable to suggest that Log-surprise works when its buffer has a length between $N = 64$ and $N = 128$.

On the contrary, for Bayesian Surprise the average F-scores are mostly smaller than the ones for Log-surprise, at least for both DCASE-T2 and UPC-TALP, being always below $F < 0.25$. The MIVIA dataset seems to be an exception, since for the proper values of N, Bayesian Surprise is capable of challenging and even slightly outperforming Log-surprise, as it occurs for $N = 512$. However, the global results leave no doubt about the performance of both systems: with an adequate value of N within the range that we introduced in the previous paragraph, Bayesian Log-surprise outperforms Bayesian Surprise.

Another conclusion that we can draw from this analysis is that, as expected, the performance of the system varies with the dataset. However, the buffer size proves to be relatively invariant for most of them for the case of Bayesian Log-surprise as it can be observed from the global results. Moreover, for the frame length that we are considering to compute the cochleograms it seems reasonable to use a buffer memory close to N

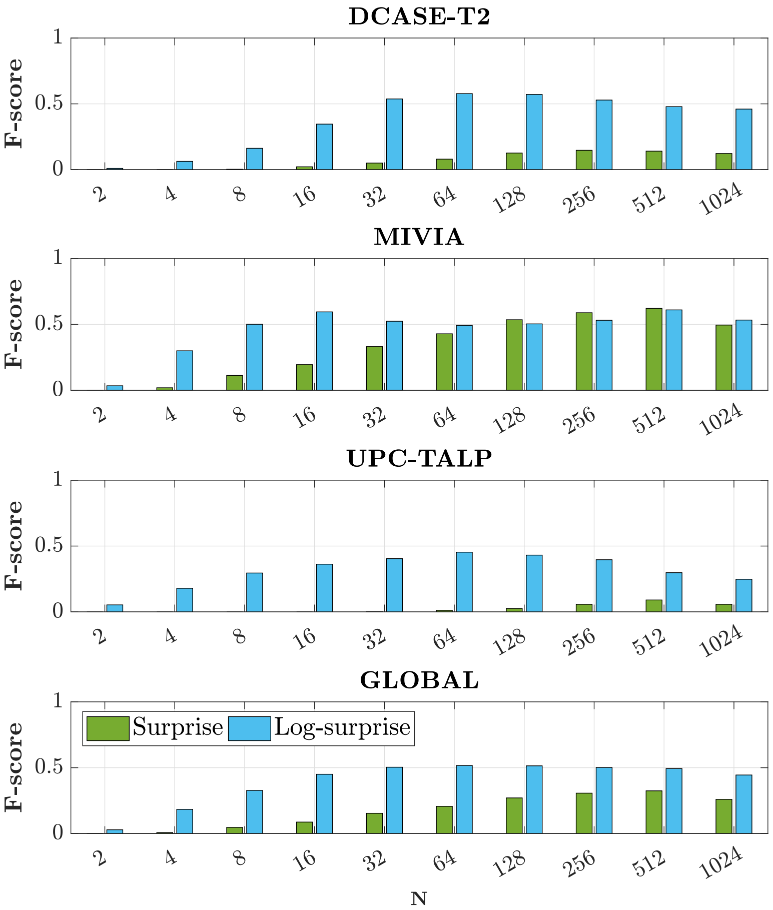


Figure 5.11: Average F-score obtained for Bayesian Surprise and Log-surprise using three different datasets. The variable under analysis is the size of the circular buffer used to compute both techniques.

$\in \{64, 128\}$. However, we consider that this parameter should be validated for other setups and datasets.

From the previous analysis we may deduce a potential configuration of N for both techniques that will be useful in later analysis of this thesis. The procedure followed to obtain such configuration consists in obtaining the maximum F-score for each file for each dataset, whose N values are grouped to form a histogram. Thanks to this methodology we are capable of determining which values of N are more advantageous for each dataset and technique, and how frequently they occur. With these histograms we obtain a set of weights that represent the most common values of N for each possible configuration of datasets and Surprise/Log-surprise techniques.

We use such information in the bar diagram shown in Figure 5.12, where we represent the F-scores for some classical detection techniques, such as

Energy and VAD, in addition to Kayser and Kalinli saliency algorithms. Results were obtained as the average F-score for the files of each dataset and technique. On the other hand, for Surprise and Log-surprise we use the weights that were obtained in the previous paragraph from the values of N, which represent the most common values of N that produced optimal results. We multiply these weights by the F-scores of Surprise and Log-surprise to obtain a weighted F-score for each file, instead of focusing on the results for a particular value of N. Then, the weighted F-scores of the files are globally averaged. The underlying idea is that systems working in real life scenarios are subject to uncertainty, and we consider that the best way of taking this factor into account is by obtaining a weighted score rather than focusing on the best score possible. In addition, the error bars for each configuration are depicted considering a confidence interval of 95%. These are computed using the standard error of the mean, which is directly related with the dispersion of the files whose scores were grouped.

In a first instance, it can be observed that Kayser is only capable to detect events from DCASE-T2. On the contrary, Kalinli detection capabilities are noticeable for every dataset with respect to Kayser, even if the F-scores that both produce are lower in comparison to some of the other techniques under analysis. What seems interesting is that both techniques are similar, with the exception of some of the features (respectively, pitch and orientations to model ripples) and an iterative normalization algorithm, which were added by Kalinli et al.

Energy and VAD perform similarly for DCASE-T2 and UPC-TALP reflected in the fact that their confidence intervals overlap. Nevertheless, there is a surprisingly high F-score for Energy when it is used to process the MIVIA dataset, which actually produces the top score for such data even in comparison to Log-surprise. In fact, the F-scores of Energy and Surprise for MIVIA seem to be abnormally big in comparison with the results for the rest of the techniques and datasets, which might be related to the anomalous silent gap distribution of such dataset that we analyzed previously in Section 4.3.5. With respect to Surprise, what seems interesting is that it performs poorly both for DCASE-T2 and UPC-TALP, but it produces a competitive average F-score for MIVIA, although with a bigger than usual confidence interval in comparison with any other of the techniques of this analysis.

The case of Log-surprise is remarkable in the sense that the average F-scores it produces for DCASE-T2 and UPC-TALP are bigger than in the rest of the techniques of this study with non-overlapping confidence intervals. However, it seems to perform similarly to Surprise for the MIVIA dataset and its performance is clearly below that of Energy for that particular dataset even if the latter actually has a bigger confidence interval.

In addition to the F-score, we consider that evaluating both Precision (P) and Recall (R) scores independently might help to understand how

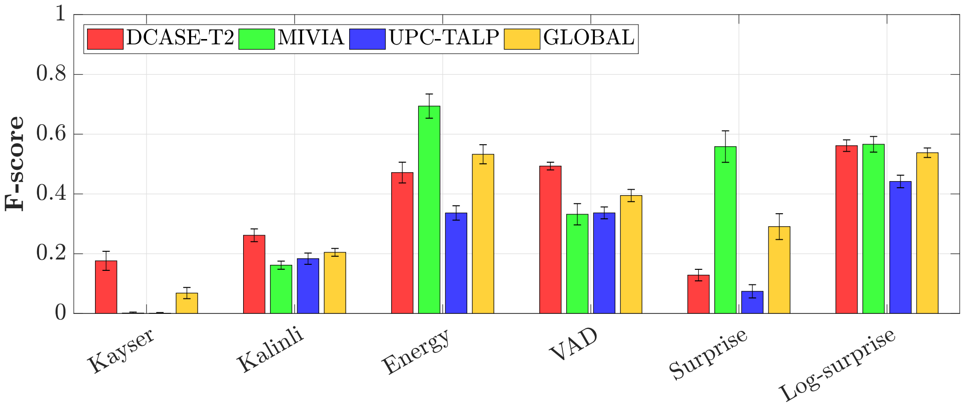


Figure 5.12: Comparison of the average F-score of each classical detection technique vs Log-surprise, considering the three datasets and the result of grouping the results of all their files.

the systems behave, considering that the scores reflect the influence of false positives and false negatives (see Section 4.1 for more details about P and R). A relevant remark about the variation of P and R is that depending on the evaluation criteria, one of the two results can be considered to be better. If a system is required to produce a decent number of properly detected events but a really small number of false detections, then a high P is required. On the contrary, if there is some laxity and the system is allowed to detect as many points as possible with the condition that some of them are correct then a big R and a small P is recommended.

Figure 5.13 depicts a scatter plot where the weighted Precision-Recall pairs for each file are represented, considering independently each of the detection techniques and datasets. These weighted P and R scores are computed similarly to the aforementioned weighted F-scores, where the weights for each value of N are used instead of a single value of N. In addition, the diagonal line representing $P = R$ is plotted for reference purposes. A remarkable fact is that for every dataset, Surprise produces values mostly below such diagonal line, implying that most of the time it occurs that $P > R$, meaning that it produces less false positive values than false negatives. The rest of the techniques are positioned in the other side of such boundary most of the time.

Kayser produces results close to the diagonal line for the three datasets, although its P and R are always small. In fact, for MIVIA and UPC-TALP most of these points are positioned at the null corner.

For DCASE-T2, Kalinli produces results with a similar P to Kayser, although its R values are higher and mostly concentrated in $R \in [0.3, 0.6]$. VAD and Energy have most of their scores concentrated in the same region of P and R, although Energy seems to produce worse results with a smaller

P score. Bayesian Surprise produces a cloud of points where P goes from 0.2 to 1 whereas $R \in [0, 0.2]$, which implies that the number of false negative values that it produces is extremely high. The case of Log-surprise seems to be mostly concentrated close the diagonal, with a slight trend towards $P > R$. Its position in the scatter plot justifies why in the 5.12 graph it shows the highest F-score.

The results for MIVIA are different, and the scores for Kalinli and VAD seem to concentrate in some straight lines where there is a big variation of R whereas $P \in [0, 0.2]$ for most of the points for Kalinli, and $P \in [0.1, 0.3]$ for VAD. This means that there is a big variation in the number of false negatives for each file, although the number of false positives remains high. Respecting to Energy, in contrast with the results for the previous dataset, it seems that for MIVIA its P increases for all the files, whereas its R stays in a similar range of values. For Surprise the P values remain similar, whilst its R values disperse into a cloud of points where most of them concentrate around $R \in [0.25, 0.8]$. This means that for this technique the number of false positives remains in a similar level, although the number of false negatives decreases dramatically for most of the audio files. For Log-surprise the average P decreases whereas its R increases, implying that the number of false negatives is smaller than for DCASE-T2, although the number of detected false positives arises.

For UPC-TALP Kalinli and VAD had a similar P score to the one observed for MIVIA. However, VAD concentrated in an area with higher R scores whereas Kalinli did the opposite. For Energy the majority of points get concentrated in $P \in [0.15, 0.4]$ and $R \in [0.3, 0.7]$, which means that there is a significant increase in the number of both false positive and negative points, being this effect more noticeable for the latter. In the case of Surprise the files show a big number of false negative points. Finally, for UPC-TALP Log-surprise detection produces a small number of false negatives, although the number of false positives is remarkably high with $P \in [0.2, 0.4]$.

5.5 Conclusions

In this Section the concept of Bayesian Surprise has been introduced, an acoustic saliency detection algorithm that processes each frequency band independently in order to measure the prominence of occurring events considering the Kullback-Leibler divergence. We have studied one of its weaknesses: the lack of compression in the output signal that it generates, which produces peaks with extremely different magnitudes that hinder the detection of the ones with smaller amplitudes.

We have introduced Bayesian Log-surprise as a potential improvement, which thanks to a compression stage based on the logarithmic operator produces output signals whose peaks amplitudes are more evenly distributed,

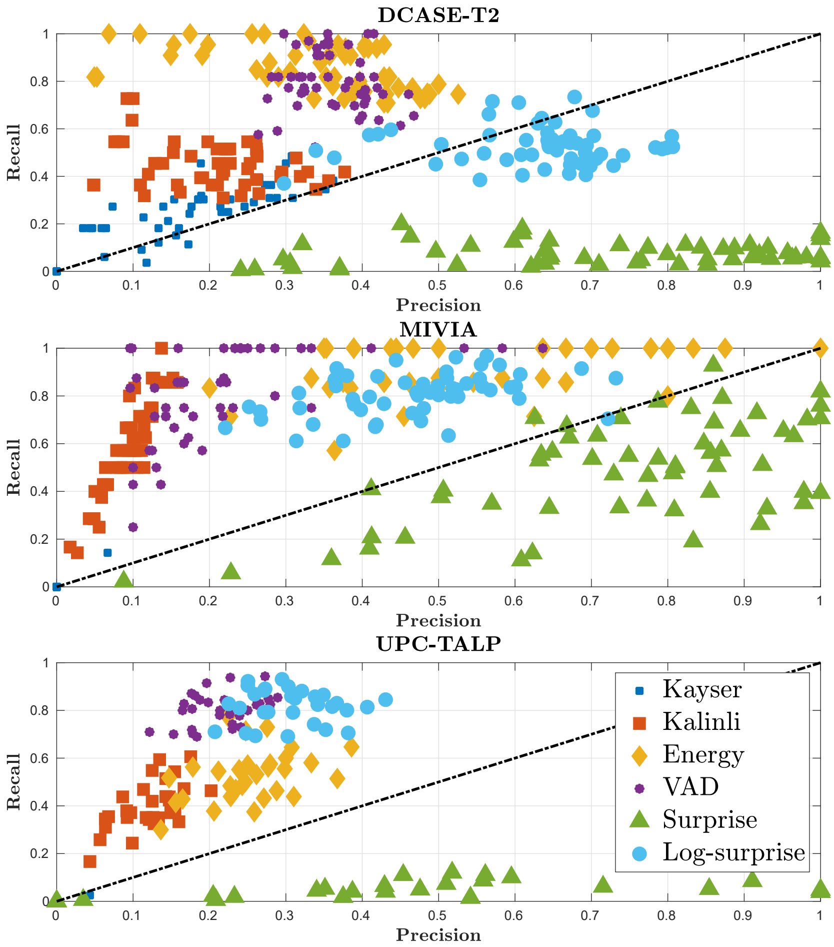


Figure 5.13: Weighted Precision-Recall points for every detection algorithm and dataset, obtained after representing the scores for each one of the points that conform each dataset. The black color line represents the points where $P = R$.

easing their detection. In addition, a filterbank formed by Mel triangular filters is included into the Bayesian Log-surprise scheme, which introduces some perceptual concepts related to how the HAS works.

The performance of Bayesian Surprise and Log-surprise, in addition to some other classical detection techniques, has been assessed using both qualitative and quantitative analyses. The qualitative analysis allows to understand two different concepts: first, how Surprise and Log-surprise are affected by the frame length used to obtain the spectro-temporal representation. A larger frame length allows to produce more appealing curves, whereas finer but also noisier results are obtained for smaller temporal windows. Secondly, a visual analysis of the buffer size suggests that increasing its length helps to reduce the effect of background noise. In addition, this first analysis hints that Log-surprise might perform better than regular Surprise, although this theory needs to be assessed using quantitative analysis.

The quantitative analysis has been performed considering a fixed frame length of 20 ms, which should provide a precise temporal resolution of the acoustic events. The three datasets described in Section 4.3 were used during these analyses all originally developed for the task of AEC/D. In a first instance, the F-score has been computed from the onsets considering several buffer lengths exclusively for Surprise and Log-surprise. Experiments suggest that the optimal results are obtained around $N = 64$ and $N = 128$. Moreover, the scores favors Bayesian Log-surprise over Surprise most of the times, which hints that the logarithmic compression is an effective mechanism to cope with the compression issue detected for Bayesian Surprise. On the contrary, the results for Surprise shows that it only performs properly for the MIVIA dataset.

Another objective analysis has been performed once that a set of candidate values for N has been obtained for Surprise and Log-surprise, and we have compared them against some detection techniques. We have observed that Kayser and Kalinli saliency algorithms are mostly outperformed by the rest of the techniques. Particularly surprising is the contribution of Energy, a simple algorithm that computes the energy directly from the cochleogram and is capable to outperform the rest of the techniques for one of the datasets. However, the global best results are obtained by Log-surprise.

We have also introduced a Precision-Recall analysis to understand how each one of the techniques behaves, and we concluded that there are severe variations depending on the dataset used for the analysis. As a consequence, it seems that further progress needs to be done so results become more uniform along datasets. In addition, a robustness analysis needs to be performed to make sure that the results are still useful under noisy conditions. Moreover, we conclude that Bayesian Log-surprise seems suitable for the development of our Echoic Log-surprise algorithm, as will be explained en Chapter 6.

Chapter 6

Echoic Log-surprise

Sensory memory is defined as the ability of the human brain to retain information temporarily after a stimulus is perceived until it is forgotten or learned and stored into another sort of memory. The particular case of Auditory Sensory Memory (ASM), also commonly known as echoic memory, will be of our concern in this chapter, although there is certainty that similar stores of information also exist for vision and haptic information. Visual sensory memory, usually referred to as iconic memory, has a capacity close to 1 s and one of the tasks where it is detected is blindness to change, which was introduced in Section 3.2. In the particular case of touch, it is named haptic memory and is capable of storing sensory data for as long as 2 s, approximately.

The particularity of echoic memory is that in comparison with the previous examples it lasts for a longer period of time. Several researchers tried to determine its capacity, which depending on the author and the experiments carried out was estimated to last from milliseconds to 10 or 20 seconds, with even shorter durations [Cowan, 1984; Bottcher-Gandor and Ullsperger, 1992; Sams et al., 1993; Gomes et al., 1999; Glass et al., 2008]. According to [Cowan, 1984], there exist two types of ASM: short storage, lasting from 150 to 350 ms, and long storage, from 2 to 20 s. They would differ in some aspects, such that the short one would be populated by weighted spectral information where relevance is determined by recency, whereas the long storage would be capable of storing longer sequences of continuous data. In addition, the duration of this memory seems to be related with the maturing level of the brain, and some authors discovered that younger people under a certain age seemed to lack this storage, whereas older children and adults were capable of storing auditory information for longer periods of time [Gomes et al., 1999; Glass et al., 2008].

Consequently, when there is a flow of auditory information it is first stored by means of echoic memory, whose content would be processed by bottom-up attention depending on its physical features without taking into consideration its semantics. This would occur afterwards, when information had been selected and finally processed by higher-level neural mechanisms, as it would happen with top-down attention. Inspired by this short duration storage, we propose a saliency algorithm that manages information from different and concurrent memory levels, taking into account data lasting from the range of 10 ms to 10 s. The core of our implementation is Bayesian Log-surprise as described in Chapter 5, where we observed that its most influencing factor in terms of performance is the circular buffer length, which represents the memory of the system. Therefore, it should be validated in different datasets. In this chapter we propose a multi-scale alternative that solves the memory issue by combining data from several memory durations, yielding the technique that we named **Echoic Log-surprise**.

6.1 Structure of Echoic Log-surprise

We proposed Bayesian Log-surprise as an alternative to Bayesian Surprise, which dealt with the compression problem that was previously described in Section 5.2. In fact, some of the analyses that we performed suggested that it was able to outperform Bayesian Surprise with the proper selection of its control parameter: the circular buffer size N, a term that we will denote as memory onwards. However, as we verified using Precision-Recall scatter plots there exists a dependence between the dataset under analysis and the capability of the system to detect events, affecting directly the number of false positives and/or false negatives.

In addition, from a subjective analysis, it was suggested that a longer circular buffer helped to reduce the effect of background noise, following the line of thought that the longer the buffer the smoother the Log-surprise saliency signal shall be. However, our analysis of the F-score showed that the optimal value was not the biggest possible as it could have been expected. In fact, an excessively big buffer might include several events at once, degrading the detection of new appearances. Furthermore, the optimal values of N proved to be $N = \{64, 128\}$ for two of the datasets, and variable for the other one.

These results inspired this line of work, and our goal consists in designing an adequate acoustic saliency technique that is able to overcome the disadvantages observed for Bayesian Log-surprise. We established that such technique should be able to combine the detection capabilities of Bayesian Log-surprise considering different memory values at once, and if possible, also reduce the effects of common noise data between Log-surprise signals at different scales. In [Rodríguez-Hidalgo et al., 2018a] we proposed a technique that fused the data available from several Log-surprise signals obtained with different buffer lengths, which we denoted Echoic Log-surprise. The stages that conform it are depicted in Figures 6.1, 6.2 and 6.3.

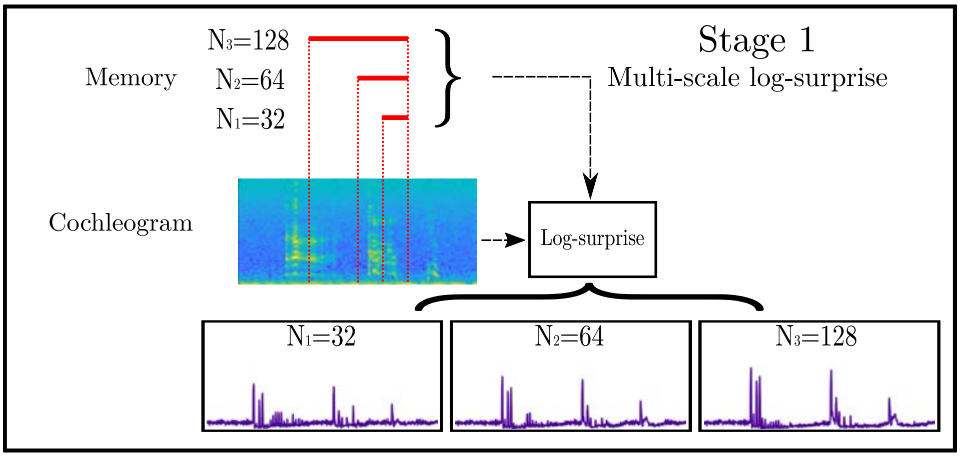


Figure 6.1: Graphical representation of the first stage of Echoic Log-surprise, where several Log-surprise signals are computed considering different memory values from the cochleogram of the input signal.

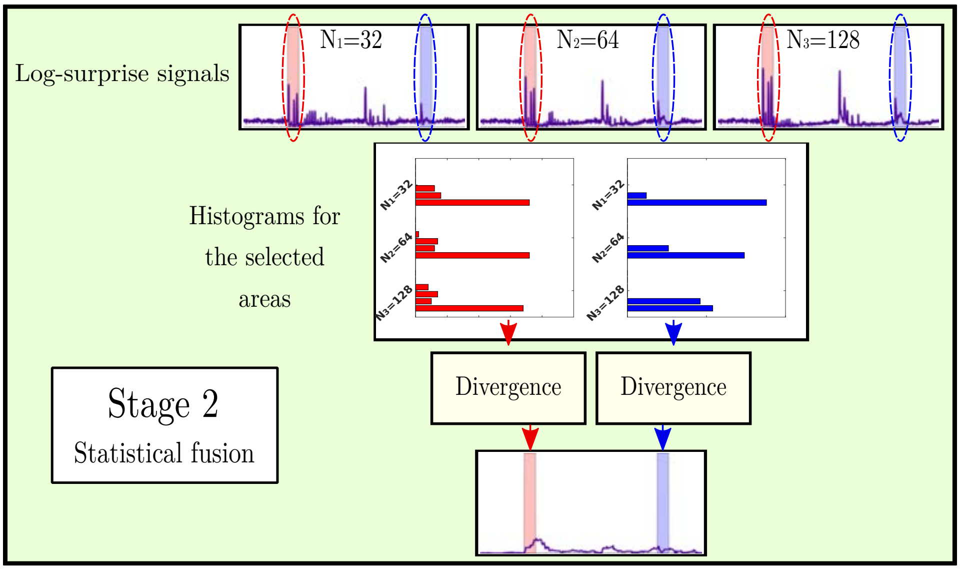


Figure 6.2: Graphical representation of the second stage where the statistical fusion takes place showing three Log-surprise signals obtained previously in the first stage. In the example, two areas represented in blue and red are selected to be fused, considering first the histogram for each one of the Log-surprise scales, whose data are finally fused using a statistical divergence.

The first stage, shown in Figure 6.1, takes as input the classical cochleogram of the audio signal obtained through the combination of the spectrogram with a Mel-scaled filterbank, as it was described in Section 2.4.3. From such representation of the signal it computes several Bayesian Log-surprise cues at different temporal scales, a task that is performed considering a set of memory values for the calculation of Log-Surprise. The number of temporal scales considered, and therefore, the number of computed Bayesian Log-surprise signals with different buffer lengths is called *depth* of the analysis and denoted as dth. These are finer or rougher depending on the memory established during the computation of Log-surprise. We consider that each Log-surprise signal is obtained with an exponential growth according to:

$$N_z = 2^{z-1} \cdot N_1, \tag{6.1}$$

where z is a reference to the depth of the system within the range $z \in \{1, \cdots, dth\}$, and indicates how many Log-surprise blocks are used to compute saliency. N_1 represents the length of the buffer for the first Bayesian Log-surprise block, and it is used to obtain the memory of successive blocks depending on dth.

The second stage, depicted in Figure 6.2, focuses on the fusion of the information available from the multi-scale data. This task can be considered the most relevant, since by means of statistical divergences the system is able to determine what information is actually more prominent.

Finally, as depicted in Figure 6.3 after the fusion process the outcome signal is thresholded in order to obtain a binary signal indicating the peaks that are salient according to the algorithm.

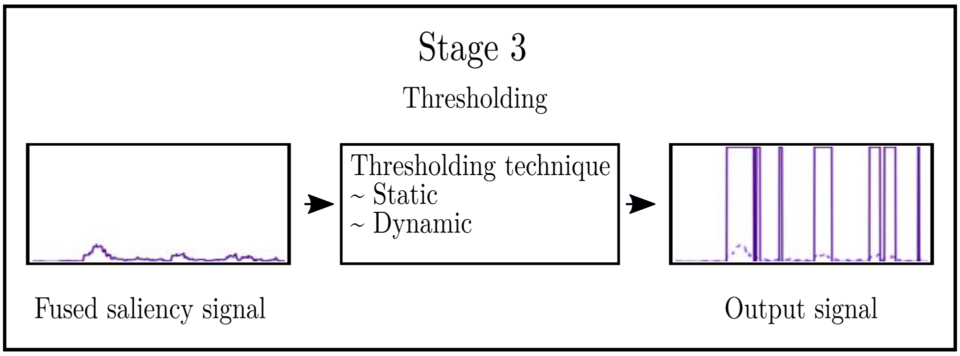


Figure 6.3: Representation of the third stage of Echoic Log-surprise, where the output signal from the second stage is thresholded according to the selected thresholding mechanism, which might be static or dynamic.

If we think about the similarities of Echoic Log-surprise with echoic memory and saliency, the multi-scale stage would be our equivalence to this memory register, since it is in charge of storing the statistical distribution of information acquired in different temporal instants and durations. From the data available at these registers anomalies are detected by means of statistical divergences, which provides the output signal that will be thresholded afterwards.

A visual example showing how an Echoic Log-surprise signal looks like is shown in Figure 6.4. Bayesian Log-surprise is depicted in blue, and is obtained using the memory values $N = N_1$ indicated in the subplot titles, whereas Echoic Log-surprise uses $dth = 3$, and consequently $z \in \{1, 2, 3\}$ considering that memory N_z for each scale is computed according to Equation 6.1. This means that each subplot for Echoic Log-surprise was obtained using its own three Log-surprise signals with different values of N_z that depended on their own N_1. Similarly to what was observed in Figures 5.7 and 5.9, the signals for both techniques become smoother when N_1 increases. In addition, one of the differences that can be observed for $N_1 = 64$ and $N_1 = 512$ is that Echoic Log-surprise signals become null when they are in a silence gap. In addition, with bigger values of N_1 the magnitudes of most of the peaks decrease significantly.

6.2 Fusion strategies and statistical divergences

All the fusion techniques used in this work always combine dth temporal signals, each one representing a different Log-surprise saliency signal (see Section 5.3 for more details), by means of the computation of the statistical divergence between their corresponding probability distributions. However, by definition some of the divergences are designed to work exclusively with two different distributions at once, whereas our system handles dth signals with $dth \geq 1$. As a consequence, two different strategies are applied for fusion.

The first strategy that we considered is pairwise fusion or as we notate, *local fusion*, where the statistical divergences only admit two distributions. In this alternative, the statistical divergences are computed in a pairwise manner and finally added to obtain the final fused signal. In this case, Echoic Log-surprise would be defined as follows:

$$s^L_{echoic}(n) = \sum_{z=1}^{dth-1} d_{local}(h^L_{d^z_{log-surp}}(n), h^L_{d^{z+1}_{log-surp}}(n)), \tag{6.2}$$

where $h^L_{d^z_{log-surp}}(n)$ and $h^L_{d^{z+1}_{log-surp}}(n)$ represent the probability distribution estimates to be fused, and they are obtained from the frame n an the previous L frames from each signal. To simplify the notation we will just write $h^L_z(n)$ and $h^L_{z+1}(n)$. A clearer vision of this fusion approach is depicted in Figure 6.5, which shows explicitly how the different divergences are computed and finally added together to conform the outcome saliency signal for

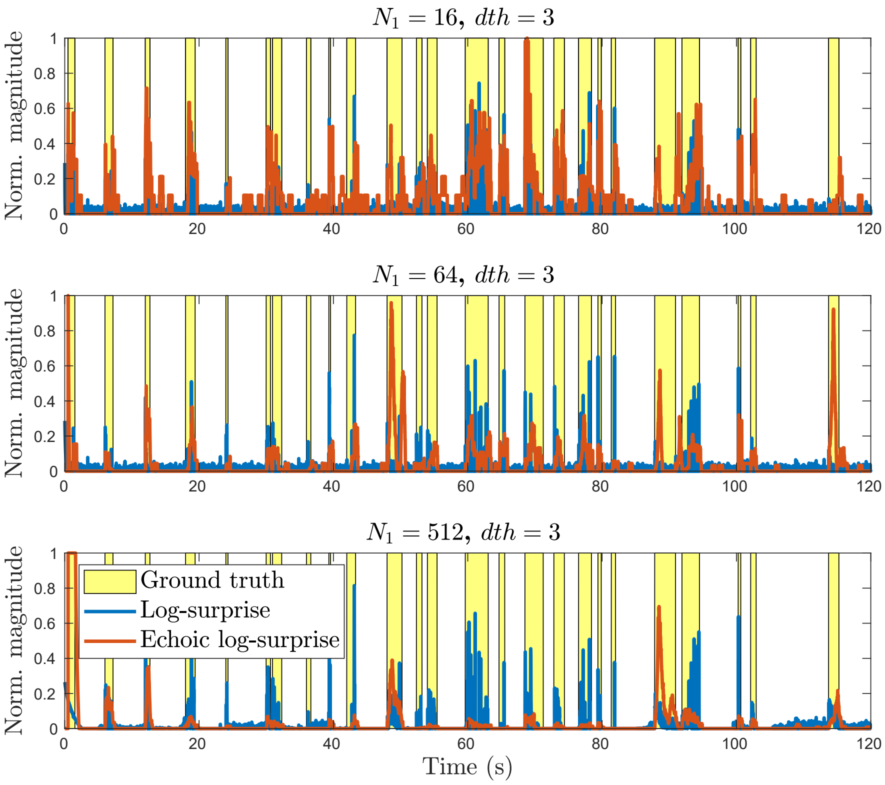


Figure 6.4: Analysis of Echoic Log-surprise (red) in comparison with Bayesian Log-surprise (blue) and ground truth AED labels (yellow), considering $dth = 3$ and three different initial buffer lengths. The plots were obtained using a signal that belongs to the DCASE-T2 dataset for which the cochleogram was computed with an overlapping factor of 50% and considering $n_{Mel} = 30$ and $F_s = 22$ kHz.

a frame n. Note that the pairwise computations are obtained for consecutive Log-surprise signals considering increasing memory values, a criteria that we selected so differences are extracted from relatively similar signals in terms of memory. These differences that are measured by the divergences and that, as a consequence, fuse the available information are the detected salient events.

Using the simplified notation explained above, local fusion becomes:

$$s^L_{echoic}(n) = \sum_{z=1}^{dth-1} d_{local}(h^L_z(n), h^L_{z+1}(n)). \tag{6.3}$$

The second fusion strategy, *global fusion*, measures the differences among a theoretically unlimited number of statistical distributions. In our proposal, this means that this kind of divergences are by definition able to fuse the

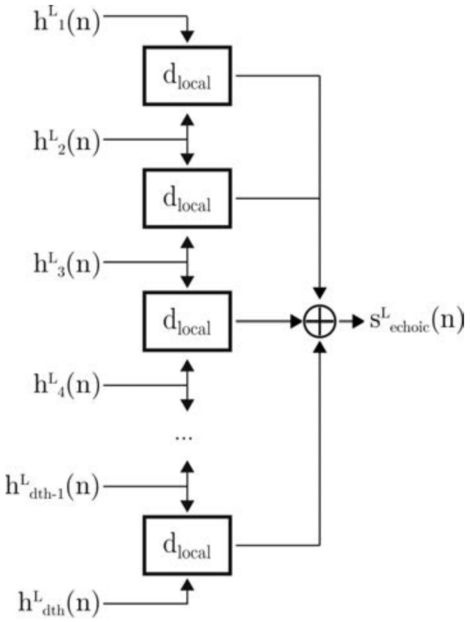


Figure 6.5: Schematic of the local fusion strategy, where the histograms of consecutive Log-surprise signals are combined pairwise by means of their divergences, which are finally added to get the output saliency signal.

dth estimated probabilities from *dth* Log-surprise signals. An example is depicted in Figure 6.6 and in the following Equation:

$$s^L_{echoic}(n) = d_{global}(h^L_1(n), h^L_2(n), \cdots, h^L_{dth-1}(n), h^L_{dth}(n)). \tag{6.4}$$

As observed, in both fusion strategies the computation of the probability distributions of each temporal sequence is required. For a temporal frame n and considering a temporal window with a duration of L frames, the underlying dth distributions are estimated by means of the histograms, which are denoted as $h^L_z(n)$ where $z \in \{1, \cdots, dth\}$. We simplify the notation of these histograms to h_z, considering that they must be recomputed for each temporal frame n and they have the same window length. In addition, we refer to the i-th bin of a histogram as $h_{z,i}$ when necessary.

In this work we have considered the following statistical divergences, some of which were analyzed in [Rodríguez-Hidalgo et al., 2018a]:

Cramer distance: It is obtained as the Euclidean distance from two distributions [Székely, 2002], which in our case are estimated by means

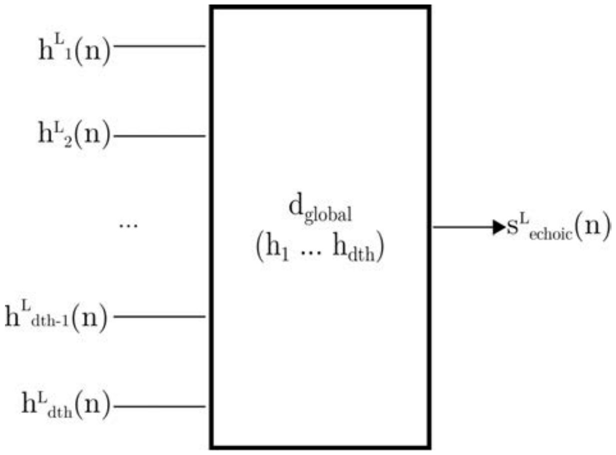


Figure 6.6: Schematic of the global fusion strategy, which combines several Log-surprise signals concurrently to obtain the output saliency signal.

of the histograms h_z and h_{z+1}.

$$d_{Cramer}(h_z, h_{z+1}) = \sum_{i=1}^{N_{bins}} (h_{z,i} - h_{z+1,i})^2. \tag{6.5}$$

where N_{bins} represents the number of histogram bins. Note that this distance is computed exclusively for two distributions, and belongs to the set of the so-called local fusion techniques.

Renyi-INF divergence: This divergence belongs to the set of Rényi divergences [van Erven and Harremoës, 2014], which can be controlled by a parameter α that defines some well-known divergences, such as the Kullback-Leibler one obtained when $\alpha = 1$. In the experiments presented in this work we tested two different divergences obtained independently for the values $\alpha = 2$ and $\alpha = \infty$. However since both produced similar results, from now on we will focus exclusively on the case where $\alpha = \infty$, which for simplicity we denoted Renyi-INF. For two histograms h_z and h_{z+1}, their Renyi-INF divergence is computed as the supremum of the ratios obtained for the bins i of each histogram, as defined below:

$$d_{Renyi-INF}(h_z \parallel h_{z+1}) = \log \sup_i \frac{h_{z,i}}{h_{z+1,i}}. \tag{6.6}$$

It represents another example of the local fusion strategy.

Bhattacharyya distance (Bhatta): This methodology was proposed by

[Bhattacharyya, 1943] and is defined as follows for two histograms:

$$d_{Bhatta}(h_z, h_{z+1}) = -\ln \sum_{i=1}^{N_{bins}} \sqrt{h_{z,i} \cdot h_{z+1,i}}. \tag{6.7}$$

Note that the previous equation shows that this divergence is an example of the local fusion strategy described before.

Jensen-Shannon divergence (JSD): This particular divergence is by definition capable of computing the differences between more than two distributions at the same time [Endres and Schindelin, 2003]. It is described by the following expression, where it can be observed that it is obtained as the entropy of a mixture distribution, obtained directly from the available *dth* variables, minus the entropy of each of the distributions that are meant to be fused:

$$d_{JSD}(h_1, \cdots, h_{dth}) = H(\sum_{z=1}^{dth} \pi_z h_z) - \sum_{z=1}^{dth} \pi_z H(h_z). \tag{6.8}$$

In the equation above, $H(x)$ refers to the Shannon entropy whereas π_z represents the weight of each of the histograms, that we consider to be uniformly distributed, so $\pi_z = \frac{1}{dth}$. This is one of the examples of what we called global fusion, since it is able to manage several divergences concurrently.

6.3 Thresholding

Finally, the Echoic Log-surprise signal obtained in the previous stage is introduced into the third one, depicted in Figure 6.3, which is in charge of the thresholding process.

Using an adequate threshold is critical when the algorithm is specifically designed for the task of detection. In our previous work [Rodríguez-Hidalgo et al., 2018a] we computed a threshold value for the output Echoic Log-surprise signals considering their average magnitude values. We denote this threshold as th_{ave}, and it can be computed according to:

$$th_{ave} = \frac{1}{T_{signal}} \sum_{n=1}^{T_{signal}} s(n), \tag{6.9}$$

where $s(n)$ represents the saliency signal from which events are going to be detected and T_{signal} represents the whole duration of the signal in frames.

Besides this static thresholding algorithm, we propose a dynamic methodology in Section 7.2 as an improvement for Echoic Log-surprise.

6.4 Optimal parameter setup

As we introduced in Section 6.1, the Echoic Log-surprise methodology relies on the selection of two parameters that control the initial memory and the depth of the system, denoted as N_1 and dth respectively. However, since different divergences are used to fuse information we have no a priori intuition about the points (N_1, dth) where optimal results could be obtained. Our first experiment consists in determining the optimal operation area for each of the datasets, as it is illustrated in Figure 6.7. For each of the fusion divergences, the point with optimal F-score is computed as well as its 9 nearest neighbors providing what could be understood as the optimal area of work for each of the techniques. Finally, all of these areas are added up to obtain the global operation areas depicted in Figure 6.7. This can be conceived as a rough estimation of the weight of each point (N_1, dth), where every statistical divergence contributes by stating what points produced its best scores.

For DCASE-T2 it can be observed that the optimal points are located at $N_1 = 8$ and $N_1 = 16$, considering depth going from 5 to 7, whereas for MIVIA and UPC-TALP the areas are different and situated around $N_1 = 64$ and $N_1 = 128$, with smaller depths dth from 2 to 4 for the former and from 3 to 5 for the latter. Note that we consider neither $N_1 > 1024$ nor $N_z > 1024$, since some of the signals available in our datasets are not sufficiently long to compute their saliency with such parameter values.

Consequently, it seems that the Echoic Log-surprise algorithm shows some similarities for the areas of MIVIA and UPC-TALP. These two differ from the area for DCASE-T2, where the algorithm seems to prefer deeper models with smaller N_1 values.

6.5 Performance comparison

In Section 5.4 we evaluated the performance of Bayesian Surprise and Log-surprise using the most common memory values as weighing values to obtain a weighted F-score for each one of them. Moreover, in Section 6.4 we performed a set of analyses to understand how the parameters N_1 and dth from Echoic Log-surprise evolved, as an attempt to determine if there were any common areas shared by the three datasets that could be used as a representative global setup. However, we observed that the regions that were obtained differed from one dataset to another.

The areas depicted in Figure 6.7 were obtained after counting how many statistical divergences agreed that a certain point (N_1, dth) was adequate for them to perform properly. Consequently, we decide to use these regions to compute the weight for each point (N_1, dth), considering that the darker the color the bigger shall be the weight it represents. Then, they are used to

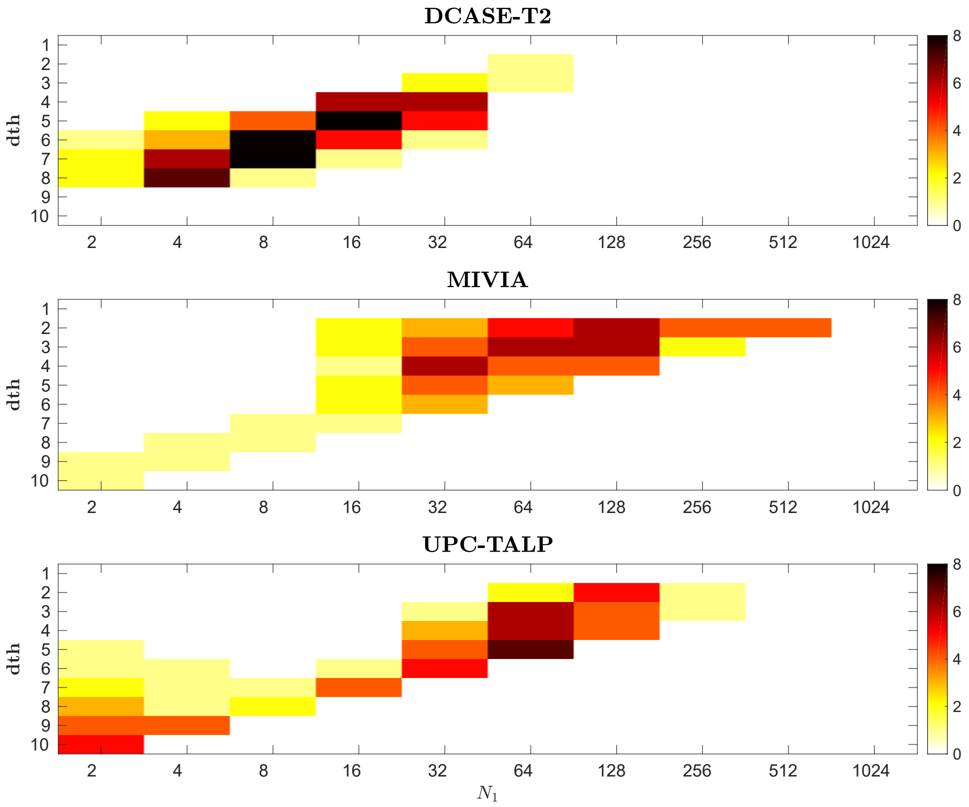


Figure 6.7: Area of work for a selection of fusion techniques for Echoic Log-surprise considering the static threshold for each dataset, computed by counting the number of techniques that performed properly in each one of the points (N_1, dth).

compute the weighted F-score for each audio file, whose values were finally represented in Figure 6.8. Similarly to how we explained in Section 5.4 for Bayesian Surprise and Log-surprise, we intend to take the uncertainty of the system into account thanks to the weighted F-score. Hence, for a certain technique instead of using only the single best point (N_1^{BEST}, dth^{BEST}) to obtain the global F-score, we work with the F-scores and their corresponding weights obtained for every point (N_1, dth) to obtain our weighted F-score. Figure 6.8 also includes the techniques that were previously analyzed in Section 5.4, since we want to verify if Echoic Log-surprise performs properly for the task of saliency detection and, moreover, whether it could be considered a real contribution to the state-of-the-art. Results are also summarized in Table 6.1.

The first thing that can be noted is that irrespective of the fusion technique, Echoic Log-surprise always produces the highest weighted F-score

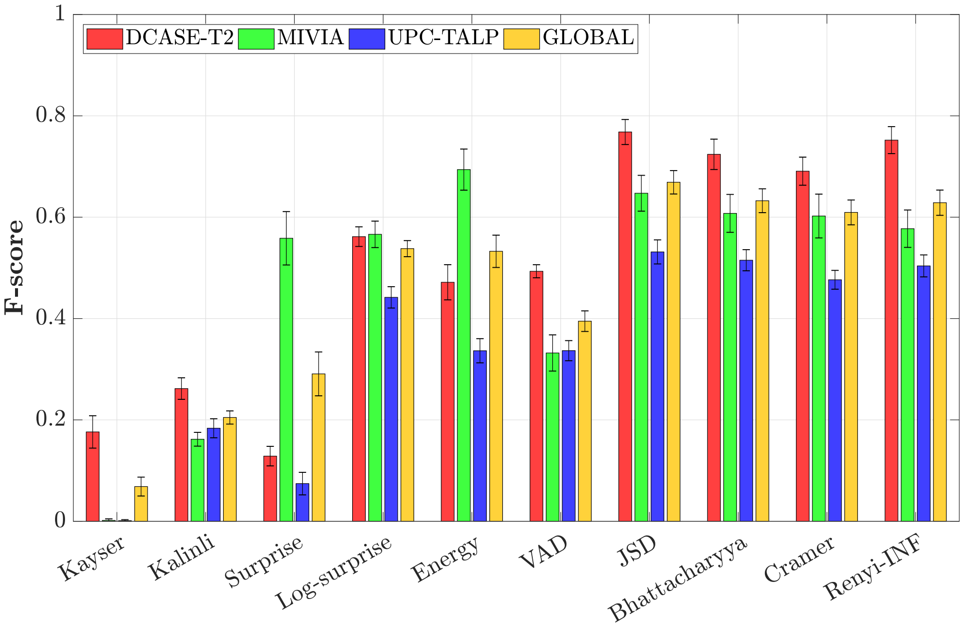


Figure 6.8: Comparison of the F-scores for some classical detection techniques against Echoic Log-surprise with four different statistical divergences, computed using a static threshold.

	DCASE-T2	**MIVIA**	**UPC-TALP**	**GLOBAL**
Kayser	0.176 ± 0.032	0.002 ± 0.003	0.001 ± 0.002	0.068 ± 0.019
Kalinli	0.262 ± 0.021	0.162 ± 0.014	0.184 ± 0.019	0.205 ± 0.013
Surprise	0.129 ± 0.019	0.558 ± 0.053	0.074 ± 0.022	0.291 ± 0.043
Log-surprise	0.562 ± 0.019	0.566 ± 0.026	0.442 ± 0.021	0.538 ± 0.016
Energy	0.472 ± 0.035	0.694 ± 0.041	0.336 ± 0.024	0.533 ± 0.032
VAD	0.493 ± 0.013	0.332 ± 0.036	0.337 ± 0.020	0.395 ± 0.020
JSD	0.768 ± 0.025	0.647 ± 0.035	0.531 ± 0.024	0.669 ± 0.023
Bhattacharyya	0.724 ± 0.030	0.608 ± 0.037	0.515 ± 0.021	0.632 ± 0.024
Cramer	0.691 ± 0.028	0.602 ± 0.043	0.477 ± 0.019	0.609 ± 0.025
Renyi-INF	0.752 ± 0.027	0.577 ± 0.037	0.504 ± 0.022	0.629 ± 0.025

Table 6.1: Table that includes the F-scores and their confidence intervals for Echoic Log-surprise and other classical detection techniques, computed considering static thresholding.

value, to the extent that the four possible statistical divergences used on this analysis offer a similar performance for the three datasets. The top performer is JSD, whose confidence interval overlaps with that of Bhattacharyya. However, this homogeneity in the results suggests that the selection of the fusion technique is secondary, since it is the fusion strategy itself what mostly matters.

Comparing the F-scores of JSD against the classical techniques shows that it clearly outperforms Bayesian Log-surprise for every dataset, and the same occurs with almost any other of the detection techniques. The only exception seems to occur with Energy, which indeed performed poorly for DCASE-T2 and UPC-TALP, but its performance for MIVIA is on average still better than any other alternative. This does not occur with the global results, where Echoic Log-surprise remains the top detection technique. Additionally, the confidence intervals from JSD and Energy are overlapping. This might suggest that they perform similarly for this dataset, which in comparison with the other two datasets is anomalous since it is conformed only by two event classes and its silent gaps are periodic, as we mentioned in Section 4.3.5. As a consequence, results suggest that the best detection results are obtained for Echoic Log-surprise, and particularly considering the JSD fusion technique.

Finally, there is a need to understand how the Echoic Log-surprise mechanism affects the process of detection in terms of false positive and negative points. We designed Figure 6.9 for such endeavor, which illustrates the Precision-Recall scatter plot for Energy and Log-surprise, as well as two representative examples of Echoic Log-surprise fusion algorithms, JSD and Renyi-INF. The comments made previously for Energy and Log-surprise remain the same: Energy has an excessively high number of false positives for DCASE-T2 and it has many points that additionally have a relatively high number of false negatives for UPC-TALP. For MIVIA, its P and R were similar to the ones obtained for Echoic Log-surprise.

On the other hand, if the performances of Log-surprise and the two examples of Echoic Log-surprise are compared it can be observed that the points for the latter have slightly higher P and R scores for MIVIA. Moreover, their R values for UPC-TALP are similar although there is an increase in the P score for Echoic Log-surprise, more significant for the particular case of JSD. Finally, the most noticeable improvement is observed for DCASE-T2, where P and R scores for JSD and Renyi-INF are concentrated around the diagonal representing $P = R$, and it occurs that both values are higher than the ones obtained for Bayesian Log-surprise.

Some conclusions that can be drawn from this analysis are that, as the Precision-Recall plots suggest, Echoic Log-surprise acts as a direct improvement of Log-surprise, considering that the scores for the three datasets are closely positioned for both techniques, although significantly better for DCASE-T2 and UPC-TALP. The Precision-Recall plots do not allow to

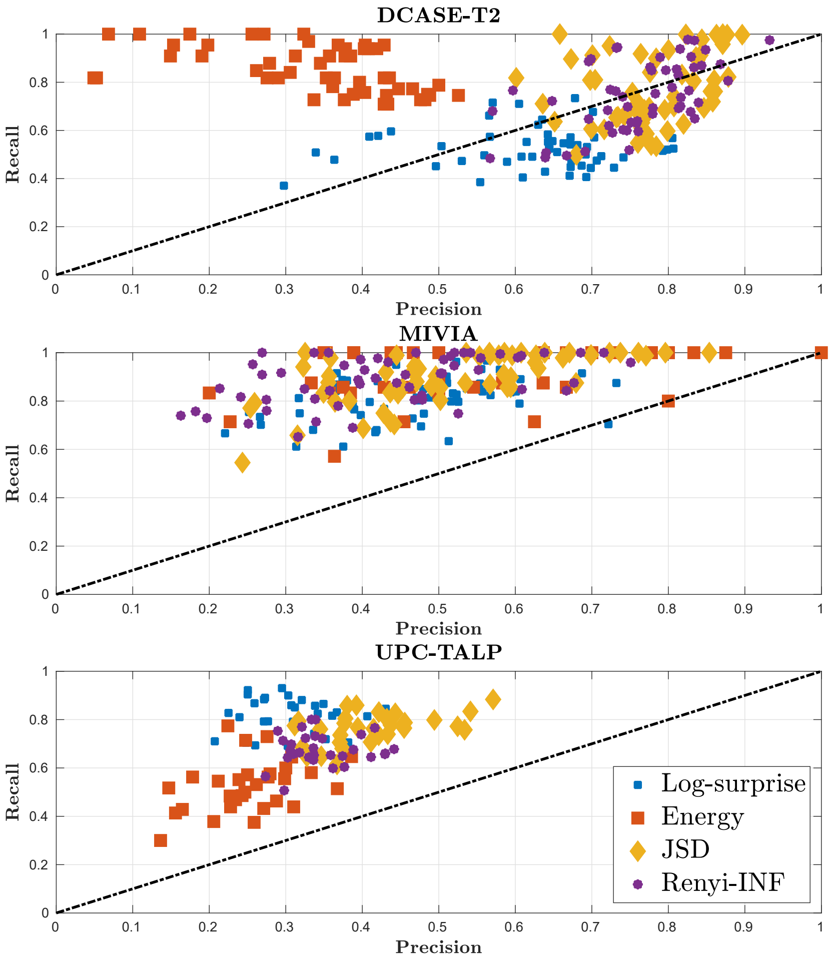


Figure 6.9: Scatter plots representing the Precision-Recall points for Log-surprise, Energy and two examples of Echoic Log-surprise, considering each individual file of the three datasets.

see this effect for MIVIA, although the F-scores showed that it also occurs for this dataset. Moreover, for MIVIA it can be observed that a significant number of files show $R = 1$ for Energy, JSD and Renyi-INF. This behavior means that every true positive point is detected for these specific files, although in exchange their P scores suggest that many false positive points are detected as well. Consequently, it would interesting to introduce some modifications in Echoic Log-surprise so the R score remains similar whereas the P score increases.

Globally speaking it seems that Echoic Log-surprise performs better than the rest of the techniques considered in this chapter, although there are still some aspects that need to be addressed. First, one of the limitations of this algorithm is that it uses a static threshold, which implies that the length of the saliency signal needs to be known beforehand. This limits the potential implementation of an online saliency algorithm. In addition, the results of this and the previous chapter were obtained in optimal recording conditions. Consequently, it is worth analyzing how all the aforementioned detection techniques behave when the audio files are polluted with noise of different nature. We will offer some solutions for these issues in the following two chapters.

6.6 Conclusions

This chapter can be summarized in a set of conclusions related to the aforementioned results. First of all, thanks to the regions representing the pairs (N_1, dth) it seems easy to determine that Echoic Log-surprise requires a different setup for each dataset. However, the necessity of validating these two parameters is less than idealistic since it is more adequate to have a value of (N_1, dth) that is common for the three datasets. Moreover, we have observed that the P and R scores of Echoic Log-surprise are disperse, as well as their areas of work, and we think that reducing this dispersion might affect positively to the performance of the system. Consequently, the need of a mechanism that eases the selection of such parameters emerges.

Second, an analysis of the global F-scores reveals that both Kayser and Kalinli saliency algorithms perform poorly, whereas Bayesian Surprise only works fine for one of the datasets. On the other hand, Energy, VAD and Bayesian Log-surprise produce reasonably good results, which suggest that they might be suitable for the task of detection. Moreover, the case of Energy is prominent since it is the technique that produces the best results for MIVIA however performing poorly for the other two datasets. Globally speaking its results are comparable to those of Bayesian Log-surprise, whose F-scores are almost equally high for the three datasets. VAD produces its best scores with DCASE-T2, although its performance compared to the rest of the datasets is not remarkable.

In global terms, and with the exception of MIVIA, the bar diagram with the F-scores suggests that all the divergence configurations proposed for Echoic Log-surprise produce better results than the rest of the algorithms. In fact, for the case of MIVIA their error bars overlap with Energy, which means that there is uncertainty about which one performs better for this dataset. Moreover, there are no important differences between the four statistical divergences for Echoic Log-surprise, which hints that there might be certain degree of freedom to choose the fusion technique.

Regarding the Precision-Recall scatter plots, where Energy, Bayesian and Echoic Log-surprise were compared, we observe that the P score for Energy tends to be low for all the audio files of DCASE-T2 and UPC-TALP, suggesting a big number of false positive points. This behavior is not observed for MIVIA, where the four techniques under analysis produce points located in similar regions of the space.

In addition, these clouds of points show that both Bayesian and Echoic Log-surprise produce similar groups of points, although with slightly bigger scores for the latter, suggesting that the multi-scale mechanism improves the detection capabilities of the system. In fact, the most usual change is an increase in the P score, which indicates that the number of false positive points decreases. In addition, for DCASE-T2 there is also an increase in the R score.

Finally, we can conclude that these results suggest that Echoic Log-surprise, with any of the statistical divergences, can be proposed as a solid mechanism to detect saliency, although its robustness against detrimental noisy effects needs to be addressed in order to assess its stability.

Chapter 7

Improving Echoic Log-surprise

In previous chapters we introduced the techniques that constituted the starting point of our analysis, Bayesian Surprise and Log-surprise, and described how they behave in three representative datasets. Moreover, in Chapter 6 we introduced our proposal, denoted Echoic Log-surprise, and explained the properties that characterize it. The analysis that we designed to test its capabilities allowed us to ascertain not only its detection power, but also some of the limitations it exhibits. Specifically, there are two main difficulties to deploy our saliency detector in real scenarios: first, the optimal free parameters need to be adjusted to the target scenario and the large dispersion observed in the optimal areas of operation is not desirable and second, the static threshold used hinders the on-line functioning of the algorithm since the whole signal needs to be processed to obtain it.

In this chapter we introduce some additional statistical divergences for Echoic Log-surprise, besides a new fusion strategy that we name mixture fusion which emulates the mechanisms underlying JSD. Moreover, we modify the thresholding stage of Echoic Log-surprise and include a dynamic thresholding algorithm, whose performance is thoroughly tested in the analysis section of this chapter. We show that the latter helps to reduce the dispersion in the areas of work and the Precision-Recall values for the three datasets when Echoic Log-surprise is used. Finally, we enlarge our benchmark with a set of algorithms that were specifically designed to detect onsets on music, in an area of research known as Music Information Retrieval (MIR).

7.1 Fusion strategies and statistical divergences

As we explained in Section 6.2, Echoic Log-surprise combines *dth* Log-surprise signals to obtain a global saliency descriptor from an audio signal, since it fuses information from different temporal scales varying from finer to rougher magnitudes of time depending on the size of the circular buffer N_z used in each one of them. Moreover, it should be noticed that the mechanism in charge of fusion is in fact conformed by statistical divergences or distances, from which we choose four well-known examples. We make the distinction between two fusion strategies, depending on how they manage scales when $dth > 2$. On the one hand, what we call *global fusion* does not take any special consideration and combines the dth Bayesian Log-surprise cues at once, whereas *local fusion* performs fusion pairwise using data from consecutive scales, since the divergences that belong to this group are designed to compare differences from only two distributions.

Interestingly, the results that we presented in Section 6.5 demonstrate that the technique that has the best detection capabilities is Echoic Log-surprise with JSD, which is an example of global fusion, and essentially compares each one of the scales computed using different buffer sizes against a reference that is obtained as their mixture distribution. However, it is not clear if the good performance of such fusion strategy has to do with the idea of combining all the scales at the same time, since the Precision-Recall plot does not show an anomalous behavior when compared with other divergence-based combination. For this purpose we propose another fusion strategy, *mixture fusion*, inspired both in local and global fusions. First, we define a mixture probability that is computed from the dth estimations as follows:

$$h^L_{MIX}(n) = \frac{1}{dth}\sum_{z=1}^{dth} h^L_z(n), \tag{7.1}$$

where $h^L_z(n)$ represents the histogram estimated using data from L temporal frames for the scale z for the current frame n.

As it was explained in Section 6.2, JSD is obtained by getting the Kullback-Leibler divergence of every distribution with respect to a reference one, which is computed as the mixture of all the dth incoming probabilities similarly to Equation 7.1. In our mixture fusion, we implement pairwise fusion by calculating the divergence among every distribution and that of the mixture:

$$s^L_{echoic}(n) = \sum_{z=1}^{dth} d_{local}(h^L_z(n), h^L_{MIX}(n)). \tag{7.2}$$

Figure 7.1 shows a clearer view of our proposal, where it can be observed that histograms $h^L_z(n)$ for $z \in \{1, \cdots, dth\}$ are first used for the computation of $h^L_{MIX}(n)$ and later on reused for pairwise fusion, just as in the schematic of Figure 6.5. From the statistical divergences that were described in Section 6.2, we only modify Renyi-INF and Cramer to work with the mixture scheme, since JSD is a global fusion technique and Bhattacharyya has a global version that will be introduced in this Section, denoted as Bhatta-N.

In addition to the new fusion strategy we consider the inclusion of four less well-known statistical divergences, which we introduced in [Rodríguez-

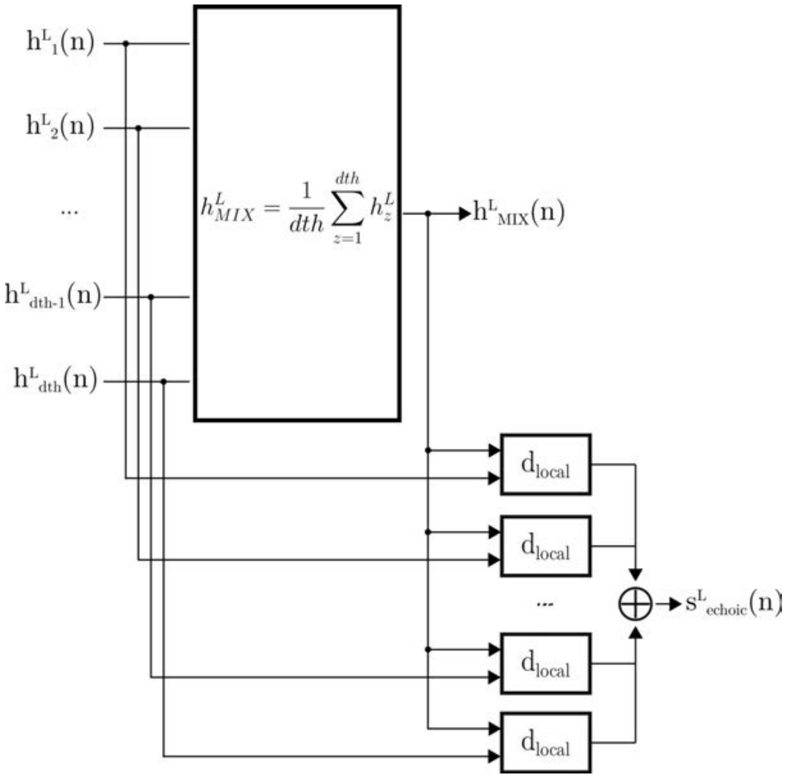


Figure 7.1: Schematic of the mixture fusion strategy, an attempt to combine the global and local fusion strategies described above by means of computing local fusion of every Log-surprise signal distribution as compared to a reference distribution obtained as a mixture of all the incoming distributions.

Hidalgo et al., 2019]. We study these new proposals since the initial four divergences that we presented in the previous chapter showed no significant difference in their performances:

Earth Mover's distance (EMD): also known as Wasserstein distance can be formally defined according to [Rubner et al., 2000; Rabin et al., 2008; Martinez et al., 2016] among other authors. If we consider that our distributions are modeled using histograms with the same number of bins N_{bins}, the computational implementation can be defined as:

$$d_{EMD}(h_z, h_{z+1}) = \sum_{i=1}^{N_{bins}} |\psi_i|, \tag{7.3}$$

where:

$$\psi_i = \sum_{j=1}^{i} (h_{z,j} - h_{z+1,j}), \tag{7.4}$$

and $h_{z,j}$ represents the j-th histogram bin of the z-th signal to be fused.

This technique is an example of the local fusion strategy, but can be modified to be compatible with mixture fusion considering that $h_{z+1,j} = h_{MIX,j}$.

Total Variation distance (TVD): This distance can be defined as:

$$d_{TVD}(h_z, h_{z+1}) = \max_i |h_{z,i} - h_{z+1,i}|, \tag{7.5}$$

so its value is obtained after computing the L_1-norm of the difference of the bins from both histograms, from which we keep only the largest difference [Levin et al., 2009].

This local fusion algorithm is also compatible with the mixture fusion strategy.

Hellinger distance: The Hellinger distance [Beran, 1977; Nikulin, 2001] is defined according to:

$$d_{Hellinger}(h_z, h_{z+1}) = \frac{1}{\sqrt{2}} \parallel \sqrt{h_z} - \sqrt{h_{z+1}} \parallel_2, \tag{7.6}$$

for discrete distributions, such as the histograms that we are considering in our work.

This distance can be used for both the local and mixture fusion strategies.

Bhattacharyya distance for n-distributions (Bhatta-N): In the particular case of Bhattacharyya distance, we make use of the following expression:

$$d_{Bhatta-N}(h_1, h_2, \cdots, h_{dth}) = -\log \sum_{i=1}^{N_{bins}} \sqrt[dth]{\prod_{z=1}^{dth} h_{z,i}}, \tag{7.7}$$

whose properties were described in [Kang and Wildes, 2015], and it allows to combine all the histograms from h_1 to h_{dth} at once. As a consequence, it is the second global fusion algorithm that we propose in addition to JSD.

7.2 Dynamic thresholding and peak-picking

As we mentioned in the previous chapter, the original Echoic Log-surprise algorithm makes use of a static threshold to determine what parts from the output saliency signal are onsets, discarding the rest. It is a fact that this implementation produces certainly adequate results, although it has a drawback that might limit its operability: the saliency from the whole signal needs to be computed beforehand, since the threshold is computed as the average value of the magnitude as it was defined in Equation 6.9. Consequently, the system can only process the signals that have been completely recorded and whose saliency has been computed, that is, it requires two passes.

We propose to find an adequate trade-off between the global performance of the system and an alternative that only requires a limited number of saliency frames to compute the threshold. This alternative could lead to an online saliency detection algorithm, where the length of the input signal is irrelevant. An interesting proposal was designed by [Rosão et al., 2012], whose dynamic threshold is defined as:

$$th_{out}(n) = th_{init} + \frac{\lambda}{2 \cdot M_{th} + 1} \sum_{n_{th}=-M_{th}}^{M_{th}} |s(n - n_{th})|, \tag{7.8}$$

where $s(n)$ represents the saliency signal that we are thresholding, th_{init} is a predefined minimum value for the threshold and M_{th} is the number of frames we are considering in order to update the threshold value, obtained from a window with a length of $2 \cdot M_{th} + 1$ frames. The parameter λ controls the update weight.

After the thresholding stage Rosão et al. include a peak-picking algorithm. This algorithm is in charge of selecting the values of the thresholded saliency signal that should be labeled as onsets, which gives another temporal signal that represents the onset detection signal $s_{th}(n)$:

$$s_{th}(n) = \begin{cases} 1 & \text{if } s(n) > th_{out}(n) \\ & \text{and } s(n) > s(n \pm m), \forall m \in \{1, \cdots, M_{th}\}, \\ & \\ 0 & \text{otherwise.} \end{cases} \tag{7.9}$$

The previous conditions establish that in order to be considered an onset a signal value $s(n)$ needs to be greater than the threshold $th_{out}(n)$ at the same temporal instant, and it also forces $s(n)$ to be the local maximum of a timespan of length $2 \cdot M_{th} + 1$.

However, it should be noticed that our implementation of this dynamic threshold differs with the work of [Rosão et al., 2012]. Our reference equation is:

$$th_{out}(n) = \frac{1}{M_{th} + 1} \sum_{n_{th}=0}^{M_{th}} |s(n - n_{th})|, \tag{7.10}$$

where we update the treshold th_{out} considering uniquely M_{th} past values of the saliency signal $s(n)$, instead of values from $n - M_{th}$ to $n + M_{th}$. We also remove the parameter th_{init} that states the minimum value reachable by $th_{out}(n)$, since a preliminary test showed that it was detrimental for Echoic Log-surprise and made it unable to get adapted to the dynamics of the datasets. The threshold that we propose is essentially computed as a moving average, although the peak-picking algorithm is kept without any modification.

Needless to be said, M_{th} is a parameter that needs to be validated, since it might perform differently depending on the dataset under analysis. To determine its value we select a validation subset of three files from each dataset and study the evolution of the performance with respect to this parameter. In a first instance we note that there is no coincidence between the three subsets, being $M_{th} = 32$ for DCASE-T2, $M_{th} = 256$ for MIVIA and $M_{th} = 64$ for UPC-TALP, the optimal values. In addition, it seems that the threshold requires a similar window for DCASE-T2 and UPC-TALP, although setting both of them to a common one only decreases the global performance. As a consequence, we decide to use a different M_{th} value for each dataset.

7.3 Optimal parameter setup

We perform the same analysis that was previously introduced in Section 6.4, but this time the matrices with the regions of N_1 and dth for optimal F-scores are obtained using the dynamic threshold and the peak-picking algorithm introduced previously in this chapter. Again, due to some limitations in the duration of certain audio files we avoid configurations where $N_1 > 1024$ or $N_z > 1024$. The results are depicted in Figure 7.2, where it can be observed that the area for DCASE-T2 remains similar to the one obtained for the static threshold configuration. On the contrary, there are some noticeable changes for MIVIA and UPC-TALP. The area for MIVIA becomes similar to the one observed for DCASE-T2, where the optimal point of work is found for $dth \in \{2, 6\}$ for small values of N_1, from 4 to 16. UPC-TALP evolves differently when the dynamic threshold is used, and it can be observed that most of the optimal points are situated at $dth = 2$ with a wide range of N_1 values.

Globally speaking, we observe that the optimal operation areas obtained after including a dynamic threshold seem to be more prone to use lower values of N_1, which mostly concentrate at $N_1 \in \{8, 16\}$. The depth dth still seems to be dataset dependent, and DCASE-T2 seems to prefer deeper models than the other two datasets.

There is still a need to verify if this thresholding mechanism improves the performance of the system in terms of the F-score. Again and similarly to what we explained in Section 6.5 for the static threshold results, these regions are used to obtain the weighted F-score for each one of the fusion

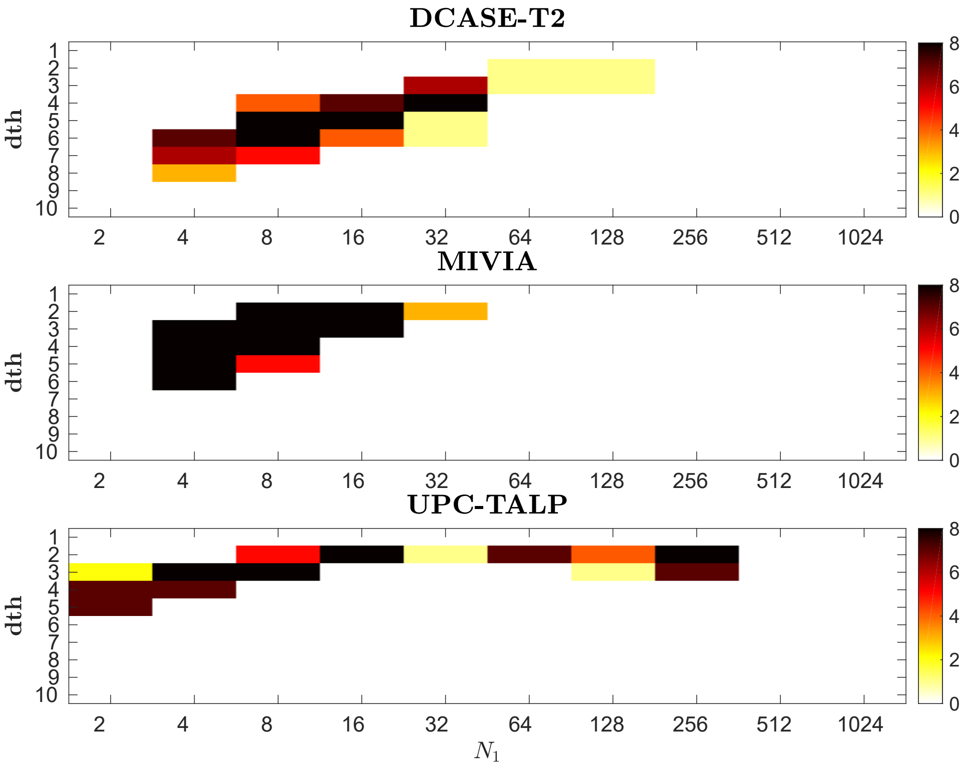


Figure 7.2: Area of work for a selection of fusion techniques considering the dynamic threshold for each dataset.

techniques explained in this chapter.

7.4 Comparison of static and dynamic thresholding

After the analysis of the regions for optimal F-score for both static and dynamic thresholding it seems necessary to evaluate the Precision-Recall points for each dataset and both thresholds. We show such assessment in Figure 7.3, where we start obviating the individual contribution of each file of the datasets in order to understand how the areas of work influence the results globally. For a specific dataset and threshold configuration, we start averaging the P and R scores from the available fileset for every divergence under analysis. These averages will be computed for the points (N_1, dth) of the associated optimal region of work, explained in Section 6.4 and 7.3, which means that each divergence will be represented by 10 different points. Since we are studying 8 divergences every cloud of points will be formed by 80 points. Then, we replicate this procedure for the rest of datasets

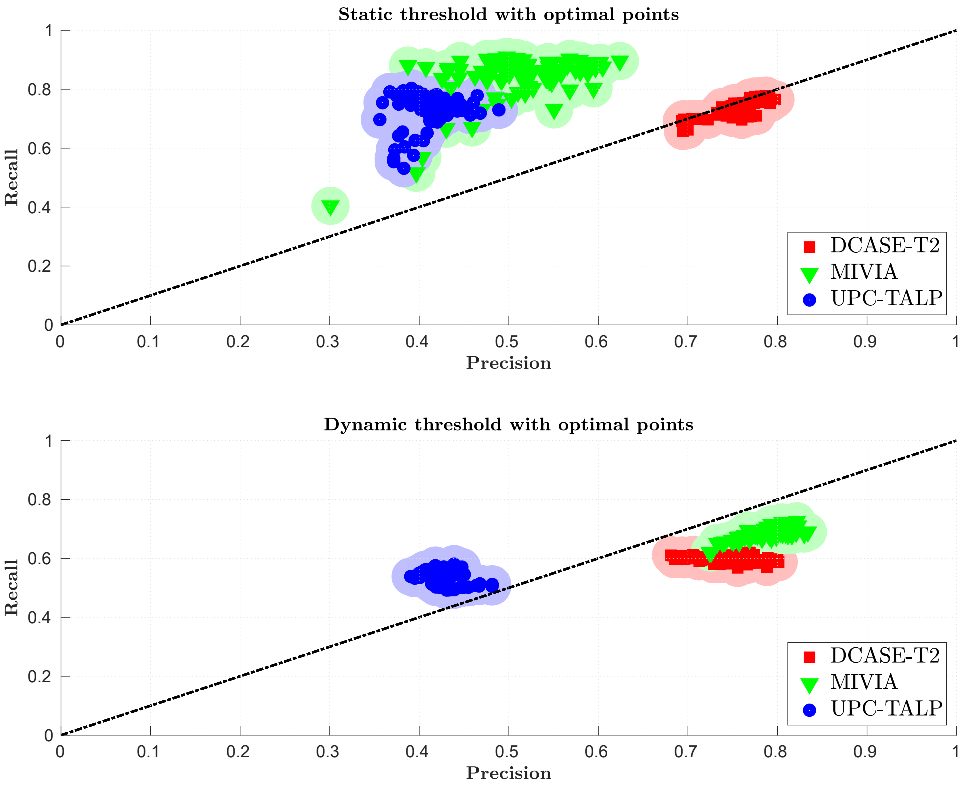


Figure 7.3: Precision-Recall scatter plots obtained from the optimal areas of work of each dataset, considering both static (top) and dynamic (bottom) thresholds.

and thresholds which produces the two scatter plots of Figure 7.3, each one formed by three clouds of points.

On the top scatter plot the results for the static threshold are depicted with different colors for each dataset. We can observe that there are clear differences between DCASE-T2 and the other two datasets. Particularly, it can be observed that DCASE-T2 seems to be isolated respecting to the other two datasets, sharing a similar R than the one of UPC-TALP and having a higher P score. This suggests that for DCASE-T2 the number of detected false positives is smaller in comparison to the other two datasets. In addition, the optimal points seem to be more concentrated in comparison to UPC-TALP and MIVIA, where the points are sparsely positioned in both axes. Although no conclusion about the F-score can be deduced with clarity from this graph, it seems that DCASE-T2 outperforms the other two. This observation is aligned with the conclusions that we draw from Section 6.5, where a subset of the current fusion techniques were used.

For the results with the dynamic threshold depicted in the second subplot

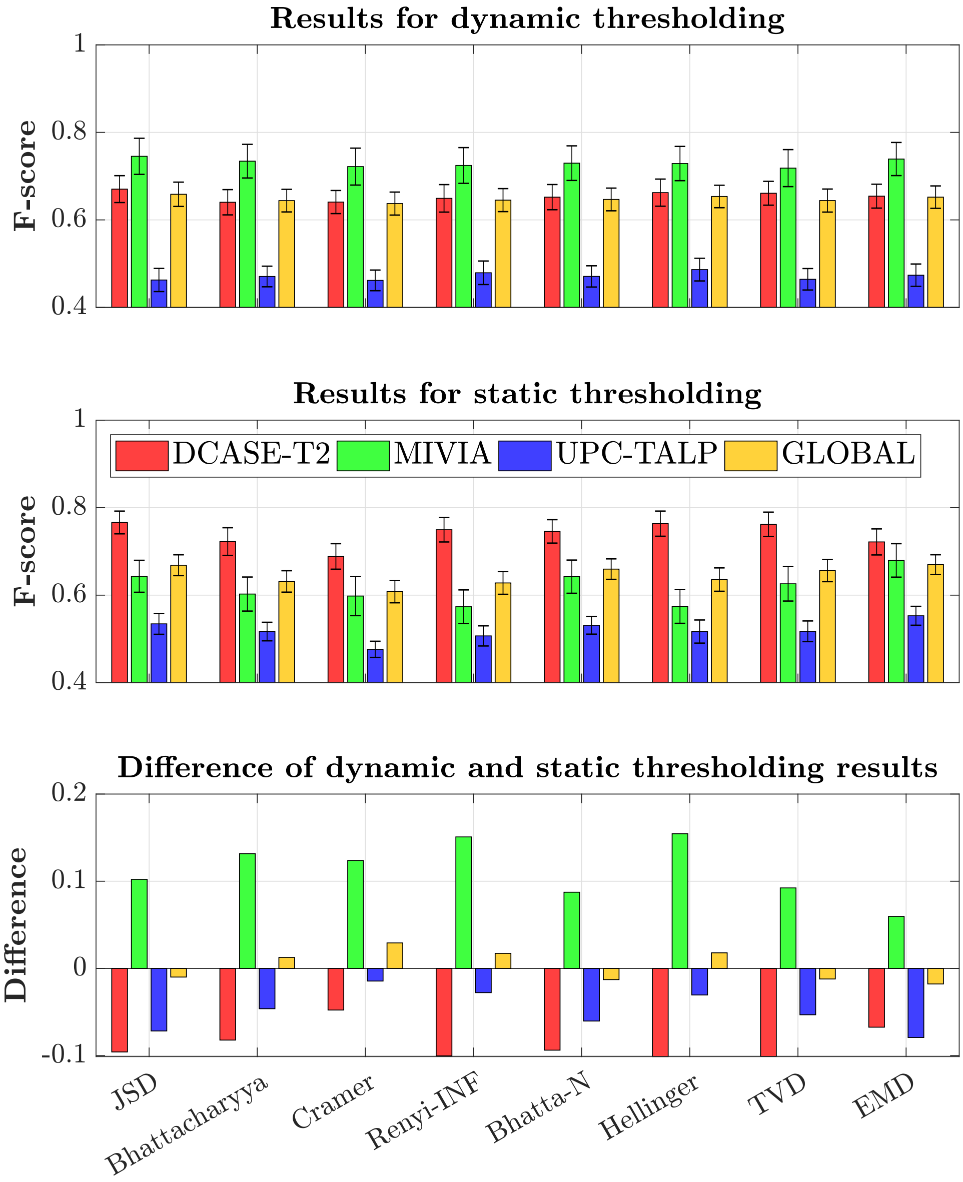


Figure 7.4: F-scores obtained for each dataset and fusion technique using dynamic and static thresholds for global and local fusion strategies. The last bar diagram shows the differences obtained after subtracting dynamic and static results.

of Figure 7.3 it can be observed that the behavior of the detectors changes globally. First, we note that R decreases for every dataset, which means that the number of false negative events increases implying that the systems become unable to detect some of the true positives. Concerning the P score, it remains similar for DCASE-T2 and UPC-TALP. Hence, for these datasets there is a clear degradation in the performance. However we observe that it improves for MIVIA, suggesting a global increase in the F-score. We also observe that for the three datasets the points in the scatter plots get more concentrated in comparison to the results obtained for the static threshold.

We interpret that this global reduction in R, the increase in P for MIVIA, and the concentration of points in the scatter plots imply that the dynamic threshold behaves more conservatively for such dataset, since it allows to reduce the number of false positive points without compromising and increasing the number of false negatives, which should be reflected in the F-score results for this particular dataset as a noticeable improvement of its performance.

To assess such results we designed Figure 7.4, which is divided in three bar diagrams. The top one represents the weighted F-scores for all the Echoic Log-surprise fusion techniques proposed in this thesis, obtained using the dynamic threshold. The middle subplot includes the same analysis but using the static threshold instead, whereas the bottom one shows the differences in the scores obtained after subtracting dynamic and static threshold F-scores. In addition to the weighted F-score, a global F-score can be computed using exclusively the optimal point for each audio file and divergence technique. These results can be consulted in the appendix A.

A priori it can be observed that there are no significant differences between the four fusion techniques proposed in the previous chapter and the ones introduced in the current one, at least in terms of F-score. Moreover, if we analyze the bar diagram that shows the existing differences between dynamic and static thresholding it can be observed that globally speaking the effect of using dynamic thresholding does not decrease the performance of the system. In particular, the F-score increases significantly for MIVIA whereas it decreases for DCASE-T2 and UPC-TALP. Echoic Log-surprise with JSD remains being the technique with the highest F-score in terms of global performance.

Figure 7.5 shows a comparison of the classical detection techniques studied before in Chapters 5 and 6, in addition to the fusion technique that shows the highest F-score for static and dynamic threshold, which is JSD. The results show that thanks to the dynamic threshold and the peak-picking algorithm Echoic Log-surprise is capable of overtaking the anomalously high F-score that Energy shows for MIVIA. On the other hand, there is a clear decrease in the performance for DCASE-T2 and UPC-TALP for the dynamic threshold.

We conclude that using the dynamic threshold introduces some interest-

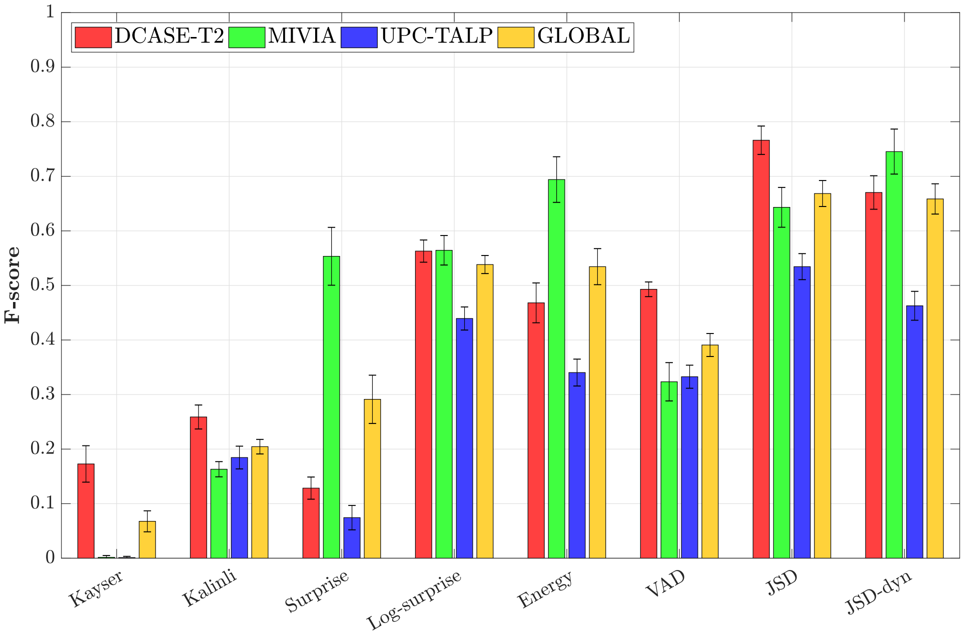


Figure 7.5: Comparison of classical detection techniques against Echoic Log-surprise with JSD, with static and dynamic thresholding.

ing improvements, since it increases the performance for MIVIA, it helps to homogenize the values of N_1 and allows the whole algorithm to become independent of the duration of the saliency signal, in exchange for a decrease in the performance observed for DCASE-T2 and UPC-TALP.

7.5 Comparison of fusion strategies considering local and mixture fusion methodologies

In Chapter 6 we devoted some time to explain two fusion strategies that we denoted as *global* and *local*. We explained the need of using local fusion for the divergences and distances that are designed to compute statistical differences from only two distributions. Alternatively, since results showed that the best detection performance is obtained using Echoic Log-surprise with JSD, a global fusion technique, we decided to investigate if such behavior has to do with the fact that this technique is computed with respect to a mixture distribution, obtained from the distributions whose differences are supposed to be obtained. As a solution, in this chapter we introduce an alternative fusion strategy that we name *mixture* fusion, which in fact attempts to emulate the behavior of JSD by computing local pairwise fusion with a common reference distribution, obtained as a mixture of the incoming distributions.

In Figure 7.6 we carry out a comparison, where three bar diagrams are represented. The first diagram conveys the F-scores for mixture fusion for the compatible techniques, which is followed by the diagram obtained for local mixture and the same set of techniques. In both cases the dynamic threshold is selected, since we agreed that this improves how Echoic Log-surprise behaves, in exchange for a significant reduction in the weighted F-score for DCASE-T2 and UPC-TALP. Finally, the diagram at the bottom illustrates the differences existing between mixture and local fusion. A careful view of the last diagram shows that the magnitude of such differences is insignificant, to the extent that the biggest difference is $\Delta F = -0.0115$. In addition, there is no clear pattern in the differences obtained for any dataset. For some of the techniques global performance slightly increases, whereas for other examples such as Renyi-INF or Hellinger there is a global decrease. In fact, for the latter there is a clear decrease in the F-scores for every dataset.

Additionally, Figure 7.7 shows the Precision-Recall plots for the mixture fusion techniques shown previously in this Section, and also the results for Hellinger with the dynamic threshold using local fusion. Our starting hypothesis is that, as it is observed for the F-scores, no noticeable variation should occur in the distribution of P and R. The results confirm our hypothesis, since the majority of points for every dataset and technique are concentrated around the same regions, which leads to the conclusion that the mixture fusion strategy does not contribute to improve the results obtained previously for local fusion.

7.6 Comparison with onset detection techniques for MIR.

Finally, in addition to the algorithms explained in Section 7.4 we decide to evaluate the performance of some specific onset detection methodologies designed for MIR, which were introduced in Section 3.7. We start studying their F-scores, illustrated in Figure 7.8, where we compare two significant examples of Echoic Log-surprise against four classical MIR onset detection techniques: CD, SF, WPD and NWPD. The weighted F-scores are computed for Echoic Log-surprise considering our dynamic threshold, whereas for the MIR methodologies we keep the threshold designed by [Rosão et al., 2012], which produced better results for these techniques during the validation process.

Globally speaking three of the MIR algorithms produce relatively high F-scores, with the exception of NWPD. Considering that it is quite similar to WPD with the exception of its normalization procedure, this mechanism seems to cancel the peaks that both algorithms are capable of detecting. In

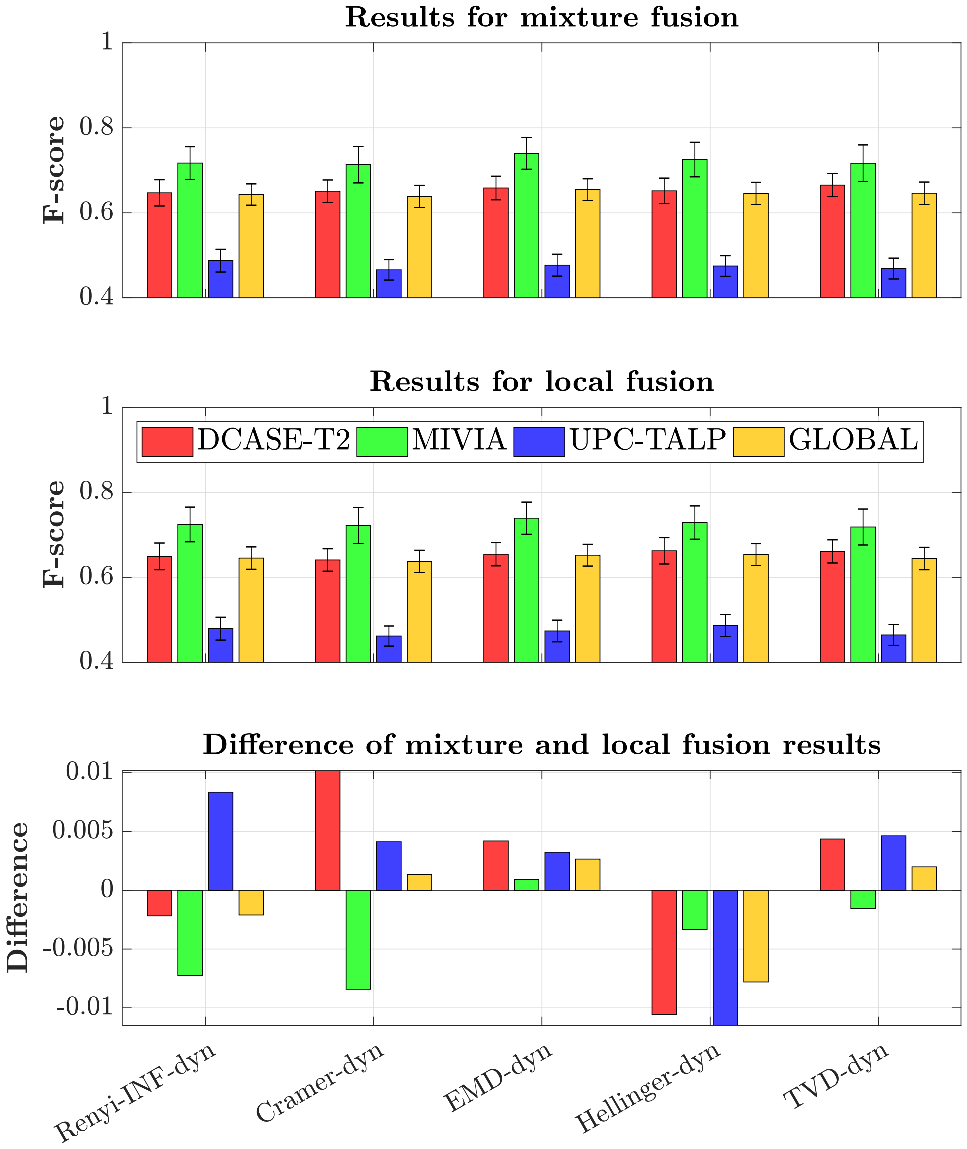


Figure 7.6: F-scores obtained for each dataset and fusion technique using mixture and local fusion strategies. The last bar diagram shows the differences obtained after subtracting the F-scores of both strategies.

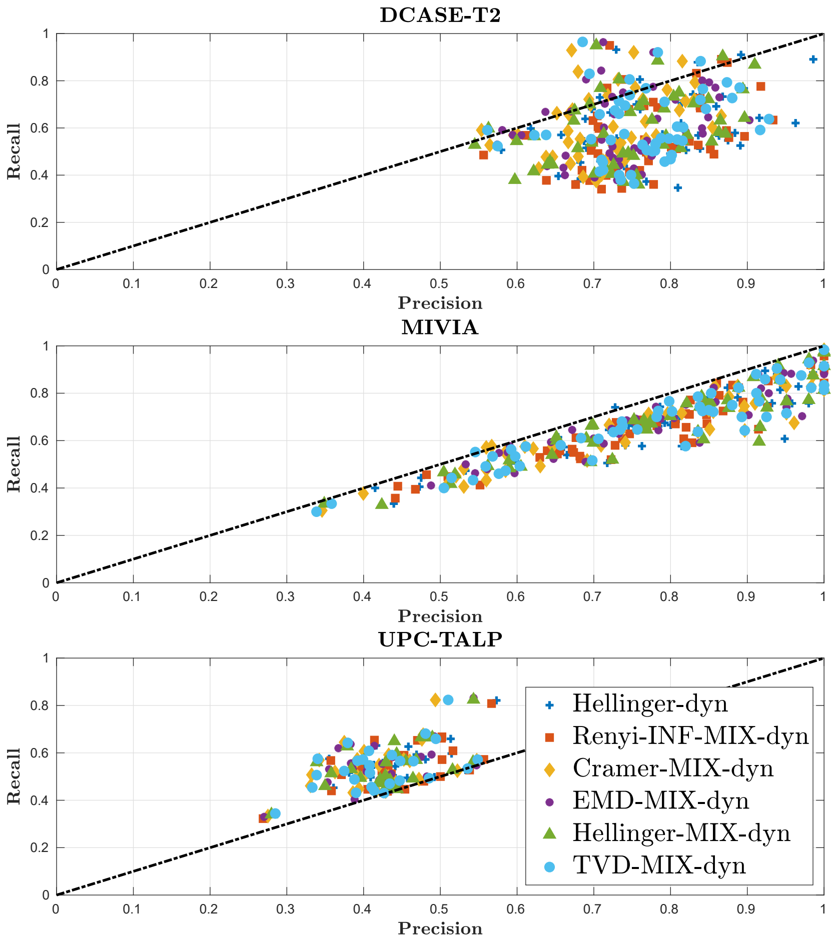


Figure 7.7: Precision-Recall scatter plots computed for all the files of each dataset considering mixture and local fusion strategies.

fact, if we study the scores of the rest of the MIR techniques WPD produces the highest ones for the three datasets. However, the differences between such techniques is not that obvious, since their confidence intervals overlap.

None of the two Echoic Log-surprise implementations presented in this diagram are surpassed by the MIR onset detection techniques, an effect that occurs thanks to the performances obtained for DCASE-T2 and MIVIA. In fact, some of the MIR techniques manage to obtain higher results for UPC-TALP, although confidence intervals show that these are not significant. This might be a consequence of the dynamic threshold, since we observed in previous analyses that it decreases the performance for DCASE-T2 and UPC-TALP. However, globally speaking results suggest that for the task of saliency detection Echoic Log-surprise manages to detect better the events.

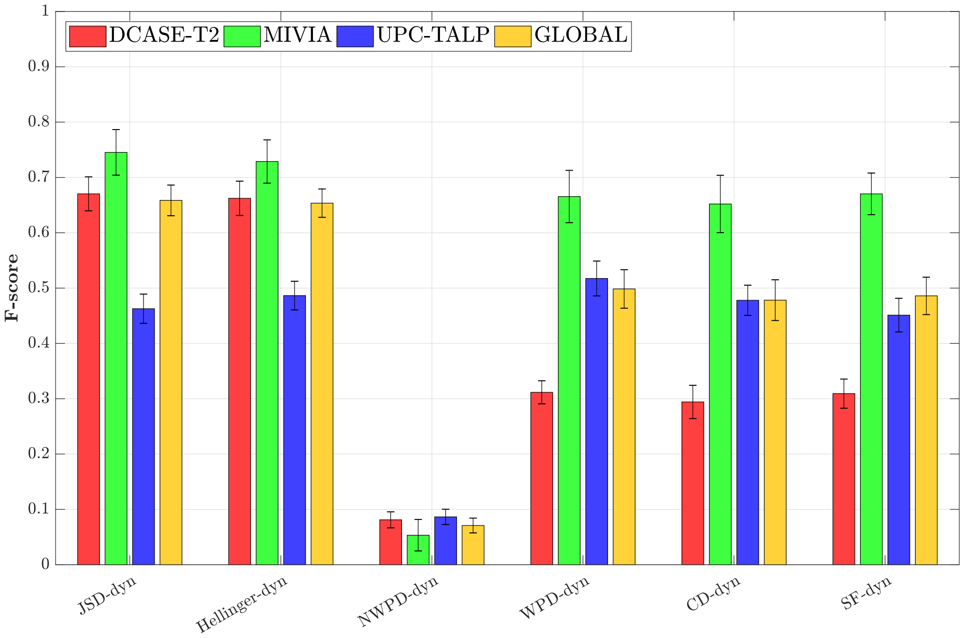


Figure 7.8: Weighted F-scores obtained for MIR onset detection techniques and two Echoic Log-surprise examples, namely JSD and Hellinger, all of them computed using a dynamic threshold.

If we represent their Precision-Recall scatter plots as in Figure 7.9 the first noticeable fact is that MIR techniques are concentrated in similar areas. In contrast, for the depicted Echoic Log-surprise technique, P and R are clearly higher for DCASE-T2 and, the majority of times, also for MIVIA. For UPC-TALP we observe that points concentrate in the same area than MIR techniques. Consequently, and with the exception observed for UPC-TALP, results suggest that Echoic Log-surprise seems to be a better detector than these techniques for the task under analysis.

7.7 Conclusions

In this chapter some novel mechanisms are introduced to improve the behavior of Echoic Log-surprise, being the most remarkable the addition of a dynamic threshold. The regions representing (N_1, dth) for the three datasets with such threshold scheme evolve to have some common values for N_1. However, Echoic Log-surprise still requires different dth values for each dataset, with deeper values for DCASE. Consequently, these regions suggest that taking a common value of $N_1 \in \{4, 8, 16\}$ suits any of the datasets, reducing the validation process to determining the proper value of dth.

In addition, a global Precision-Recall scatter plot analysis shows that

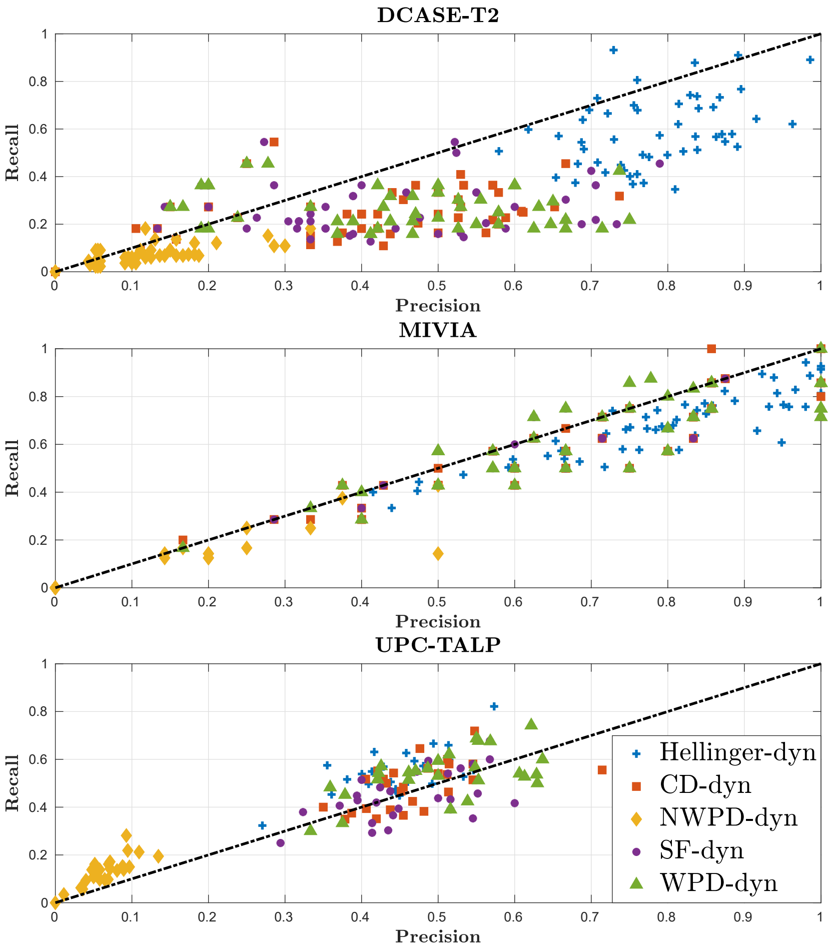


Figure 7.9: Precision-Recall scatter plots computed for all the files of each dataset considering Hellinger Echoic Log-surprise and the MIR onset detection techniques.

points for each dataset concentrate in different regions depending on the thresholding methodology. Particularly, the common effect of dynamic thresholding is a significant reduction in the R value, which translates into a greater number of false negative points. Nevertheless, dynamic thresholding also favors an increase in P for MIVIA, implying that the number of false positives decreases for this particular dataset. Moreover, the regions where the points for each dataset concentrate are more compact for the dynamic threshold, even for MIVIA and UPC-TALP, showing that there is a lower variance in the scores of the audio files. Consequently, we deduce that the dynamic threshold influences positively the detection performance of the system, forcing it to be more conservative and reducing the number of detected false positive points for MIVIA. In exchange, there is a reduction in the R score for the other two datasets.

Studying the variations in the F-scores suggest that between the four additional divergences studied for Echoic Log-surprise in this chapter and the ones of Chapter 6, there is no significant difference in terms of F-score for any of the thresholding methodologies.

Comparing one threshold to the other shows that dynamic thresholding increases the F-scores for MIVIA, an effect that can be observed for any of the divergences, whereas there is a significant reduction in the scores of the other two datasets, specially for DCASE-T2. The technique that produces a slightly better F-score is JSD, which shows a superior performance in comparison with the classical detection techniques studied in previous chapters. Moreover, it is capable of outperforming Energy for the MIVIA dataset, in addition to the fact that Echoic Log-surprise performs clearly better for the global results even without this new threshold. Consequently we consider that including a dynamic threshold is a significant addition to Echoic Log-surprise, since it eases the selection of N_1, reduces the variability in the P and R scores and also allows the algorithm to work online.

We also study a new fusion methodology that we named mixture fusion, inspired in the divergence mechanism of JSD. The analysis that we performed shows no significant differences between the methodology that we introduce and local fusion, and we conclude that this contribution is unfortunately insignificant for our work.

Finally, we decide to study some alternative algorithms that are designed specifically to detect onsets in music. We assessed their detection capabilities and the results suggest that they are capable of producing remarkably good performances, although Echoic Log-surprise manages to produce the highest F-scores.

Chapter 8

Assessing the robustness of the detection techniques

In the previous chapters we introduced several detection techniques designed to detect saliency from audio signals. Their properties and detection capabilities were extensively studied in an attempt to characterize them and to determine which one could be more useful for the detection of salience.

However, our previous analysis did not take into account the fact that on every real scenario the acoustic environment is plagued by distortions, whose effects in the detection process cannot be disregarded. Consequently, we devote this chapter to test how the aforementioned techniques perform when the datasets we previously employed are contaminated with noise of different nature. Our final goal consists in determining which one of these algorithms is more robust against noisy interferences. This chapter expands the conclusions of our previous work [Rodríguez-Hidalgo et al., 2018b], where we thoroughly studied the behavior of Echoic Log-surprise against noise considering exclusively static thresholding.

8.1 Noise contamination

As we explained in Section 4.3.4 and in our previous work [Rodríguez-Hidalgo et al., 2018b], for the contamination of the audio signals we have used the DEMAND collection of noises [Thiemann et al., 2013]. It comprises different real-world noise files acquired using an array of microphones at $F_s = 48$ kHz, from which we have chosen the second channel. The noise collection is divided into six categories, four of them captured indoors and the other two, outdoors. From a total of 18 noise files, we have selected six different ones for our analysis, one per category:

- DKITCHEN: belongs to the 'Domestic' category, and contains audio recorded in a kitchen during the preparation of a meal.

- NFIELD: was captured from a sports field where there were several people. It belongs to the 'Nature' category.

- OHALLWAY: contains the sounds of groups of people passing by along a hallway. It belongs to the 'Office' category.

- PCAFETER: from the 'Public' category. As its name indicates, it was captured on a cafeteria placed inside an office.

- SCAFE: was also acquired on a cafeteria, but placed on a public square instead. It is included into the 'Street' category.

- TBUS: contains sounds captured inside a public bus. Its category is 'Transportation'.

There are mainly two reasons why we choose this dataset. First, the sounds that it contains were captured considering a wide variety of real-life scenarios, allowing to test the behavior of all the analyzed systems in a diversity of acoustic environments. Second, other noise datasets developed for speech-related tasks such as Noisex-92 [Varga and Steeneken, 1993] and Chime-4 [Vincent et al., 2017] have a lower sampling frequency of $F_s = 16$ kHz, or $F_s = 8$ kHz in the case of Aurora-2 [Pearce and Hirsch, 2000]. However, since we are working with higher sampling frequencies we consider that DEMAND is more appropriate.

In addition, for the sake of comparison with other robustness studies, we have also included white Gaussian noise in our tests. We denote this modality *WHITE*.

In summary, we have seven different noise types, which are added to the audio signals using the Voicebox Toolbox [Brookes, 1997] considering SNR values from -5 dB to 20 dB in 5 dB steps. The noise addition algorithm computes the signal level using the P.56 ITU-T recommendation [ITU-T, 1994]. Finally, we also obtain the results for the *Noiseless* condition. Although we performed our analysis for seven different noise files, six from the DEMAND dataset and one with white noise, in this work we simplified it to the extent that the results for these seven contaminating noise types were grouped together and averaged. We prefer to perform the analysis in such a way since results shall be clearer to analyze, besides the fact that we have already performed this thorough study in our previous work [Rodríguez-Hidalgo et al., 2018b]. As a consequence, for every SNR level the scores for the seven noise types are averaged to produce a F-score that represents the global performance of the technique under analysis.

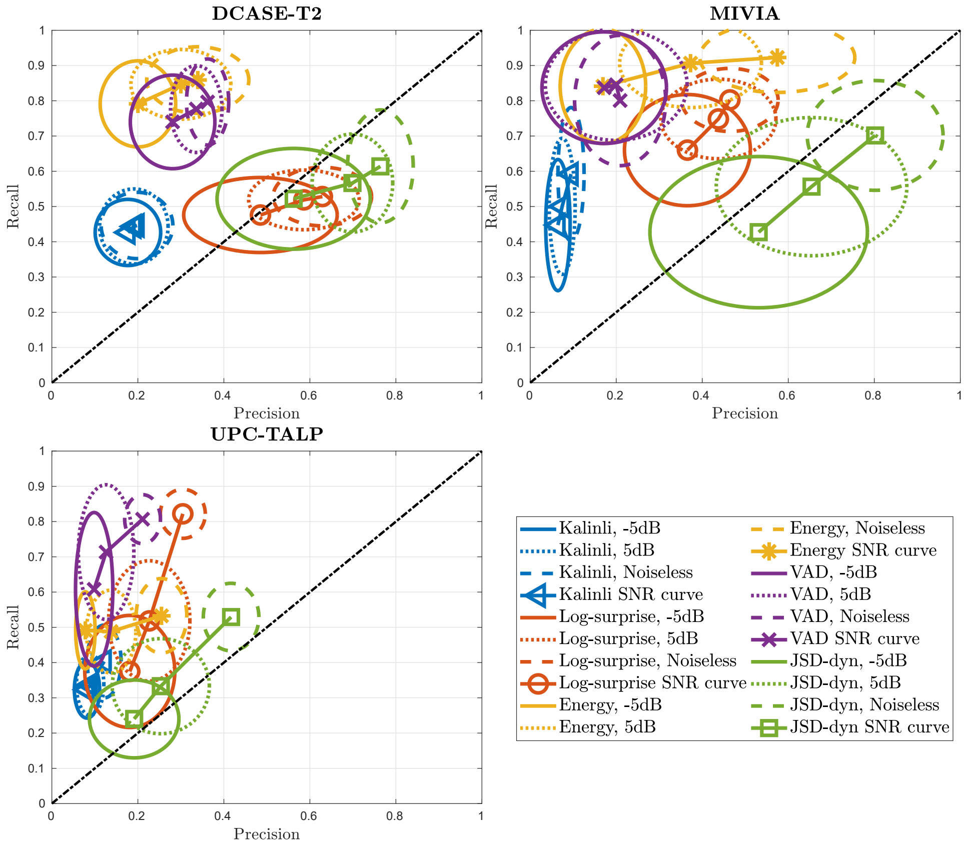


Figure 8.1: Precision-Recall scatter plots showing some detection techniques at different SNR values.

8.2 Robustness for classical and salience detection techniques

We start our analysis of the robustness against noise by considering the scatter plots depicted in Figure 8.1 for P and R, which represent the performance evolution with respect to the SNR for five detection algorithms: Kalinli, VAD, Energy, Bayesian Log-surprise and Echoic Log-surprise based on JSD with the dynamic threshold. Bayesian Surprise is not included in this analysis to keep the visualization of the plots as clear as possible. The values of P and R represented by this set of ellipses are calculated as the weighted sums for each one of the files conforming each dataset, which are obtained after grouping the results for the seven noisy files used to contaminate audio. Then, from this set of points we compute the mean and the standard deviation (stdev) for P and R, which are shown for three different SNR values, namely $SNR \in \{-5, 5\}$ and *Noiseless*. The center of each ellipse represents the aforementioned mean value, whereas the width of the ellipse in the horizontal and vertical axes represent the standard deviation for P and R respectively.

For DCASE-T2 we observe that Kalinli shows little changes in its P and R for the three SNR configurations. Regarding Energy, it can be observed that both the mean and the standard deviation increase their values for P, which means that the global number of false positives seems to decrease, although the uncertainty about them increases since this does not occur for all the audio files. VAD evolves similarly, although its standard deviation gets lower for P.

If we focus on the worst noise configuration, obtained for $SNR = -5$ dB, the plot suggests that normally Log-surprise and Echoic Log-surprise have their average P and R scores in the diagonal, with bigger scores for the latter as it has been seen in previous chapters. In fact, it can be observed that the relative positions of all the techniques with respect to each other remain similar for the three noise conditions already considered for DCASE-T2, which globally means that JSD seems to be the algorithm that produces the biggest scores in all these conditions.

In the case of MIVIA with $SNR = -5$ dB we observe that Energy and VAD show similar results. On the contrary, Log-surprise shows a bigger P and a smaller R, which means that it produces less false positive points and more false negatives. JSD is in a more advantageous position along the diagonal, although its standard deviation values are clearly bigger in comparison with the rest of the techniques. For the *Noiseless* condition what seems to occur is that VAD almost changes its P and R values, whereas Energy clearly improves its P score. Log-surprise and JSD increase their performance uniformly on both P and R scores, being the latter technique the one that seems to produce the biggest F-score. What these points suggest is that JSD and Energy perform similarly when there is no noise, although when the SNR conditions worsen, Energy starts to underperform. This is a hint that for this dataset JSD might be more robust against the influence of noise.

For UPC-TALP there is a quite noticeable transition from $SNR = -5$ dB to the *Noiseless* condition, and the P and R scores are usually smaller in comparison with the other two datasets. Consequently, a smaller F-score is expected for this dataset for any of the SNR configurations.

This Precision-Recall analysis is complemented by the F-score results depicted in Figure 8.2, which are divided in two columns of plots. The first column shows a set of bar diagrams, one for each dataset, which contain the same information than the previous analysis of this work: the weighted F-score for the detection algorithms, obtained for six different SNR values in addition to the *Noiseless* condition. In the second column we represent a set of graphs denoted candlestick charts, which are commonly used in stock trading analysis. The red rectangles represent the differences existing

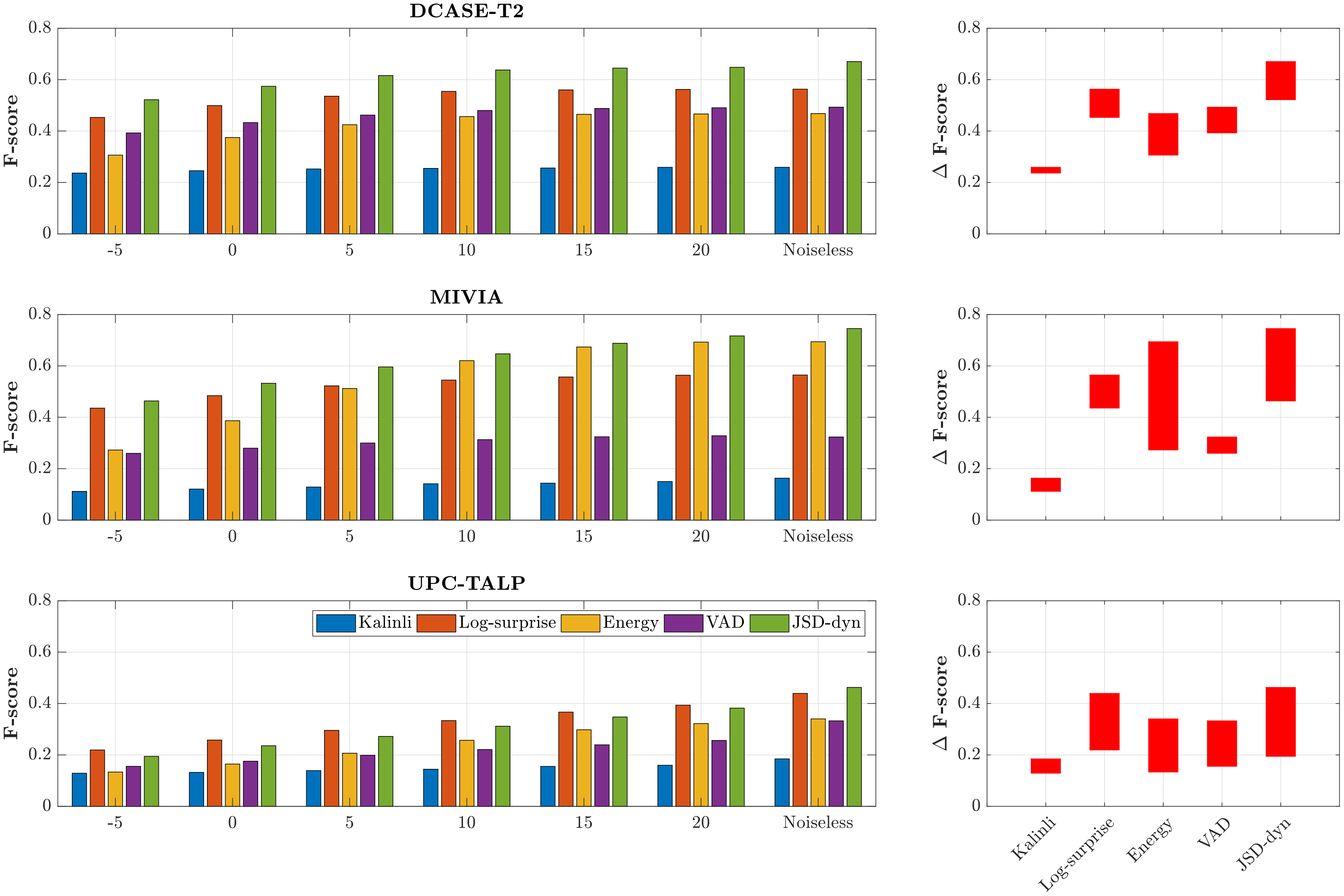


Figure 8.2: Bar diagrams and candlestick charts representing the F-scores and their variations for a set of onset and saliency detection techniques.

between the F-scores for $SNR = -5$dB and the results for the *Noiseless* condition. There are two properties that a robust technique should accomplish to be categorized that way. First, it is necessary that the minimum value of the rectangle is as high as possible, since this restriction implies that its F-score for $SNR = -5$dB is high. Secondly, although not critical it is desirable that the area of the rectangle is small, because this means that the variations observed in the F-scores for the two extreme SNR configurations are small. Jointly with the previous condition for $SNR = -5$dB, this implies that a technique behaves properly for all the SNR values.

A technique that shows little robustness for the stimuli used to contaminate audio is Kalinli, where for the two extreme SNR values the F-scores are low. What this behavior suggests is that the technique performs poorly for any noise condition. Log-surprise is more robust than Energy and VAD for UPC-TALP and DCASE-T2, since its rectangles are equal or smaller and always positioned with a better minimum value than the two other techniques. For MIVIA, it clearly outperforms VAD, although Energy shows a bigger F-score for the *Noiseless* condition. However, it does not mean that for this dataset Energy is more robust, since the observed rectangle for such technique is huge and offers a small F-score for $SNR = -5$ dB. Consequently, we can confirm that Log-surprise is clearly more robust than VAD, and is capable of outperforming Energy in the worst noisy scenarios.

Optimal conditions are observed for JSD, where the rectangle minimum values are higher than for the rest of the detection techniques for both DCASE-T2 and MIVIA, and close to the one of Log-surprise for UPC-TALP. The rectangle areas for JSD and Log-surprise are similar for DCASE-T2 and UPC-TALP, although it is higher for the former technique in the case of MIVIA. However, what seems to occur is that the results for $SNR = -5$ dB and *Noiseless* are normally higher for JSD. Consequently, this analysis suggests that Echoic Log-surprise computed using JSD and dynamic thresholding is more robust against the influence of noise than the other detection techniques presented in this section.

8.3 Robustness for static and dynamic thresholding

In Section 7.3 we evidenced the existing differences between static and dynamic thresholds, and we concluded that despite the scores suffered certain degradation, the system was capable to improve its performance for some datasets, in addition to other advantages related to the usage of a dynamic threshold algorithm. However, the robustness of such methodology needs to be assessed.

Figure 8.3 represents the Precision-Recall scatter plots obtained for JSD, Hellinger and Bhatta-N considering static and dynamic thresholds. In addi-

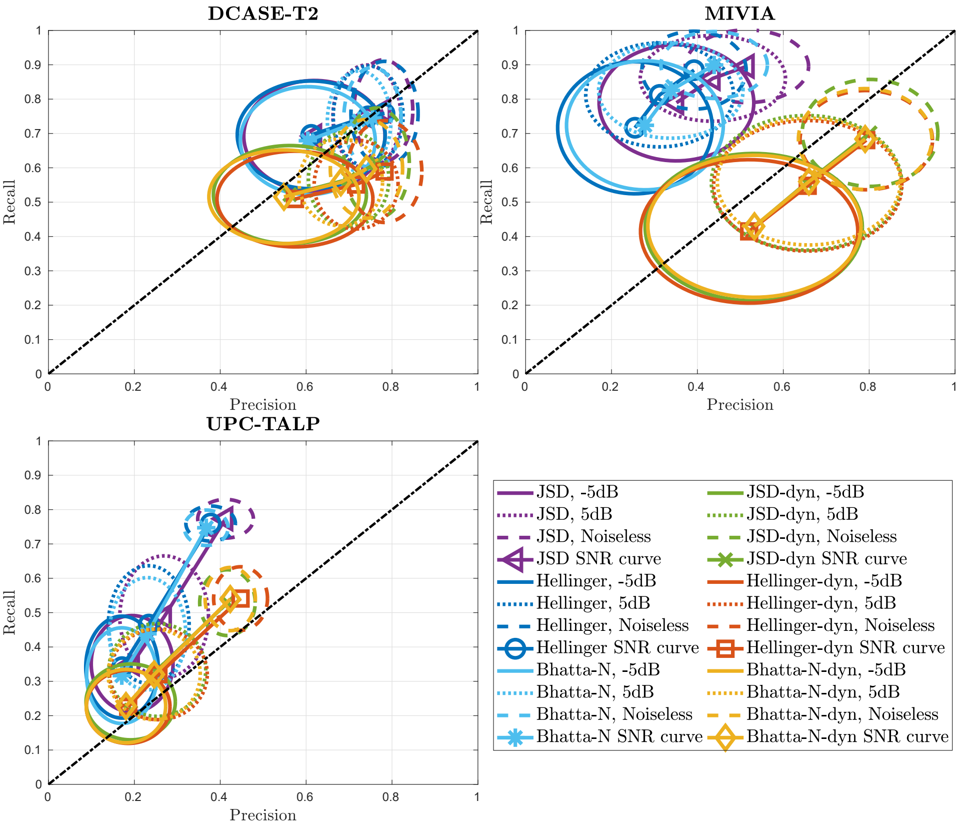


Figure 8.3: Precision-Recall scatter plots show the evolution of Echoic Log-surprise for static and dynamic thresholding for different SNR configurations.

tion to the comments made previously for Figures 6.9 and 7.3, what seems remarkable is the fact that the three techniques perform similarly for every SNR configuration. Additionally, an interesting effect that can be observed is the reduction of the standard deviation of P when the SNR increases. Since the behavior of both static and dynamic thresholds is similar to the one depicted for Figure 7.3 we conclude that the robustness of Echoic Log-surprise holds when any of the thresholds is used. Moreover, using static thresholding grants bigger R scores in general, which for DCASE-T2 and UPC-TALP seems to produce better F-scores.

This conclusion is confirmed by the bar diagrams and candlestick charts depicted in Figure 8.4. The charts allow to see that the static threshold seems to be clearly more robust for DCASE-T2 and UPC-TALP, since its F-scores for $SNR = -5$ dB are commonly bigger than for dynamic thresholding, although the areas of the rectangles remain similar in both conditions. In contrast, for MIVIA the rectangles of the dynamic threshold show similar bottom values to the ones for static thresholding, although the top F-scores for the former threshold configuration are clearly better, implying that for this dataset the dynamic threshold seems to be more robust.

8.4 Robustness of MIR onset detection techniques

In addition to the robustness analysis that we performed for the detection techniques of Section 8.2, we consider necessary to measure how the MIR onset detection algorithms perform when the conditions are worse than ideal. To carry out this task we introduce the Precision-Recall scatter plots of Figure 8.5 that show some differences between how NWPD behaves with respect to the other three onset techniques as it occurred previously in Section 7.6. For the rest of the techniques, it is obvious that their performances vary from one dataset to another. However, the three techniques seem to perform in a similar way and our comparisons will be done with respect to them.

Echoic Log-surprise based on the JSD with dynamic threshold produces remarkable scores for DCASE-T2, where its P and R values are the best for every SNR configuration. This behavior is not that clear for MIVIA, where the top three MIR techniques and Echoic Log-surprise produce similar ellipses for low and medium SNRs. However, for higher values of SNR the trend changes and JSD gets bigger values of both P and R. Finally, for UPC-TALP it occurs that JSD produces the highest scores for $SNR = -5$ dB and $SNR = 5$ dB. However, for the *Noiseless* configuration the scores of all the techniques are similar, to the extent that WPD gets slightly better P-R values than the rest.

Nevertheless, globally speaking and thanks to the behavior observed for small values of the SNR we can conclude that Echoic Log-surprise is more robust against the distortions that contaminate the audio files.

We also introduce an analysis of two of the classical detection techniques that show the best F-scores, VAD and Energy, and compare them with WPD, the MIR algorithm that produces the best detection results. Our analysis is depicted in the Precision-Recall scatter plots of Figure 8.6, where we represent the ellipses where the techniques produce their average scores in comparison to Echoic Log-surprise, computed using JSD and the dynamic threshold.

After analyzing the results for DCASE-T2 a remarkable effect is that for WPD there is little variation in the values of P and R for different SNR configurations. These are more noticeable for VAD and Energy, which due to their higher R scores should produce a better F-score than WPD. For MIVIA we observe that VAD remains in a similar location of the Precision-Recall graph for all the SNR configurations. However, there is a noticeable degradation for Energy. In general, it can be observed that WPD produces

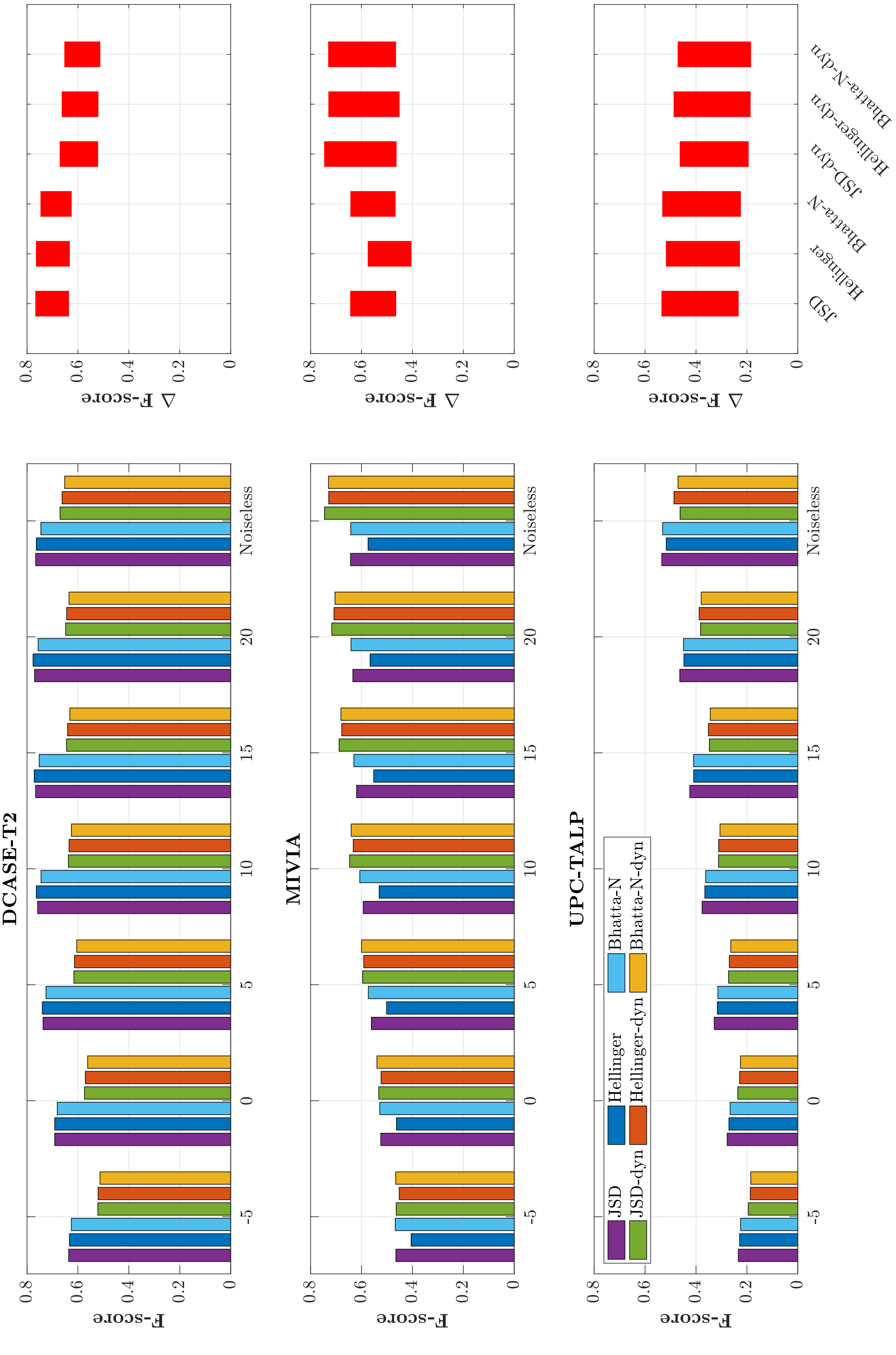


Figure 8.4: Bar diagrams and candlestick charts to compare some Echoic Log-surprise techniques considering both static and dynamic thresholding.

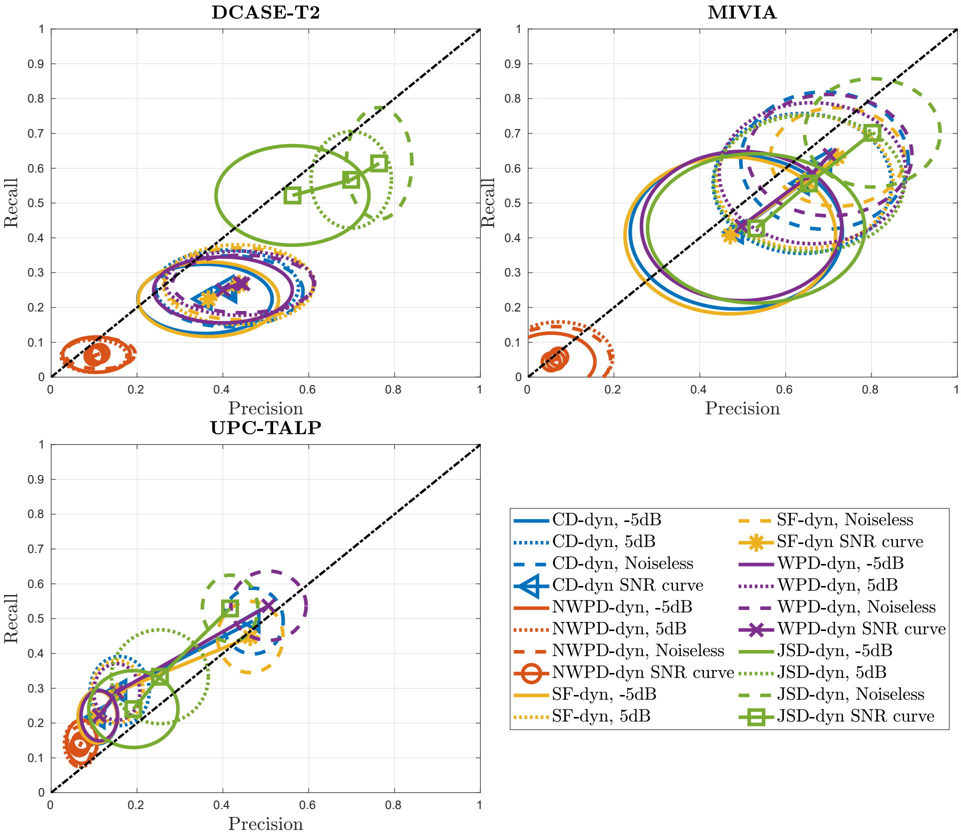


Figure 8.5: Precision-Recall scatter plots representing the performance variation of some MIR onset detection techniques against Echoic Log-surprise considering different SNR values.

scores close to the diagonal, which indicates that the number of false positives and negatives remain similar. Thanks to these locations it can be asserted that its F-score is usually bigger for any SNR than for the other two techniques, but only for this dataset. The case of UPC-TALP can be controversial, since for $SNR = -5$ dB and $SNR = 5$ dB WPD and Energy produce the smallest F-scores.

From the conclusions above, we can deduce that none of the classical or MIR techniques outperforms the rest in the detection of saliency, since all of them behave quite differently for each dataset. What seems clear is that Echoic Log-surprise shows, in general, more robust results, specially for DCASE-T2 and MIVIA.

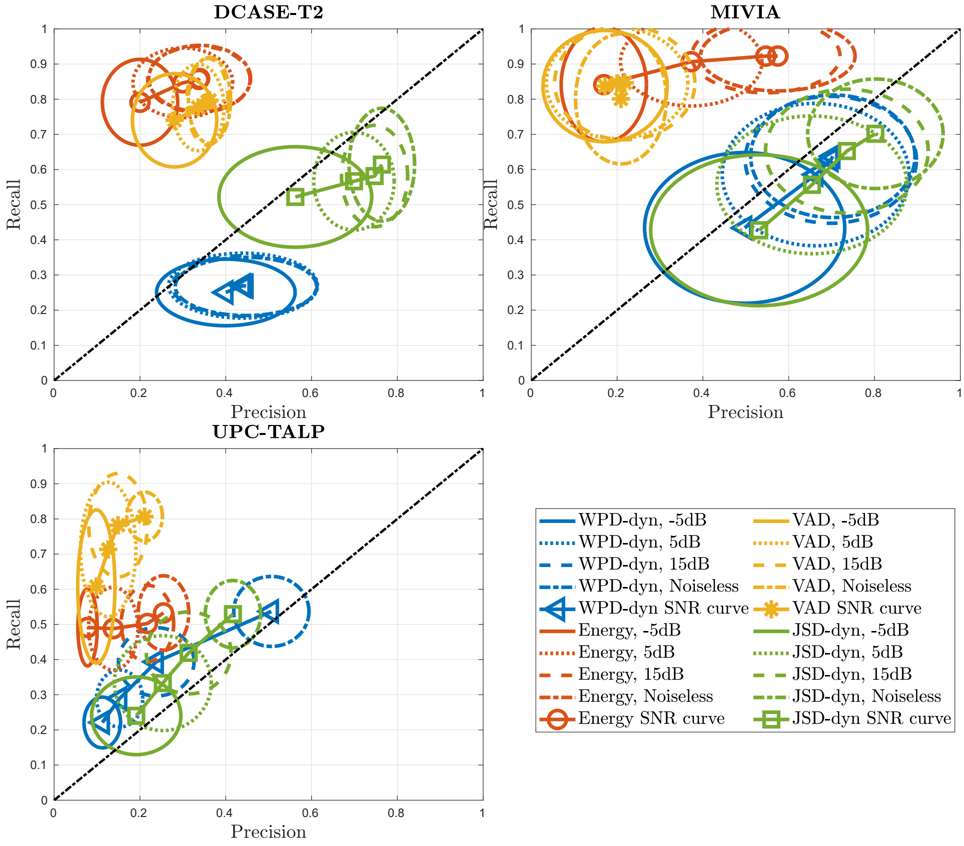


Figure 8.6: Precision-Recall scatter plots created to compare the performances of two classical detection techniques against WPD and JSD.

Chapter 9

Conclusions and future lines of work

9.1 Conclusions

This thesis is devoted to the study of techniques for auditory saliency detection, being saliency the property of auditory events or objects that makes them prominent in a particular acoustic context. We proposed two algorithms that we named Bayesian Log-surprise and Echoic Log-surprise.

In the first place, a significant challenge is related with how the performance of such algorithms should be measured due to the absence of any kind of device that could measure our target signal. Therefore, we opted for a proxy that allows us to overcome such issue. After a thorough analysis of the state-of-the-art, we concluded that there exists a clear relationship between the capacity of human beings to detect the appearance of sounds and human auditory attention, understanding the latter as the mechanism in charge of prioritizing the brain resources in order to guarantee our survival and avoid potential threats for individuals. The proxy that we chose consists in detecting onsets from audio signals, a task that in the case of a human being would be solved thanks to attention. Fortunately, there exists an area of research where this data is commonly used, named AEC/D, where onsets and offsets are used to locate relevant acoustic events in audio signals. We decided to use three representative examples of these datasets for the experiments of this thesis, namely DCASE-T2, MIVIA and UPC-TALP. However, note that we are only using AEC/D labels for testing the performance avoiding their use for training given the significant risk of overfitting to this particular task.

Therefore, for testing purposes we conformed a benchmark with some classical detection techniques, which helps to understand if our proposed algorithms are capable of performing adequately for the task of auditory saliency detection. In fact, we select some unsupervised acoustic saliency detection algorithms, such as Kayser and Kalinli models. Moreover, we consider some techniques that were originally designed to detect voice activity, such as Energy and VAD. Finally, we became aware of the existence of some algorithms that were specifically designed for the detection of onsets in music, and we decide to integrate them in our evaluation benchmark. These techniques are NWPD, WPD, SF and CD.

The algorithms that we propose in this thesis are inspired in the notion of Bayesian Surprise. An implementation of it models saliency using a circular memory buffer from which it determines statistical anomalies that are interpreted to be salient events. From this technique we develope our first approach, called Bayesian Log-surprise, since our analyses determined that the original Surprise has some drawbacks that need to be solved when the frame length of the output saliency signal is sufficiently small. In fact, this is a determining factor since salient events can occur spontaneously and with a really small duration, in the scale of milliseconds.

We introduce two modifications for Bayesian Surprise: first, our implementation opts for a cochleogram to represent spectro-temporal information instead of the well-known spectrogram, since we agreed that the former includes implicitly some perceptual concepts related to the HAS that can be useful for detection. Secondly, in order to solve the compression problem observed in the output saliency signal we consider a logarithmic operator to compress the temporal data from each frequency band, just before obtaining the global saliency signal of Bayesian Log-surprise. Our analyses demonstrate that the proposed modifications improve significantly the capabilities of the system to detect salient events in audio. Bayesian Log-surprise clearly outperforms Bayesian Surprise, and is capable of producing better detection results than the rest of the classical detection techniques with a single exception: for one of the datasets, MIVIA, Energy, a really simple technique, is capable of getting a better score than Bayesian Log-surprise. However, as we explained in Section 4.3.5 this dataset is peculiar in the sense that its silence gaps are clearly periodic and comprises only two classes of events, making it more unnatural than the two other datasets.

The main drawback of Bayesian Log-surprise is its dependence on the value of the memory length for the circular buffer. When we use short memory values, abrupt signals capable of detecting the onsets of anomalies more precisely are obtained, although noise crowds the rest of the saliency signal, whereas large memory values produce smoother signals with a small degree of distortions, where some onsets are likely to be missed.

In our second proposal, named Echoic Log-surprise, we fuse multiple Bayesian Log-surprise signals computed considering different memory lengths, in order to overcome the aforementioned issue. We propose a scheme that estimates the histograms from each Log-surprise signal and fuses their data by means of statistical divergences, which allows us to combine data from *dth* signals with different memory scales at the same time. Results suggest that the system is capable of producing better event detection scores than the rest of the classical saliency and detection techniques. Moreover, we study four different statistical divergences for the fusion stage, but no significant differences are observed between them.

However, one of the restrictions of our algorithm is related to the configuration process, since it relies on two parameters that model the depth and memory of the scales whose information is going to be fused. We observe that these two parameters differ depending on the nature of the dataset where they are tested. At the same time, our algorithm used a static threshold to determine whether the events from the saliency signal are actually salient or not. This threshold produced remarkably good results, although it requires the signal to be preprocessed and its whole temporal duration known, limiting the options to design an online saliency algorithm. In addition, we still observe that Energy is capable of outperforming our proposal for the same dataset than before, although globally speaking the results of Echoic Log-surprise are much better even in comparison with Bayesian Log-surprise. Despite these negative aspects, we consider that our technique has a lot of potential if these drawbacks are overcome.

Therefore, we introduce some modifications in its scheme. We add some specific statistical divergences to perform the fusion process. Additionally, we propose a new methodology that we denoted mixture fusion, which attempts to combine different Log-surprise scales in a more global manner instead of pairwise. However, the most noticeable addition to our algorithm is a dynamic threshold computed by means of a moving average, whose advantages are remarkable: the algorithm becomes one-pass and independent of the duration of the whole acoustic signal. Since this threshold only requires a small portion of audio to determine its value, it reduces the dispersion of the Precision and Recall values obtained for the optimal regions of operation, and performs better than any of the other detection alternatives. In addition, the choice of the control parameters gets easier, since a similar memory value is required for all of the datasets to obtain the best scores, although the depth varies for each case. Unfortunately, our mixture fusion proposal, in spite of its good performance does not provide a significant improvement in the performance of Echoic Log-surprise in comparison with our previous fusion methodologies. Finally, after comparing the results obtained for all the alternative statistical divergences employed in Echoic Log-surprise we observe once again that none of them seems to perform better than the rest, which leads us to conclude that the multi-scale fusion scheme clearly improves the detection performance, although the divergence metric choice seems to be secondary.

During our last analysis we assess if the previous results hold for noisy acoustic environments, contaminated with different signals that include white noise and recordings from common real-life scenarios, such as cafeterias, metro stations, etc. The analysis performed shows that the best classical technique might be VAD, whereas Energy shows great sensitivity to low SNR values. We also observe that most of the MIR onset detection techniques that were studied in this work perform similarly for all the SNR configurations, with the exception of NWPD which clearly underperforms. We try to determine if these MIR techniques are actually more robust to noise effects than the classical detection techniques, although our results are inconclusive since they heavily depend of the particular dataset under analysis. Regarding Echoic Log-surprise, computed using a dynamic threshold, the analysis determines that globally our technique seems to be more robust for the detection of salient events, since it produces better F-scores than any of the detection techniques for most of the SNR levels and datasets that were used, suggesting that our methodology is less sensitive to the influence of noise.

Thus we can conclude that the second saliency detection methodology that we propose in this work, Echoic Log-surprise, performs better than the rest of the techniques that were used in the evaluation benchmark, considering the results for three representative datasets of the state of the art. Moreover, our robustness analysis shows that the performance of our system holds reasonably well even in adverse noisy conditions. As a consequence, we agree that the initial objectives proposed for this thesis have been successfully fulfilled.

9.2 Future lines of work

We believe that there are still improvements that could be proposed as future lines of work. Recent progress in the area of machine learning have been massive, and novel algorithms have become a relevant part of the daily life of people. Some of the most influential algorithms are neural networks, which thanks to the overwhelming volume of data that can be acquired nowadays have shown remarkable results in tasks such as automatic speech recognition, autonomous driving, object detection, etc. We believe that research on acoustic saliency might take advantage of these advancements, since some authors demonstrated that it was possible to use pre-trained models as feature extractors, a methodology known as representation learning [Bengio et al., 2013]. Consequently, a future line of work consists in training models for a task where the volume of data is sufficiently big, such as acoustic event classification, and use their initial layers to get relevant features that characterize salient information in a more representative manner. Such features shall be used as the input data for our saliency model, or even for an updated version that takes into account other proposals used for anomaly detection. It needs to be stated that the models presented in this work are unsupervised, and might take advantage of these features without further modifications nor training process.

Moreover, one of the most limiting challenges that auditory saliency detection faces is the lack of labeled datasets, which could be helpful to design and train more sophisticated algorithms and features. We understand that some recent progress in the area of neuroscience are essential for the development of these algorithms, such as the research made by [Zhao et al., 2018] who suggested that there might exist a relationship between micro-saccadic movement and auditory saliency, which would mean that an eye-tracker could be also used to determine if an acoustic event is salient or not. Consequently, it would be possible to acquire specific and more precise data to develop computational attention models.

Finally, another future line of research that has been considered is the development of an audiovisual saliency model, which can be helpful to perform tasks such as video summarization, vigilance and driving assistance, among many others. We believe that it could be interesting to integrate our progress in auditory saliency into the structure of an audiovisual saliency model, and test it for real-life applications such as driving assistance.

Appendices

Appendix A

Echoic Log-surprise using optimal operation points

In Chapters 6 and 7 we depicted the areas of operation for Echoic Log-surprise considering both static and dynamic thresholding. These bidimensional graphs showed that the systems can perform optimally considering some close values of dth and N_1, which implies that different files from each dataset require different values for these two control parameters to achieve the best F-scores. Hence, we proposed to use such regions as weights to obtain a weighted F-score for each technique and dataset, as the results of Figure 7.4 show.

We consider that the aforementioned approach to present the results is more representative of its real behavior and ensures the stability of the assessment against inaccuracies in the configuration parameter settings. This is due to the fact that the best parameters have a bigger weight than their suboptimal counterparts when the F-score is computed. In addition, all the regions contribute to this final value which ensures that the results are robust taking into consideration all the possible scenarios. However, with this approach the weighted F-score is below its optimal value, since the contribution of suboptimal parameters deteriorates the global behavior of the results.

This is why we decided to also show the potential of the systems in terms of their optimal parameters alone, considering only the best configuration for every statistical divergence used to obtain the matrices of Figures 6.7 and 7.2. Therefore, for the sake of completeness, we present here the results obtained using the optimal parameters selected for each one of the files. We are aware that this approach is not representative of a real configuration, since parameters should be adjusted globally and remain constant for each file of every dataset. However, we consider this unrealistic scenario in order to obtain the upper-bound performance of Echoic Log-surprise.

The results produced by our algorithm are depicted in Figure A.1, which

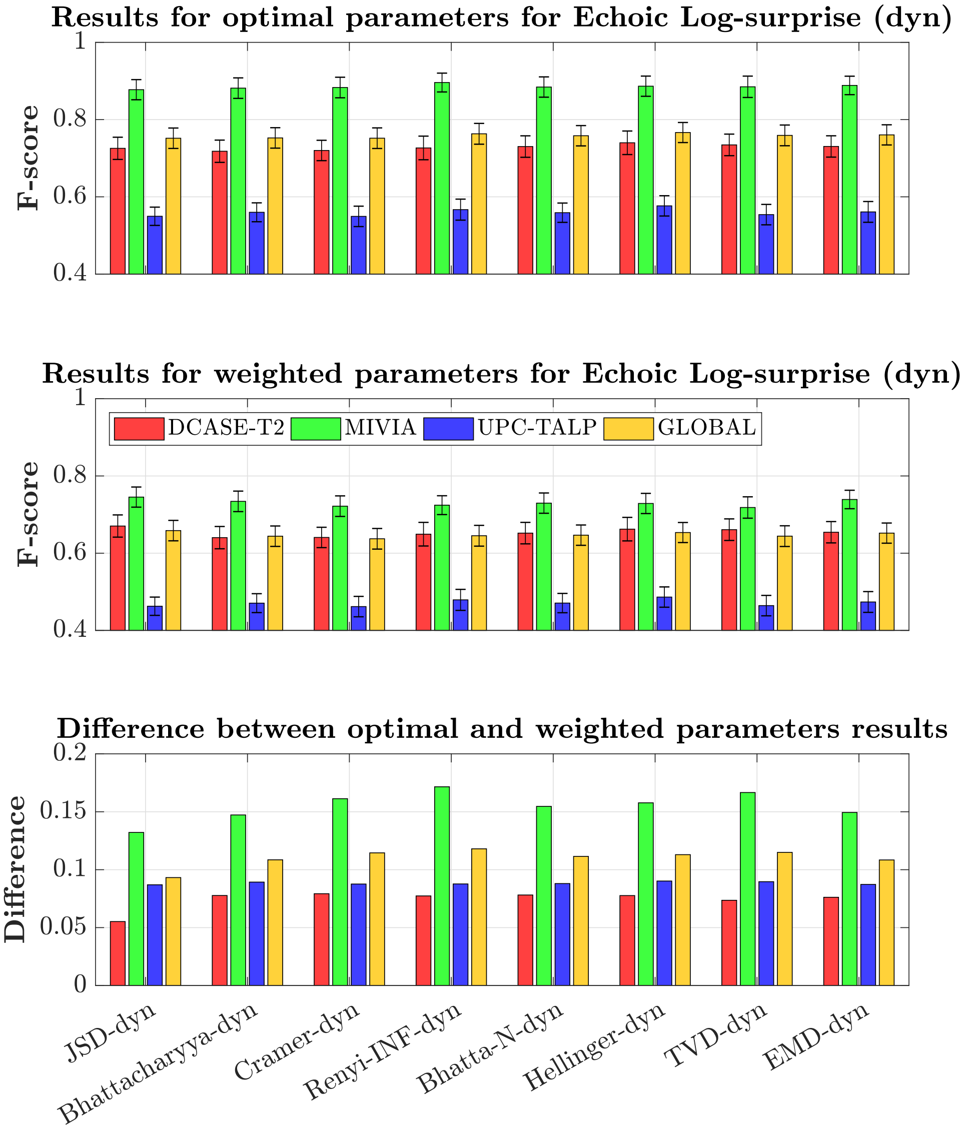


Figure A.1: Comparison of the F-score for the optimal parametric setup of Echoic Log-surprise (top) and using configuration areas (middle). The bottom diagram shows the differences between their F-scores. These results are obtained using global and local fusion strategies.

is divided into three bar diagrams. The top diagram illustrates the F-score results obtained for the optimal parameters that we propose in this appendix, which conveys the best results for each statistical divergence and file of every dataset. The second diagram summarizes the results that were included in Section 7.4, where the weighted F-score was obtained using the configuration areas for each one of the datasets. Finally, the last diagram illustrates the difference between these two configurations.

If we focus on the last diagram, the difference shows that the most remarkable improvement is observed for MIVIA, suggesting that this dataset is more sensitive to the configuration parameters than the other two datasets. The results for DCASE-T2 and UPC-TALP show less variations, which are quite independent of the divergences for UPC-TALP and remain in the range $\Delta F \in [0.08, 0.09]$, whereas for DCASE-T2 they lie within $\Delta F \in [0.05, 0.09]$.

Moreover, in terms of the global F-score no statistical divergence seems to outperform the rest according to their confidence intervals, as it occurred previously in the analysis performed in Chapter 7. However, if we focus on the results obtained for the optimal parameters the best performing technique differs from the one for weighted F-scores. In the weighted scenario, the top performing statistical divergence is JSD, whereas for the optimal parameters the best divergences are Hellinger and Renyi-INF, although their confidence intervals show that the differences with respect to JSD are not significant. Consequently, one of the conclusions that were deduced for the weighted F-scores holds for the optimal parametric setup: the selection of the statistical divergence does not seem to be very relevant, since all of them perform similarly. In addition, it is necessary to remark that although the F-scores obtained for such optimal configuration are better than the ones for the weighted F-score, they are not realistic since they are obtained using an ideal configuration where for every file and statistical divergence the best parameters are selected.

Appendix B

Spanish summary and conclusions

B.1 Motivación de la tesis

La atención se define como el mecanismo del cerebro que se encarga de categorizar la información percibida mediante los sentidos y actuar conforme a la misma dependiendo del contexto y de los recursos disponibles. La hipótesis de partida es que el cerebro humano es un órgano cuya capacidad de procesado es inmensa, y sin embargo existen escenarios donde al realizar varias tareas al mismo tiempo este es incapaz de realizar una gestión eficiente de sus recursos internos. Por ejemplo, tareas tan habituales como la conducción requieren una gran cantidad de recursos atencionales, y una tarea aparentemente tan simple como leer un mensaje de texto al volante puede derivar en un accidente de tráfico. Otro ejemplo ocurre cuando tratamos de escuchar a dos personas que están emitiendo su discurso al mismo tiempo, lo cual deriva en la frustración de notar que somos incapaces de recordar los mensajes completos emitidos por ninguno de los dos oradores. Por otra parte, podemos realizar sin dificultad tareas tan simples como ver una película, donde se perciben imágenes en movimiento y sonidos al mismo tiempo. Por lo tanto, se puede observar que en función de la naturaleza de las tareas que se pretenden realizar de forma simultánea nuestro cerebro será más o menos eficiente al gestionar sus recursos.

Existen diversas formas de clasificar la atención, si bien prestaremos atención exclusivamente a una de ellas que distingue entre las dos siguientes categorías: *top-down* y *bottom-up*. La atención *top-down* se define como aquella que va asociada a una determinada tarea que se trata de resolver, e implica que se posee un determinado conocimiento previo sobre la misma. En resumidas cuentas, lleva asociada una determinada intención. Por el contrario, la atención *bottom-up* depende exclusivamente de las características de la señal percibida y el entorno donde se ha adquirido. En contrapartida con la modalidad *top-down*, la atención *bottom-up* es automática y algunos autores la definen como un mecanismo del cerebro que garantiza nuestra supervivencia, dado que permite reaccionar de forma rápida frente a estímulos que pueden suponer una amenaza. Existen múltiples ejemplos de ambas modalidades de atención. Un caso de modalidad *top-down* se observa cuando una persona busca un objeto, por ejemplo un lápiz, en el caos de su escritorio. También podemos referirnos a atención *top-down* cuando un individuo trata de localizar un sonido muy molesto que escucha desde su despacho, pero es incapaz de determinar qué es exactamente ni de dónde procede. Por otra parte, un ejemplo de atención *bottom-up* se observa cuando en una calle muy transitada por vehículos a motor se escucha repentinamente el sonido de una ambulancia, debido a que la sirena de la misma contrasta fuertemente con el contexto acústico del oyente. Otro ejemplo de saliencia *bottom-up* se puede observar cuando en un texto escrito por ordenador se percibe una sección marcada en rojo, cuyo color contrasta fuertemente con el blanco y negro del resto del documento.

La atención *bottom-up* es habitualmente conocida como saliencia, y se puede definir como una característica de las señales que percibimos a través de nuestros sentidos, la cual representa la prominencia de la información que se ha adquirido del entorno. Así, para el ejemplo anteriormente expuesto sobre la ambulancia, el contexto acústico estaría plagado por los sonidos de tráfico tan habituales en las ciudades. Sin embargo, el sonido de una sirena sería prominente al contrastar fuertemente con el tráfico de fondo.

Este trabajo está relacionado con el concepto de saliencia y su detección automática mediante algoritmos. En los últimos años se han producido grandes avances en la detección de saliencia visual, la cual pretende determinar qué objetos de una determinada escena captan la atención de un espectador. Este progreso se debe a dos grandes grupos de contribuciones: por una parte, gracias a nuevos algoritmos que proporcionan mapas cuyas detecciones se asemejan cada vez más a las de espectadores humanos, y por otra gracias a las numerosas bases de datos existentes para entrenar dichos modelos. Dichas bases de datos se adquieren mediante un dispositivo denominado *eye-tracker*. Tal y como su nombre indica, este dispositivo se encarga de medir la trayectoria seguida por los ojos de un participante humano mientras este visualiza un determinado vídeo o imagen. Los datos extraídos a partir de múltiples participantes permiten generar un gran volumen de información, la cual puede utilizarse para entrenar modelos de atención visual de gran calidad.

Sin embargo, los avances en otras modalidades de saliencia han resultado ser menos fructíferos. Tal es el caso de la saliencia auditiva, donde si bien es cierto que existen algunos algoritmos que se encargan de realizar su detección, no queda claro cuál de los mismos produce los resultados más fidedignos. Esto se debe a la total ausencia de bases de datos etiquetadas con dicha información. Es habitual que los investigadores produzcan sus propias bases de datos utilizando a un determinado número de participantes, cuyo comportamiento tratan de emular los algoritmos propuestos. Sin embargo, no existe un dispositivo similar al *eye-tracker* que permita determinar de forma precisa qué elementos de una escena acústica son más prominentes.

B.2 Contribuciones del trabajo

En este trabajo proponemos dos modelos para la detección de la saliencia auditiva. Tal y como acabamos de exponer, uno de los principales obstáculos que deben afrontar este tipo de algoritmos está relacionado con la carencia total de datos para realizar las mediciones pertinentes de rendimiento. En consecuencia, se debe establecer una metodología que permita determinar si un modelo de saliencia auditiva se comporta mejor que otras alternativas del estado del arte. Para ello, planteamos una hipótesis de partida que tiene que ver con la capacidad del ser humano para detectar la aparición y desaparición de sonidos en un determinado entorno o contexto acústico. Existen numerosos estudios previos que han determinado que los seres humanos son más sensibles a la aparición de eventos acústicos, si bien esta ocurrencia no se replica con tal magnitud para la desaparición de los mismos. En consecuencia, la hipótesis de partida de este trabajo consiste en aceptar que aquellos eventos acústicos que aparecen repentinamente poseen la etiqueta de salientes frente al entorno acústico. La principal ventaja de dicha hipótesis radica en el hecho de que existen múltiples bases de datos diseñadas específicamente para la detección y clasificación de eventos acústicos. Esto es, están formadas por sonidos que aparecen espontáneamente, de los cuales se conocen sus instantes de aparición y desaparición, así como los sonidos que representan. Por lo tanto, evaluaremos nuestros sistemas de detección de saliencia y otros pertenecientes al ámbito utilizando dichas bases de datos, y estableceremos que el instante de aparición de un evento acústico, conocido técnicamente como *onset*, determinará la ocurrencia de un evento saliente.

Una vez se ha establecido el procedimiento objetivo para comparar las capacidades de diversos sistemas de detección de saliencia auditiva, se procederá a diseñar múltiples experimentos que permitan determinar cómo se comportan en las situaciones más adversas posibles. La comparativa se realiza con respecto a otras técnicas de saliencia auditiva previamente implementadas: los modelos de *Kayser* y *Kalinli*, respectivamente. Al mismo tiempo, proponemos la utilización de técnicas empleadas para la detección automática de habla en tramos de señal acústica, por lo que incluimos el *Voice Activity Detector* (VAD) propuesto por [Sohn et al., 1999] así como un detector basado en un umbral energético, que denominamos *Energy*. Por último, proponemos utilizar técnicas de detección de *onsets* en señales musicales, entre las cuales destacamos NWPD, WPD, SF y CD.

Al mismo tiempo, planteamos la utilización de tres bases de datos de detección y clasificación de eventos acústicos ampliamente conocidas en el estado del arte: *DCASE 2016 (Task 2)*, *MIVIA road audio events* y *UPC-TALP*, contando así con más de 3400 eventos acústicos etiquetados.

Nuestro primer algoritmo se inspira en una implementación previa denominada *Bayesian Surprise*, aplicable tanto para la detección de saliencia visual como auditiva. Para el caso auditivo, dicha metodología mide la saliencia modelando la información de instantes consecutivos de tiempo mediante distribuciones Normales de probabilidad, las cuales se comparan mediante la utilización de la divergencia de *Kullback-Leibler*. Dicha divergencia proporciona un valor de disimilitud, donde aquellos valores supuestamente más distantes son los que representarían una prominencia o valor saliente en la señal acústica. Esto es, al medirse la divergencia entre instantes consecutivos lo que se consigue es determinar si ha aparecido algún tipo de patrón acústico anómalo en comparación con el entorno conocido hasta el momento. La señal de saliencia de salida es umbralizada mediante un algoritmo estático, que producirá una señal binaria indicando si los eventos son salientes o no.

Tal y como se ha indicado previamente, se modela el contenido acústico de la señal mediante sendas distribuciones Normales, las cuales representan respectivamente la información pasada (probabilidad *a priori*) y la información actual (probabilidad *a posteriori*), cuyas medias y varianzas se determinan utilizando ventanas (o *buffers*) de una determinada longitud N. Ventanas de mayor longitud implicarían predicciones más suaves donde las prominencias podrían llegar a pasar desapercibidas, mientras que ventanas menores producirían señales de saliencia más ruidosas, pero al mismo tiempo más sensibles y capaces de detectar anomalías con una mayor precisión temporal.

Sin embargo, en uno de nuestros trabajos previos determinamos que *Bayesian Surprise* produce señales de saliencia con niveles de compresión inaceptables entre los distintos picos salientes. Esto es, los eventos más salientes poseen una magnitud tan grande que impiden la visualización y detección de aquellos eventos salientes de menor magnitud. Nuestra primera propuesta consiste en aplicar el operador logarítmico para provocar que eventos salientes con mayor y menor magnitud pasen a tener valores similares, de forma similar a como operan la *A-law* y la *μ-law* propuestas en ITU-T G.711. Denominamos esta técnica *Bayesian Log-surprise.* Los resultados muestran que al comparar *Log-surprise* frente a *Surprise* y al resto de técnicas de detección que planteamos para este trabajo, *Log-surprise* es capaz de producir las mejores puntuaciones de detección. Sin embargo, detectamos que para una de las bases de datos su rendimiento queda ensombrecido por *Energy.*

En consecuencia, proponemos un nuevo algoritmo inspirado en dicha técnica, el cual calcula un número dth de señales utilizando *Bayesian Log-*

surprise con *buffers* de distinta longitud N_z. Esto es, se calculan varias señales mediante *Bayesian Log-surprise* utilizando *buffers* con longitudes $[N_1, N_2, \cdots, N_{dth}]$, lo que equivale a contar con diversas escalas temporales de saliencia auditiva. Seguidamente se estiman sus distribuciones mediante sendos histogramas, los cuales son fusionados utilizando divergencias estadísticas. Denominamos a esta técnica *Echoic Log-surprise*. Las señales producidas están caracterizadas por una serie de ventajas:

- Deja de observarse el ruido de fondo característico de las señales de *Bayesian Log-surprise*, ya que queda caracterizado por los histogramas y atenuado por las divergencias.
- El sistema detecta las prominencias de las señales acústicas en distintas escalas, lo cual favorece su detección.
- Las puntuaciones de rendimiento muestran mejorías muy significativas al aplicar este nuevo algoritmo.

Sin embargo, también detectamos una serie de inconvenientes:

- Dado que trabajamos con un modelo multi-escala es necesario fijar un mecanismo que facilite la validación del tamaño del *buffer* de cada una de las escalas, así como el número de las mismas.
- El algoritmo para umbralizar que se utiliza sigue siendo estático, lo cual impide medir la saliencia *online*.

Es por ello que planteamos una serie de mejoras, incluyendo una nueva estrategia de fusión, nuevas divergencias estadísticas y la implementación de un umbralizador dinámico, siendo esta última modificación la que produjo la mejora más significativa en los resultados.

Finalizamos este trabajo replicando los experimentos anteriores tras contaminar las señales disponibles en las tres bases de datos utilizando ruido de diversa índole: estacionario, donde consideramos el ruido blanco Gaussiano, y no estacionario, donde contaminamos las señales acústicas mediante audios grabados en entornos acústicos diversos, tales como estaciones de tren, cafeterías, parques, etc. Contaminamos las señales utilizando seis valores de SNR, desde -5 dB hasta 20 dB, parámetro que compara la magnitud de la señal original frente a la magnitud de la señal de ruido.

Los resultados obtenidos nos permiten determinar la robustez frente al ruido de todas las técnicas que se han evaluado en este trabajo, tanto aquellas de detección de habla u *onsets* musicales como los distintos algoritmos de saliencia auditiva. Tal y como era previsible, las puntuaciones son diversas en función de la naturaleza de cada uno de los algoritmos. Sin embargo, con respecto a *Bayesian Log-surprise* determinamos que es más robusta que VAD, *Energy* y *Kalinli*. Por otra parte, a nivel global observamos que *Echoic*

Log-surprise es más robusta frente a ruido que cualquiera de los algoritmos estudiados en este trabajo, ya que produce las mejores puntuaciones de detección para todas las posibles configuraciones de ruido.

B.3 Conclusiones

Esta tesis se dedica al estudio de las técnicas de detección de saliencia acústica, siendo esta una propiedad de los eventos u objetos acústicos que los hace prominentes en un determinado contexto. Contribuimos con dos algoritmos que denominamos *Bayesian Log-surprise* y *Echoic Log-surprise*, ambos inspirados en la noción de *Bayesian Surprise*. Esta metodología modela la saliencia mediante un *buffer* circular a partir del cual detecta anomalías estadísticas, las cuales son interpretadas como eventos salientes. A partir de la misma desarrollamos nuestra primera propuesta, denominada *Bayesian Log-surprise*, dado que nuestros análisis determinaron que *Surprise* poseía algunos inconvenientes destacables. Las mejoras introducidas en *Bayesian Log-surprise* son, primero, la utilización del cocleograma como representación espectro-temporal de la señal acústica, dado que de esta forma la información espectral contenida incluye un procesamiento por bandas de frecuencia similar al del sistema auditivo humano (HAS). La segunda mejora consiste en utilizar el operador logarítmico para comprimir las bandas de frecuencia de la señal de saliencia obtenida al aplicar *Bayesian Surprise*. Nuestros análisis demostraron que dichas modificaciones mejoraron significativamente las capacidades de detección del sistema respecto al resto de algoritmos de detección que forman parte del *benchmark*.

En nuestra segunda propuesta, denominada *Echoic Log-surprise*, fusionamos múltiples señales de saliencia obtenidas a partir de *Bayesian Log-surprise* considerando diferentes tamaños de memoria. Los resultados sugieren que el sistema es capaz de producir mejores detecciones que el resto de técnicas clásicas de saliencia y detección. Además, las cuatro divergencias estadísticas empleadas en la etapa de fusión mostraron resultados muy similares.

Decidimos incluir algunas modificaciones en nuestro esquema: añadimos cuatro divergencias estadísticas adicionales para llevar a cabo el proceso de fusión. Además, incluimos una nueva metodología para llevar a cabo dicha labor de forma global, la cual denominamos *mixture fusion*. Sin embargo, la mejora más significativa es un algoritmo de umbralización dinámico obtenido mediante una media móvil, el cual permite a *Echoic Log-surprise* operar *online*. Dicho umbralizador redujo la dispersión en *Precision* y *Recall* para las regiones óptimas de funcionamiento, a la par que el algoritmo mantuvo un rendimiento adecuado y superior al del resto de propuestas de detección. Al mismo tiempo, facilitó la elección de los parámetros de control del sistema. El resto de mejoras propuestas no supuso ninguna variación significativa respecto a la versión inicial del algoritmo.

En el último análisis que propusimos verificamos si los resultados anteriores se mantenían en entornos acústicos ruidosos, contaminados con señales entre las cuales incluimos ruido blanco y grabaciones de escenarios reales, tales como cafeterías, estaciones de metro, etc. El análisis mostró que los mejores resultados para las técnicas clásicas fueron producidos por VAD, mientras que *Energy* mostró una gran sensibilidad frente al ruido. También observamos que las técnicas de MIR se comportaron de forma similar. Respecto a *Echoic Log-surprise*, obtenido utilizando el umbralizador dinámico, los resultados mostraron que a nivel global e individual dicha metodología resultó ser la más robusta para la detección de eventos salientes, dado que produjo los mejores F-scores para cualquier valor de SNR, sugiriendo que nuestra propuesta es la más insensible frente a los efectos perjudiciales del ruido.

Así, concluimos que la segunda propuesta para la detección de saliencia que realizamos en este trabajo, *Echoic Log-surprise*, mostró un rendimiento superior al resto de técnicas analizadas considerando tres bases de datos representativas. Al mismo tiempo, nuestro análisis de robustez mostró que dicha metodología también resultó ser la menos sensible a la influencia del ruido. Por lo tanto, confirmamos que los objetivos iniciales propuestos para esta tesis doctoral quedan satisfactoriamente cumplidos.

B.4 Líneas futuras de trabajo

Existen múltiples líneas en las cuales se puede trabajar para mejorar este proyecto. El área del aprendizaje máquina ha avanzado enormemente en los últimos años, y especialmente la utilización y desarrollo de redes neuronales artificiales. Estos algoritmos se encuentran en nuestro día a día en tareas como el reconocimiento automático del habla, la conducción autónoma, la detección de objetos, etc. Pensamos que la saliencia auditiva puede aprovechar algunos de estos avances, dado que algunos autores han demostrado que dichas estructuras pueden emplearse a modo de extractores de características [Bengio et al., 2013]. En consecuencia, una línea inmediata de investigación consistiría en entrenar modelos para la tarea de clasificación de eventos acústicos, y emplearlos a modo de extractores de características. Estas serían introducidas en modelos de detección de saliencia acústica, los cuales consideramos que podrían aprovechar dicha información para mejorar la detección de eventos prominentes.

Por otra parte, uno de los desafíos que afronta la detección de saliencia auditiva tiene que ver con la carencia de bases de datos específicas etiquetadas para tal labor. [Zhao et al., 2018] sugiere en uno de sus trabajos que existe una relación entre el movimiento micro-sacádico de los ojos y la saliencia acústica, por lo que pretendemos utilizar *eye-trackers* para generar nuevas bases de datos que permitan desarrollar mejores algoritmos de detección automática.

Por último, otra línea de investigación que proponemos consiste en desarrollar un modelo de saliencia audiovisual, con el cual pretendemos mejorar la realización de tareas tales y como son la sumarización de vídeos, videovigilancia y asistencia durante la conducción.

Bibliography

[Abka and Pardede, 2015] Abka, A. F. and Pardede, H. F. (2015). Speech recognition features: Comparison studies on robustness against environmental distortions. In *2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 114–119, Bandung, Indonesia.

[Anderson, 2015] Anderson, J. (2015). *Cognitive Psychology and Its Implications.* Worth Publishers, New York, United States.

[Auvray et al., 2008] Auvray, M., Gallace, A., Hartcher-O'Brien, J., Tan, H. Z., and Spence, C. (2008). Tactile and visual distractors induce change blindness for tactile stimuli presented on the fingertips. *Brain Research*, 1213:111–119.

[Barascud et al., 2014] Barascud, N., Griffiths, T. D., McAlpine, D., and Chait, M. (2014). "Change Deafness" Arising from Inter-feature Masking within a Single Auditory Object. *Journal of Cognitive Neuroscience*, 26(3):514–528.

[Bello et al., 2005] Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. B. (2005). A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

[Beran, 1977] Beran, R. (1977). Minimum Hellinger Distance Estimates for Parametric Models. *The Annals of Statistics*, 5(3):445–463.

[Bhattacharyya, 1943] Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109.

[Blausen, 2014] Blausen (2014). The Anatomy of the Ear. Medical gallery of Blausen Medical 2014. *WikiJournal of Medicine*, 1(2). [Online; accessed 17/01/2019].

[Böck et al., 2012] Böck, S., Krebs, F., and Schedl, M. (2012). Evaluating the Online Capabilities of Onset Detection Methods. In *International Society for Music Information Retrieval Conference*, Porto, Portugal.

[Borji et al., 2019] Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., and Li, J. (2019). Salient Object Detection: A Survey. *Computational Visual Media*, 5(2):117–150.

[Borji and Itti, 2013] Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207.

[Bottcher-Gandor and Ullsperger, 1992] Bottcher-Gandor, C. and Ullsperger, P. (1992). Mismatch Negativity in Event-Related Potentials to Auditory Stimuli as a Function of Varying Interstimulus Interval. *Psychophysiology*, 29:546–550.

[Broadbent, 1954] Broadbent, D. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, 47(3):191–196.

[Brookes, 1997] Brookes, M. (1997). VOICEBOX: Speech processing toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[Bruya and Tang, 2018] Bruya, B. and Tang, Y. Y. (2018). Is attention really effort? Revisiting Daniel Kahneman's Influential 1973 Book Attention and Effort. *Frontiers in Psychology*, 9(1133).

[Bylinskii et al., 2019a] Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., and Torralba, A. (Accessed 15/01/2019a). MIT Saliency Benchmark. http://saliency.mit.edu/results_mit300.html.

[Bylinskii et al., 2019b] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2019b). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757.

[Cervantes Constantino et al., 2012] Cervantes Constantino, F., Pinggera, L., Paranamana, S., Kashino, M., and Chait, M. (2012). Detection of Appearing and Disappearing Objects in Complex Acoustic Scenes. *PLOS ONE*, 7(9):1–13.

[Cherry, 1953] Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5):975–979.

[Coutrot and Guyader, 2015] Coutrot, A. and Guyader, N. (2015). An efficient audiovisual saliency model to predict eye positions when looking at conversations. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1531–1535, Nice, France.

[Cowan, 1984] Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, 96(2):341–370.

[Cowan, 2008] Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In *Essence of Memory*, volume 169 of *Progress in Brain Research*, pages 323 – 338. Elsevier.

[Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

[Deutsch and Deutsch, 1963] Deutsch, J. and Deutsch, D. (1963). Attention: Some Theoretical Considerations. *Psychological Review*, 70(1):80–90.

[Dixon, 2006] Dixon, S. (2006). Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects*, pages 133–137, Montreal, Canada.

[Drugman et al., 2015] Drugman, T., Stylianou, Y., Chen, L., Chen, X., and Gales, M. J. F. (2015). Robust excitation-based features for Automatic Speech Recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4664–4668, Brisbane, Australia.

[Duxbury et al., 2003] Duxbury, C., Bello, J. P., Davies, M., and Sandler, M. (2003). Complex Domain Onset Detection for Musical Signals. In *Proceedings of the 6th International Conference on Digital Audio Effects*, pages 1–4, London, UK.

[Encyclopædia Britannica, 2019] Encyclopædia Britannica (2019). Inner ear. https://www.britannica.com/science/inner-ear/images-videos. [Online; accessed 30/05/2019].

[Endres and Schindelin, 2003] Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

[Eramudugolla et al., 2005] Eramudugolla, R., Irvine, D. R., McAnally, K. I., Martin, R. L., and Mattingley, J. B. (2005). Directed attention eliminates 'change deafness' in complex auditory scenes. *Current Biology*, 15(12):1108–1113.

[Evangelopoulos et al., 2013] Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., and Avrithis, Y. (2013). Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

[Fastl and Zwicker, 2007] Fastl, H. and Zwicker, E. (2007). *Psychoacoustics: Facts and Models*. Springer-Verlag, Berlin, Heidelberg.

[Fernández-Torres et al., 2016] Fernández-Torres, M., González-Díaz, I., and Díaz-de-María, F. (2016). A probabilistic topic approach for context-aware visual attention modeling. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6.

[Fernández-Torres et al., 2019] Fernández-Torres, M., González-Díaz, I., and Díaz-de-María, F. (2019). Probabilistic Topic Model for Context-Driven Visual Attention Understanding. *IEEE Transactions on Circuits and Systems for Video Technology*.

[Fletcher, 1940] Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, 12(1):47–65.

[Fletcher and Munson, 1933] Fletcher, H. and Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *The Bell System Technical Journal*, 12(4):377–430.

[Foggia et al., 2015] Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., and Vento, M. (2015). Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, 65:22–28.

[Foggia et al., 2016] Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., and Vento, M. (2016). Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):279–288.

[Foggia et al., 2014] Foggia, P., Saggese, A., Strisciuglio, N., and Vento, M. (2014). Cascade classifiers trained on gammatonegrams for reliably detecting audio events. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 50–55, Seoul, South Korea.

[Forster and Spence, 2018] Forster, S. and Spence, C. (2018). “What Smell?” Temporarily Loading Visual Attention Induces a Prolonged Loss of Olfactory Awareness. *Psychological Science*, 29(10):1642–1652.

[Gallace et al., 2007] Gallace, A., Tan, H. Z., and Spence, C. (2007). Do "mudsplashes" induce tactile change blindness? *Perception & Psychophysics*, 69(4):477–486.

[Glasberg and Moore, 1990] Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1):103 – 138.

[Glass et al., 2008] Glass, E., Sachse, S., and von Suchodoletz, W. (2008). Development of auditory sensory memory from 2 to 6 years: an MMN study. *Journal of Neural Transmission*, 115(8):1221–1229.

[Goldstein, 2009] Goldstein, E. B. (2009). *Sensation and Perception*. Wadsworth, Cengage Learning, 8th edition.

[Gomes et al., 1999] Gomes, H., Sussman, E., Ritter, W., Kurtzberg, D., Cowan, N., and Vaughan, H. J. (1999). Electrophysiological evidence of developmental changes in the duration of auditory sensory memory. *Developmental Psychology*, 35(1):294–302.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Harel, 2012] Harel, J. (2012). Saliency map algorithm : Matlab source code. Accessed 15/01/2019. `http://www.vision.caltech.edu/~harel/share/gbvs.php`.

[Harel et al., 2006] Harel, J., Koch, C., and Perona, P. (2006). Graph-Based Visual Saliency. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 545–552, Vancouver, Canada.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

[Ishida and Matsuura, 2001] Ishida, T. and Matsuura, T. (2001). The effect of cellular phone use on driving performance. *IATSS Research*, 25(2):6 – 14.

[Itti and Baldi, 2005] Itti, L. and Baldi, P. F. (2005). A principled approach to detecting surprising events in video. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 631–637, San Diego, CA.

[Itti and Baldi, 2009] Itti, L. and Baldi, P. F. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295 – 1306.

[Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

[ITU-T, 1988] ITU-T (1988). *ITU-T Recommendation G.711: Pulse code modulation (PCM) of voice frequencies*. International Telecommunication Union.

[ITU-T, 1994] ITU-T (1994). Objective measurement of active speech level. Technical Report p.56 (12/11), International Telecommunication Union.

[Jia, 2018] Jia, S. (2018). EML-NET: an expandable multi-layer network for saliency prediction. *CoRR*, abs/1805.01047.

[Kahneman, 1973] Kahneman, D. (1973). *Attention and effort*. Prentice-Hall series in experimental psychology. Prentice-Hall, New Jersey.

[Kalinli and Narayanan, 2007] Kalinli, O. and Narayanan, S. (2007). A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech. In *INTERSPEECH-2007. 8th Annual Conference of the International Speech Communication Association*, pages 1941–1944, Antwerp, Belgium.

[Kalinli and Narayanan, 2009] Kalinli, O. and Narayanan, S. (2009). Prominence detection using auditory attention cues and task-dependent high level information. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):1009–1024.

[Kang and Wildes, 2015] Kang, S. M. and Wildes, R. P. (2015). The n-Distribution Bhattacharyya Coefficient. Technical Report EECS-2015-02, University of York.

[Katsuki and Constantinidis, 2014] Katsuki, F. and Constantinidis, C. (2014). Bottom-up and top-down attention: Different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521.

[Kaya and Elhilali, 2014] Kaya, E. M. and Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, 8(327):1–12.

[Kayser, 2018] Kayser, C. (Accessed 1/10/2018). Auditory saliency map. `http://uni-bielefeld.de/biologie/cns/resources.html`.

[Kayser et al., 2005] Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947.

[Kim and Stern, 2010] Kim, C. and Stern, R. M. (2010). Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4574–4577, Dallas, United States.

[Kruthiventi et al., 2017] Kruthiventi, S. S. S., Ayush, K., and Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456.

[Kümmerer et al., 2017] Kümmerer, M., Wallis, T. S. A., Gatys, L. A., and Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808, Venice, Italy.

[Levin et al., 2009] Levin, D. A., Peres, Y., and Wilmer, E. L. (2009). Markov chains and mixing times. *American Mathematical Society*, page 76.

[Li and Gao, 2014] Li, J. and Gao, W. (2014). *Visual saliency computation: A machine learning perspective*, volume LNCS 8408. Springer, Cham.

[Macaluso, 2018] Macaluso, E. (Accessed 1/10/2018). MT_TOOLS: Computation of saliency and feature-specific maps for visual and auditory stimuli. http://www.brainreality.eu/mt_tools/.

[Martinez et al., 2016] Martinez, M., Tapaswi, M., and Stiefelhagen, R. (2016). A Closed-form Gradient for the 1D Earth Mover's Distance for Spectral Deep Learning on Biological Data . In *ICML 2016 Workshop on Computational Biology*, New York City, United States.

[Mazza et al., 2007] Mazza, V., Turatto, M., Rossi, M., and Umiltà, C. (2007). How automatic are audiovisual links in exogenous spatial attention? *Neuropsychologia*, 45(3):514 – 522.

[Menon, 2015] Menon, V. (2015). Salience network. In *Brain Mapping*, pages 597 – 611. Academic Press, Waltham.

[Merriam-Webster.com, 2019] Merriam-Webster.com (2019). Sense. [Online; accessed 30/05/2019].

[Mesaros et al., 2016a] Mesaros, A., Heittola, T., and Virtanen, T. (2016a). Metrics for polyphonic sound event detection. *Applied Sciences*, 6(162):1–17.

[Mesaros et al., 2018] Mesaros, A., Heittola, T., and Virtanen, T. (2018). A multi-device dataset for urban acoustic scene classification. In *DCASE2018 Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 1–5, Surrey, UK.

[Mesaros et al., 2016b] Mesaros, A., Heittola, T., Virtanen, T., Benetos, E., Foster, P., Lagrange, M., Lafay, G., and Plumbley, M. D. (2016b). IEEE AASP challenge: Detection and classification of acoustic scenes and events 2016. http://www.cs.tut.fi/sgn/arg/dcase2016/.

[Moray, 1959] Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11(1):56–60.

[Nikulin, 2001] Nikulin, M. S. (2001). *Hellinger distance*. Encyclopedia of mathematics. Springer Berlin Heidelberg and New York.

[Okamoto et al., 2007] Okamoto, H., Kakigi, R., Gunji, A., and Pantev, C. (2007). Asymmetric lateral inhibitory neural activity in the auditory system: a magnetoencephalographic study. *BMC Neuroscience*, 8(33):1–6.

[Oppenheim et al., 1999] Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (1999). *Discrete-time Signal Processing (2nd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Pavani and Turatto, 2008] Pavani, F. and Turatto, M. (2008). Change perception in complex auditory scenes. *Perception & Psychophysics*, 70(4):619–629.

[Pearce and Hirsch, 2000] Pearce, D. and Hirsch, H.-G. (2000). The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In *6th International Conference on Spoken Language Processing, ICSLP 2000*, Beijing, China.

[Pinto et al., 2013] Pinto, Y., van der Leij, A. R., Sligte, I. G., Lamme, V. A. F., and Scholte, H. S. (2013). Bottom-up and top-down attention are independent. *Journal of Vision*, 13(3):1–14.

[Proakis, 2001] Proakis, J. (2001). *Digital Communications*. Electrical engineering series. McGraw-Hill.

[Rabin et al., 2008] Rabin, J., Delon, J., and Gousseau, Y. (2008). Circular Earth Mover's Distance for the comparison of local features. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Tampa, USA.

[Ramirez et al., 2007] Ramirez, J., Segura, J., and Gorriz, J. (2007). Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In *Robust Speech*. IntechOpen, Rijeka.

[Rensink et al., 1997] Rensink, R. A., O'Regan, J. K., and Clark, J. J. (1997). To See or not to see: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5):368–373.

[Rodríguez-Hidalgo et al., 2016] Rodríguez-Hidalgo, A., Gallardo-Antolín, A., and Peláez-Moreno, C. (2016). Towards aural saliency detection with logarithmic Bayesian Surprise under different spectro-temporal representations. In *Proceedings of Iberspeech 2016*, pages 99–108, Lisbon, Portugal.

[Rodríguez-Hidalgo et al., 2018a] Rodríguez-Hidalgo, A., Peláez-Moreno, C., and Gallardo-Antolín, A. (2018a). Echoic log-surprise: A multi-scale scheme for acoustic saliency detection. *Expert Systems with Applications*, 114:255–266.

[Rodríguez-Hidalgo et al., 2018b] Rodríguez-Hidalgo, A., Peláez-Moreno, C., and Gallardo-Antolín, A. (2018b). The robustness of echoic log-surprise auditory saliency detection. *IEEE Access*, 6:72083–72093.

[Rodríguez-Hidalgo et al., 2019] Rodríguez-Hidalgo, A., Peláez-Moreno, C., and Gallardo-Antolín, A. (2019). Auditory saliency detection based on information fusion by means of statistical divergences. *Manuscript submitted for publication.*

[Rosão et al., 2012] Rosão, C., Ribeiro, R., and Martins de Matos, D. (2012). Influence of peak selection methods on onset detection. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012*, pages 517–522, Porto, Portugal.

[Rubner et al., 2000] Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121.

[Sams et al., 1993] Sams, M., Hari, R., Rif, J., and Knuutila, J. (1993). The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence. *Journal of Cognitive Neuroscience*, 5(3):363–370.

[Schauerte, 2013] Schauerte, B. (2013). Gaussian Surprise and Running Windowed Mean/Variance. https://es.mathworks.com/matlabcentral/fileexchange/33573-gaussian-surprise-and-running-windowed-mean-variance.

[Schauerte et al., 2011] Schauerte, B., Kühn, B., Kroschel, K., and Stiefelhagen, R. (2011). Multimodal saliency-based attention for object-based scene analysis. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1173–1179, San Francisco, United States.

[Schauerte and Stiefelhagen, 2013] Schauerte, B. and Stiefelhagen, R. (2013). "Wow!" Bayesian surprise for salient acoustic event detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6402–6406, Vancouver, Canada.

[Schreiner et al., 2000] Schreiner, C. E., Read, H. L., and Sutter, M. L. (2000). Modular Organization of Frequency Integration in Primary Auditory Cortex. *Annual Review of Neuroscience*, 23(1):501–529.

[Sela and Sobel, 2010] Sela, L. and Sobel, N. (2010). Human olfaction: A constant state of change-blindness. *Experimental Brain Research*, 205(1):13–29.

[Seltzer et al., 2013] Seltzer, M. L., Yu, D., and Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7398–7402, Vancouver, Canada.

[Serizel et al., 2018] Serizel, R., Turpault, N., Eghbal-Zadeh, H., and Parag Shah, A. (2018). Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments. In *DCASE2018 Workshop on Detection and Classification of Acoustic Scenes and Events*, pages 1–5, Surrey, UK.

[Shamma, 2001] Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Sciences*, 5(8):340–348.

[Slaney, 1993] Slaney, M. (1993). An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Technical Report 35, Apple Computer, Inc.

[Smith, 2007] Smith, J. O. (2007). *Mathematics of the Discrete Fourier Transform (DFT)*. https://ccrma.stanford.edu/~jos/st/. Online book, 2007 edition.

[Sohn et al., 1999] Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3.

[Stevens et al., 1937] Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

[Székely, 2002] Székely, G. J. (2002). E-statistics: The energy of statistical samples. Technical Report No. 02-16, Bowling Green State University.

[Temko et al., 2007] Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., and Omologo, M. (2007). CLEAR Evaluation of Acoustic Event Detection and Classification Systems. *LNCS 4122*, pages 311–322.

[Thiemann et al., 2013] Thiemann, J., Ito, N., and Vincent, E. (2013). DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.1227121.

[Treisman, 1960] Treisman, A. M. (1960). Contextual Cues in Selective Listening. *Quarterly Journal of Experimental Psychology*, 12(4):242–248.

[Treisman, 1964a] Treisman, A. M. (1964a). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, 3(6):449 – 459.

[Treisman, 1964b] Treisman, A. M. (1964b). Selective Attention in Man. *British Medical Bulletin*, 20(1):12–16.

[Tsuchida and Cottrell, 2012] Tsuchida, T. and Cottrell, G. (2012). Auditory saliency using natural statistics. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, pages 1048–1053, Sapporo, Japan.

[Underwood, 1974] Underwood, G. (1974). Moray vs. the rest: the effects of extended shadowing practice. *The Quarterly journal of experimental psychology*, 26(3):368–372.

[van Erven and Harremoës, 2014] van Erven, T. and Harremoës, P. (2014). Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

[Varga and Steeneken, 1993] Varga, A. and Steeneken, H. J. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251.

[Vincent et al., 2017] Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J., and Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557.

[Vitevitch, 2003] Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2):333–342.

[Waibel and Stiefelhagen, 2009] Waibel, A. and Stiefelhagen, R. (2009). *Computers in the Human Interaction Loop*. Springer, London.

[Welford, 1962] Welford, B. P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420.

[Xu et al., 2005] Xu, M., Duan, L.-Y., Cai, J., Chia, L.-T., Xu, C., and Tian, Q. (2005). HMM-Based Audio Keyword Generation. In *Advances in Multimedia Information Processing - PCM 2004*, pages 566–574. Springer Berlin Heidelberg.

[Zhang et al., 2008] Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20.

[Zhao et al., 2018] Zhao, S., Yum, N. W., Benjamin, L., Benhamou, E., Furukawa, S., Dick, F., Slaney, M., and Chait, M. (2018). Rapid ocular responses are a robust marker for bottom-up driven auditory salience. *bioRxiv*.