UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

PhD Thesis

Automatic Analysis of the Acoustic Environment of a Preterm Infant in a Neonatal Intensive Care Unit

Author: Ganna Raboshchuk
Advisor: Dr. Climent Nadeu

Speech Processing Group
TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, July 2016

Моїм близьким,

Abstract

Most preterm newborns must be admitted to a Neonatal Intensive Care Unit (NICU) where they receive a specialized medical care, what, in many cases, is crucial for their survival. The acoustic environment of a typical NICU is highly diverse and may contain a large number of different sounds, which come either from various biomedical equipment or from human activities taking place in the unit. There exists a medical concern about the effect of that noisy acoustic environment on further growth and neurological development of preterm infants. The long-term effects of a NICU acoustic environment on a preterm infant could be revealed by the infant short-term reactions to auditory stimuli from it, which can be investigated by relating the presence of particular sounds with the preterm physiological variables. To carry out such statistical correlation study that uses the sound identities and their situation in time, big amounts of labelled audio data are required, which can hardly be obtained without using automatic detection from audio signals. Furthermore, automatic detection is also required for acoustic monitoring of the NICU environment to assist the medical staff in their work.

The major part of this thesis work is devoted to the challenging task of acoustic event detection in the NICU, where the goal is to develop robust systems able to detect and identify the sounds that appear in such environment. The detection of the two most relevant types of sounds is targeted in this work: equipment alarms and vocalizations. Acoustic alarms triggered by biomedical equipment play a key role in providing healthcare and are extensively present in a NICU environment. Several systems are proposed and developed in this thesis for automatic detection of particular types of alarm sounds. They are based on different approaches: a relatively simple signal processing based approach, which does not require model training; a model based approach that uses knowledge about the spectral and temporal structure of alarms and includes a specific feature extraction scheme; and, finally, an approach based on neural networks where the topology of the net is focalized to either a generic or a particular type of alarm sounds. The other type of considered sounds are vocalizations, a term used to encompass all sounds produced through a vocal tract. Vocalizations frequently happen in a NICU environment and may affect the preterm baby in various ways. The proposed binary detection system includes a prior vocalization enhancement step, and several techniques have been investigated for reduction of non-vocalization sounds. The development of the detection systems has required a design of proper evaluation metrics and the targeted medical application has been considered for that purpose.

A non-negligible part of the thesis work concerns the audio database acquisition and annotation. Due to the pioneering character of the application, a whole framework has been generated (in close collaboration with medical and engineering staff from the HSJD-Barcelona) for audio database production for the NICU environment, including key specifications like the recording setup and guidelines, and the labelling protocol. The produced database contains more than 1.5 hours of recorded audio data, and the laborious manual annotations cover roughly half of it. Finally, another contribution of this thesis work consists in an overall exploratory description of the NICU acoustic environment from the audio recordings. Unlike most previously published works, the whole content of the audio signal has been analysed, and, besides the usual measurements of sound pressure levels, the types of sounds and their spectro-temporal properties has been described. Additionally, a set of acoustic scenarios has been defined and described, and a sound taxonomy has been proposed for the NICU acoustic environment.

Resum

La majoria dels nadons prematurs han de ser ingressats a una Unitat de Cures Intensives Neonatals (UCIN), on reben l'atenció mèdica especialitzada que, en molts casos, és crucial per a la seva supervivència. L'entorn acústic d'una UCIN típica pot contenir un gran nombre de sons diferents, que provenen dels diversos equips biomèdics i de les activitats humanes que tenen lloc a la unitat. Existeix una certa preocupació mèdica per l'efecte d'aquest entorn acústic sorollós en el creixement i desenvolupament neurològic posterior dels infants prematurs. És probable que aquests possibles efectes a llarg termini de l'ambient acústic d'una UCIN sobre un infant prematur es manifestin en les seves reaccions a curt termini als estímuls auditius de l'entorn, els quals poden ser investigats relacionant la presència de sons particulars amb les variables fisiològiques del prematur. Per dur a terme aquest estudi de correlació estadística que utilitza les identitats dels sons i les seves ubicacions en el temps, es requereixen grans quantitats de dades d'àudio etiquetades, que difícilment poden ser obtingudes sense l'ús de detecció automàtica a partir dels senyals d'àudio. D'altra banda, la detecció automàtica també es requereix en la monitorització acústica de l'entorn de la UCIN per ajudar el personal mèdic en el seu treball.

La major part del treball d'aquesta tesi està dedicat a la tasca de detecció d'esdeveniments acústics en la UCIN, on l'objectiu és desenvolupar sistemes robustos, capaços de detectar i identificar els sons que apareixen en aquest entorn. En aquest treball es consideren els dos tipus de sons més rellevants: les alarmes d'equips biomèdics i les vocalitzacions. Les alarmes acústiques dels equips tenen un paper clau en la prestació de l'assistència sanitària i estan àmpliament presents en l'ambient d'una UCIN. En aquesta tesi, s'han proposat i desenvolupat diversos sistemes de detecció automàtica dels sons d'alarmes, que estan basats en diferents enfocaments: un primer enfocament, relativament simple, basat en el processament de senyals, que no requereix l'entrenament d'un model; un segon enfocament basat en modelatge, que utilitza coneixement de l'estructura espectral i temporal de les alarmes i inclou un esquema d'extracció de característiques específic; i, finalment, un enfocament basat en xarxes neuronals, on la topologia de la xarxa és, o bé generica, o bé focalitzada a un tipus particular de sons d'alarmes. L'altre tipus de sons considerat és el de les vocalitzacions, terme que abasta tots els sons produïts a través d'un tracte vocal. Les vocalitzacions ocorren freqüentment en l'entorn d'una UCIN i podrien afectar el nadó prematur de diverses maneres. El sistema de detecció binària proposat inclou un pas previ de millora dels senyals, i s'han investigat diverses tècniques per a la reducció dels sons que no són vocalitzacions. El desenvolupament de tots els sistemes de detecció ha requerit el disseny de mètriques d'avaluació apropiades i per a aquest propòsit s'ha tingut en compte l'aplicació mèdica que es persegueix.

Una part no menyspreable del treball de la tesi consisteix en l'adquisició de la base de dades d'àudio i en el seu etiquetatge. A causa del caràcter pioner de l'aplicació, s'ha generat un protocol sencer (en estreta col·laboració amb el personal mèdic i d'enginyeria de HSJD-Barcelona) per a la producció de bases de dades d'àudio per a l'entorn d'una UCIN, incloent especificacions claus com la configuració de la gravació i les directrius i el protocol d'etiquetació. La base de dades produïda conté més de 1,5 hores de d'enregistraments d'àudio, i les laborioses anotacions manuals en cobreixen aproximadament la meitat. Finalment, una altra contribució d'aquesta tesi consisteix en la descripció global exploratòria de l'ambient acústic de la UCIN a partir de les gravacions d'àudio. A diferència de la majoria dels treballs publicats anteriorment, s'ha analitzat tot el contingut del senyal d'àudio i, a més de les mesures usuals de nivells de pressió sonora, s'han descrit els tipus de sons i les seves propietats espectrotemporals. Per últim, s'ha definit i descrit un conjunt d'escenaris acústics, i s'ha proposat una taxonomia de sons per a l'entorn acústic d'una UCIN.

Acknowledgements

This document marks a completion of an important period of my life that gave a lot for my professional and personal growth. I would like to thank everybody who took part, in one way or another, in this incredible experience.

Foremost, I feel extremely grateful and fortunate for having spent this time under the tutorship of my advisor, Climent Nadeu. Not only he played a crucial role in making this thesis work possible and provided an indispensable guidance, but he also constantly encouraged me along the way and gave plenty of possibilities for my professional development. I feel grateful for having an opportunity to learn from his research experience and for all the time and effort he put in this work.

I'm very grateful to Ana Riverola de Veciana and Blanca Muñoz Mahamud for the helpful insights about the medical application and for their permanent enthusiasm, and to Santiago Navarro Hervas, Eva Bargallo and Cristina Borras Novell for their contributions at the beginning of this work. I want to express my deep gratitude to Peter Jančovič and Münevver Köküer for the opportunity of our collaboration and for the stimulating research stay at the University of Birmingham. I'm grateful to Sergi Solvez, Sergi Gomez Quintana and Alex Peiró Lilja for their contributions to the thesis and for our fruitful discussions, and to Vanessa Sancho Torrents and Francisco Alarcón Sanz for their work on the database annotation.

I want to thank all the colleagues from the TALP research center, specifically Omid Ghahabi, Pooyan Safari, Rupayan Chakraborty, Henrik Schulz, Lluis Formiga, Igor Jauk, Javier Hernando, Enric Monte, for the enjoyable research environment. I'm grateful to Martin Wolf for sharing his experience and giving valuable advices at the beginning of this work, and to Diego Lendoiro and Carlos Nistal for providing an excellent technical support.

I would like to express my gratitude to Taras Butko for having inspired me to pursue a PhD in another country and for being so helpful at the initial steps. I'm particularly grateful to Sergey for constantly supporting me in this decision. Thanks to all my beloved friends for being in my life and for supporting me all along this experience, and specifically to Nai and Ily without whom Barcelona wouldn't feel as much home as it does.

Last but not least, I would like to say a huge thank you to my family for always believing in me without a slightest bit of doubt and for making me feel special.

Contents

List of Acronyms

AED	Acoustic Event Detection
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CPAP	Continuous Positive Airway Pressure
DBN	Deep Belief Network
DET	Detection Error Tradeoff
EER	Equal Error Rate
EOP	Energy Overload Protection
FAR	False Alarm Rate
FF-LFBE	Frequency Filtered Logarithm Filter Bank Energies
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
MCRA	Minima-Controlled Recursive-Averaging
ME	Morphological Envelope
MF	Matched Filter
MFCC	Mel-Frequency Cepstral Coefficients
MR	Missing Rate
NICU	Neonatal Intensive Care Unit
NMF	Non-negative Matrix Factorization
NN	Neural Network
PCA	Principal Component Analysis
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machine
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SS	Spectral Subtraction
SVM	Support Vector Machines

List of figures

List of tables

Chapter 1

Introduction

1.1 Thesis motivation

Newborns delivered at a gestational age of 24-32 weeks (very low birth weight preterms) commonly have health problems and must be admitted to a Neonatal Intensive Care Unit (NICU), what, in most of the cases, is crucial for their survival. The increased survival and reduced neonatal morbidity of preterm infants in the past three decades has not always been accompanied by an improvement in their neurological development [1]. It is known that the noisy acoustic environment of the NICU may have adverse effects on the growth and neurodevelopment of the preterm infants as inadequate, loud, unexpected sounds replace natural hearing placental stimulation [2]. The negative or stressful environmental impact of NICUs on the developing brain has been widely documented [3–6]. An important negative effect of noise in the NICU is the one it has on sleep, which is essential for neurosensory development, learning, memory and preservation of brain plasticity [7].

The acoustic environment of a typical NICU is highly diverse and may contain a large number of sounds coming from numerous sources, such as alarm sounds generated by different biomedical equipment, noisy mechanical ventilation, telephone ring sound, people conversations, etc [8,9]. Various acoustic events are usually taking place simultaneously in a NICU and the maximum sound pressure level limits recommended [10] are exceeded frequently [11], being of a great concern in the medical literature.

Different ways that have been proposed to deal with this can be divided into two groups. The major group of methods is aimed at analysing and changing the acoustical environment of a NICU (for example, by planning a more rational distribution of the wards [12] or controlling and reducing the activities taking place in it [13]). The other group of methods directly concerns the preterm baby and implies protecting a baby with special accessories (earmuffs) [14,15]. But although the prolonged time of sleep and stress reduction were reported, the effect on the physiological variables has not been completely proved and there is no information about the long-term outcomes. This thesis work, which consists in analysing the acoustic environment of a preterm baby, is in the scope of the first group of methods.

The acoustic environment of a preterm baby admitted into a NICU has been the object of a number of reported studies during the last two decades. These works analyse the environment placing a microphone both inside [16] and outside the incubator. Usually, sound is represented only by its intensity level and just a few works weakly analyse sound spectra [9, 17]. Moreover, to our knowledge, very little studies considered the intensity levels of specific sounds [8] or analysed specific conditions of the NICU's acoustic environment [18]. Unlike most previous works, in this work the whole content of the audio signal is used and an automatic description of the whole sound landscape is pursued.

The short term effects of a NICU acoustic environment on a preterm infant could be revealed by the infant reactions to auditory stimuli from it, which can be investigated by relating the presence of particular sounds with the preterm physiological variables. Note that in such study the sounds are not produced artificially, but occur naturally in the NICU environment and are the ones actually perceived by the preterm infant. Such investigation can complement greatly the work already reported in the literature, in which only the sound pressure level is considered without taking into account the spectro-temporal properties and identity of sounds (e.g. in [19]).

To carry out a statistical correlation study that uses the sound identities, big amounts of labelled audio data are required, which can hardly be obtained without using automatic detection from audio signals. The manual annotation of audio data is usually extremely time consuming and tedious, requires specific skills, prone to errors and may be inconsistent when several independent annotators are involved [20]. A robust detection system may be used to overcome these limitations while keeping the human-involved labour at a minimum level.

In addition, automatic detection systems may be used for real-time or offline acoustic monitoring of the NICU environment. While video monitoring provides valuable information, the automatic monitoring based on audio information has several advantages: cheaper sensors, lower computational requirements, avoiding image limitations (e.g. blind spots), etc. [21]. Specifically, in the NICU environment a lot of distributed cameras would be needed.

Therefore, the major part of this thesis work is devoted to the task of Acoustic Event Detection (AED), where the main goal is to develop robust systems able to detect and identify the sounds that appear in the NICU environment. Due to the multisource nature of that environment and the fact that most sounds are simultaneous, that AED task is rather challenging. In particular, the detection of the two most represented types of sounds is targeted: equipment alarms and vocalizations.

Equipment alarms, which are extensively present in a NICU environment, are used in biomedical equipment to alert of situations requiring medical attention. The fact that a large number of sounding alarms are not clinically relevant and/or are unrelated to emergency situations [22], and also general noise and information overload may lead to unsatisfactory quality of healthcare provided by the medical staff. Intelligent alarming systems are being proposed [22, 23] in order to improve the alarm handling process in NICUs and reduce noise levels. These solutions make use of alternative alarm modalities and usually imply development of a distributed alarm system where all or almost all the medical equipment is connected to a central monitoring system and medical staff carries personal notification devices, whereas only the most critical alarms are sounding. Unfortunately, in the majority of the NICUs smart alarming solutions are yet to be developed.

Vocalizations, which encompass all sounds produced through a vocal tract either by infant or adult, are the sounds most frequently happening in a NICU environment and may affect a preterm baby [24,25]. For instance, newborns demonstrate a clear preference for the maternal voice [3], which can have a calming effect, while shouts or cries may affect them in a negative manner.

1.2 Thesis objectives

The main objective of this thesis work is the automatic analysis of the acoustic environment of a preterm infant in a Neonatal Intensive Care Unit (NICU), which can be expressed in terms of several specific objectives which are relevant for the medical application.

I. An overall description of the NICU acoustic environment: existing acoustic scenarios, sound types and their sources, sound categories, relations between sounds, etc. This is relevant for the medical application in two senses:

    (a) The knowledge of the NICU acoustic environment is required to implement policies for consistently and substantially reducing the types of noise that may affect the newborns.

    (b) To design a NICU room with better acoustic characteristics. General principles for designing a quiet NICU were proposed in [10,12], yet more studies about the effects of various NICU designs on infants, parents, clinicians and the delivery of services are needed to advance the field of design.

    Apart from that, audio data description is a required initial step for any audio classification task as it helps to explore the data domain.

II. The second objective, which actually is the main one, is detection of some relevant NICU acoustic events: equipment alarms and vocalizations. The aim of the developed Acoustic Event Detection (AED) systems is to automatically label temporal regions within the input audio where a particular sound is present, i.e. to specify the start and end time of each relevant sound occurrence. This can be useful for two medical application purposes:

    (a) To assist the medical staff in their work and facilitate the reaction to events. For example, in [13] a sound-activated light device was implemented for alerting the staff members when the sound pressure level exceeded a predefined threshold. The automatic detection system can be a part of a more sophisticated notification system allowing smart handling algorithms, which could be designed to warn about triggering of particular sounds, to take into account their clinical relevance and urgency, etc.

    (b) To detect sounds that may be affecting the preterm infant. Sound description obtained from AED systems can be correlated with the preterm physiological variables in order to investigate how a preterm infant reacts to the sounds that take place in a NICU.

III. The acquisition and annotation of an audio database is required in order to meet the above mentioned objectives. The produced database, that captures the NICU environment in various conditions, is important both for performing its acoustic description and for developing automatic detection systems for the relevant types of sounds. Due to the pioneering character of this work, a general framework of the database production for the NICU environment has to be designed, which includes specifying the recording setup (equipment, conditions, timetable, guidelines, etc.), the considered acoustic scenarios and the labelling protocol.

1.3 Thesis outline

The thesis is organized as follows.

Chapter 2 starts with reviewing the work already reported in the literature about the analysis of the acoustic environment of a preterm in a NICU. It then presents a literature review for AED from a point of view of possible applications, further discussing the state of the art approaches to feature extraction, classification algorithms, audio enhancement and source separation techniques that have been used for AED so far.

Chapter 3 reports our work on the audio database acquisition and annotation, where we provide details of the recording setup, specify the selected acoustic scenarios and describe the recording sessions carried out. Also, in this chapter we outline the annotation campaign progress, define the labelling protocol and give information about the produced annotations.

Chapter 4 contains the results of acoustic description of the NICU environment, where we provide an extensive list of acoustic events found in that environment, structure them into acoustically homogeneous groups by building a taxonomy and present the analysis of the major types of sounds. Also, the considered acoustic scenarios are characterized.

Chapter 5 and Chapter 6 describe our work on the automatic detection systems for two types of sounds that are the most represented ones in a NICU environment: alarms and vocalizations. In Chapter 5 several systems for automatic detection of equipment alarm sounds are proposed, which deal with the problem from three different perspectives: a signal processing based approach, a knowledge-based approach that employs machine learning, and an approach based on neural networks. Chapter 6 presents our work on an automatic system for detection of vocalization sounds. The focus is put on reducing the presence of irrelevant sounds prior to detection, and so several techniques for vocalizations enhancement are investigated.

Finally, Chapter 7 concludes the work. The main achievements are summarised and several promising directions for future work are highlighted in this chapter.

Chapter 2

Literature review

2.1 Chapter overview

In this chapter, the work done so far on the acoustic description of a Neonatal Intensive Care Unit (NICU) environment and in the area of Acoustic Event Detection (AED) is reviewed.

The sections of this chapter are organized as follows. Section 2.2 describes the work already reported on analysis of the acoustic environment of a NICU. Section 2.3 presents the AED task. The review of the state-of-the-art approaches in areas of feature extraction, classification algorithms, audio enhancement and source separation is provided in Sections 2.4, 2.5 and 2.6, respectively.

2.2 Acoustic analysis of a NICU environment

The acoustic environment of a NICU and its effects on preterm infants have been extensively studied during the last two decades, where the research works mainly concerned describing a NICU acoustic environment and its major sound sources, looking for possible solutions for reducing the noise levels within a NICU and evaluating their effectiveness, finding the potential adverse effects of the NICU acoustic environment on preterm babies, and studying the impact of the NICU design on well-being of neonates.

The common approach to the description of a NICU acoustic environment lies in measuring the Sound Pressure Levels (SPLs) with the microphone placed in the central location of the room [26]. In some works, like in [27], the microphone is placed inside the incubator, at some distance from the infant's ear, so as to measure what the preterm infant is perceiving. The average noise levels measured in different conditions (e.g. during day and night shifts) are reported and compared with the recommendations of the American Academy of Pediatrics [6]. The SPL values are usually presented in dBA units. For that purposes, the A-weighting is first applied to modify the audio signal spectrum, modelling the spectral response of the human auditory system, as was done, for example, in [11, 17]. Such studies usually describe the typical sound sources, which could contribute to high SPL values, and propose a potential solutions for reducing the noise levels within a NICU [28].

Attempts on characterizing the sound levels of particular sound sources have been reported in the literature. For instance, in [8] the noise level increment due to the most frequent types of sounds, like phone ring, alarms, speech, baby crying, is analysed. The measurements were made with the microphones placed both inside (1 cm near the newborn's ear) and outside (at around 50cm distance) the incubator, and also the background SPLs were measured at a central location in the middle of the unit room.

To our knowledge, few works analysed the sound spectra. In [9] the spectral analysis of the noise generated by individual equipment and activities is performed, where the SPLs across the frequency spectrum are analysed. Similar work, which concentrated on investigating the major sound sources within an incubator (incubator cooling fan and ventilator), is reported in [17]. In this work, the SPLs, measured during the routine morning clinical activity close to a dummy infant's ear and at the head level outside the incubator, are analysed at different frequencies and their cumulative values are reported.

In [5] an overview of studies looking for potential adverse effects of the noisy NICU acoustic environment on cardiovascular, auditory, nervous systems and on long-term neurodevelopment of preterm infants is given. The problem of validity and reliability of such studies is discussed in [29], where the authors provide clear criteria for research evaluation and identify possible study design problems. In particular, the importance of measuring SPLs at the infant's ear or a specific distance from it is emphasized.

In [24] the importance of the positive early auditory experience for the development of the preterm infants is described. The appropriate auditory input and careful protection against overstimulation while the stay of a neonate in the NICU are emphasized. Among the proposed recommendations are playing the mother's voice inside the incubator and introducing vocal music (such as lullabies). In [30] the effects of music therapy (recorded music, parent voices, sung lullabies) on physiological variables, sleep and feeding of preterm infants were explored. The study reported in [31] revealed that the controlled music stimulation appears to be a safe and effective way to ameliorate pain and stress in premature infants following heel sticks. On the other hand, the problem of feasibility and safety of producing the maternal sounds (i.e. voice, heartbeat) inside the incubator with the preterm baby was addressed in [32]. In [25] the acoustic environment of a preterm infant cared for in the NICU was explored in order to determine the influence of the parent's speech on the number of preterm infant vocalizations over time. For that purpose, the counts of adult words, infant vocalizations and conversational turns were related.

Due to the excessive noise levels reported in the literature [11, 33], some studies concentrated on optimizing the hospital environment for preterm newborns. The main purpose of these works is to ameliorate the negative effects of noxious stimuli and to decrease unfavourable auditory stimulation to a minimum. For instance, in [22] the authors focused on investigating the ways of reducing the noise levels and the number of sounds to which the newborns are exposed by improving the alarms handling process, which could be done by introducing technological and organizational changes. In [13] the efficacy of the specifically designed light-alarm device for reducing the staff-produced noise in a NICU was studied. The device was activated when the sound levels exceeded a threshold in order to notify staff of it.

Various studies have evaluated the impact of the NICU design on preterm infants. In particular, the literature review reported in [34] explores the main features of the NICU design and links a range of aspects of the physical environment of NICU to well-being of neonates, family comfort and caregiving process. The study reported recently in [35] investigates the relationship between the NICU room type and the primary outcome of neurodevelopmental performance at the age of 2 years.

Advancing methods in medical practice often require new healthcare facilities or improvements of the existing ones. The study reported in [18] compares the sound levels, staff perceptions and patient outcomes before, during and after the renovation project taking place near the NICU. In [36] the noise levels before and after the structural reconstruction within a NICU were compared. In this study, authors divide all the sound sources within a NICU into two groups: 1) operational (staff or equipment generated sounds); 2) structural (building generated sounds).

2.3 Acoustic event detection in real-world environments

A real-world acoustic environment represents a complex scenario, and is characterized by a huge number of sound sources, occuring spontaneously and possibly overlapping, along with different combinations of their positions. Depending on the enviromnent, the types of sounds that may be encountered spread from natural ambient sounds, through sounds produced by humans and animals, to artificial sounds coming from machines and equipment. The human activity in such environments is reflected in a rich variety of sounds, produced by a human body or by objects handled by humans. All these sounds are connected to a wide variety of objects, actions, events, and communications, thus an acoustic environment is a rich source of information on the types of activities, participants involved, and communication modes.

While speech is, obviously, the most informative sound, other kinds of sounds may also provide useful cues for context understanding. For instance, in a meeting/lecture context, we may associate a chair moving or door noise to its start or end, cup clinking to a coffee break, or footsteps to somebody entering or leaving. Furthermore, some of these acoustic events are tightly coupled with human behaviors or psychological states: paper wrapping may denote tension; laughing, cheerfulness; yawning in the middle of a lecture, boredom; keyboard typing, distraction from the main activity in a meeting; and clapping during a speech, approval.

Generally, the term *acoustic event* or *sound event* refers to a label that people would use to describe a recognizable event in a region of audio [37]. More specifically, as it is said in [38], it is “any possible audible acoustic event which is caused by motions in the ordinary human environment; they have real events as their sources; they are meaningful, in the sense that they specify events in the environment (...)”. Hence, acoustic events can be used to represent an acoustic environment in a symbolic way, e.g., a busy street environment contains events of passing cars, car horns, or footsteps of people rushing.

AED is a discipline belonging to the area of the computational auditory scene analysis [39] that consists of processing acoustic signals and converting them into symbolic descriptions corresponding to a human listener’s perception of the different sound events that are present in the signals and their sources. While acoustic event classification deals with events that have already been isolated from their temporal context, acoustic event detection aims at determining both the temporal positions and the identities of sounds in a continuous audio stream. So far, acoustic event detection and classification has been found useful in manifold applications, like surveillance [21,40], ambient assisted living [41,42], robotics [43], information indexing and retrieval [44,45], to list a few.

From a semantic point of view, AED has been addressed for recognition of generic sounds and sounds specific to a given environment or activity. For instance, in [46] the problem of detecting alarm sounds was investigated, where such alarms as phone rings, smoke alarms, sirens, car/truck horns were considered. Detection of acoustic events has been carried out in different environments like living environments [47,48], kitchen rooms [49], bathrooms [50], public places [51], offices [52], meeting rooms [53], industrial workplaces [54], etc. It has been showed, that AED systems can benefit from the incorporation of the context-related information, like count-based event priors and context-dependent acoustic models [37].

While considerable effort has been devoted to speech and music related research, the AED task has been far less studied. Nevertheless, there exists an increasing interest in this topic and several international efforts to evaluate systems designed to recognize acoustic events have already been conducted. In the framework of CLassification of Events, Activities and Relationships (CLEAR) campaigns in 2006 [55] and 2007 [56] the participating systems were evaluated on the set of acoustic event classes specific to a meeting-room environment. During the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (D-CASE) in 2013 [52] the problem of detecting acoustic events in an office enviromnent was addressed.

2.4 Audio feature extraction

Numerous studies have been devoted to the topic of audio feature extraction, and audio features proposed in the literature can be roughly divided into three categories:

- Time-domain features are computed directly from the audio waveforms. Features like frame energy, zero-crossing rate,high zero-crossing rate [57] comprise this category.
- Frequency-domain features are derived from the Fourier transform of the time signal over a fixed time period, e.g. frame. This category contains such features as fundamental frequency [58], spectral centroid, spectral flux [43, 59], spectral rolloff [57], spectral tilt [60], filter-bank energies [61], etc.
- Time-frequency or spectro-temporal features operate in time and frequency domain jointly. Some variants of those features will be discussed later in this section.

Depending on the time span considered the features can be regarded as short-term (frame-level) or long-term (segment-level). The frame-level features are usually designed to capture the short-term characteristics of the audio signal, where the signal can be considered stationary. The concept of the audio frame comes from traditional speech signal processing, where analysis over a short time interval has been found to be appropriate. To extract the semantic content, the temporal variation or evolution of signal is observed on a longer time scale. This consideration has lead to the development of various segment-level features. The segment duration, in this case, may span up to several seconds. For example, these features can be built on top of the frame-level features and characterize how they are changing over a segment with some kind of statistics, e.g. bag-of-frames approach [62] or factor analysis [63]. It is worth mentioning that the temporal evolution of the audio signal could be captured on the model level, for instance, by Hidden Markov Models (HMMs).

In [64] the authors model the temporal context by employing convolutive Non-negative Matrix Factorization (NMF) for feature extraction. NMF is used for discovering a set of spectro-temporal patches which roughly correspond to the acoustic events considered, and the features are derived from the activations of this patches in time.

As a common approach, audio signals have been often characterized with the conventional well investigated features for Automatic Speech Recognition (ASR), like Mel-Frequency Cepstral Coefficients

(MFCC) features [65], Frequency Filtered Logarithm Filter Bank Energies (FF-LFBE) [66] or linear prediction coefficients. Usually, these features are combined with the specific features designed for a particular application [43, 64], and in most of the cases this leads to a better recognition accuracy. The temporal evolution of the above mentioned features may be incorporated by adding duration and dynamics features. In the later case, usually the first and the second temporal derivatives are used [67, 68].

Several techniques, that have been proposed for describing the temporal dynamics of audio signal, work in a modulation frequency domain [69,70]. Generally, these techniques divide the audio signal to a set of sub-bands and derive the sub-band amplitude modulation envelopes, which are further converted to a modulation frequency domain. Conversely, in [70] the modulation features are obtained by means of a frequency-domain linear prediction, that provides an approximation of the temporal (Hilbert) envelope of the time domain signal.

In [71] authors explored the use of recurrence quantification analysis for providing additional information about the temporal evolution of audio. In particular, it was applied for the characterization of environmental sounds. The features based on that analysis do not require assumptions about linearity or stationarity of the time series, and are extracted from the frame-level spectral audio features (namely, MFCC features).

Many research works on audio feature extraction aimed at emulating human audio perception mechanism in order to achieve human-like performance. These approaches try to model functioning of the human auditory system, have similar structures as the human auditory pathway, yielding so-called perceptual psychoacoustic features.

In the Perceptual Linear Prediction (PLP) technique [72], the short-term spectrum of the speech is modified by several psychophysically based transformations, which make use of the three concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum: the critical-bands spectral resolution, the equal-loudness curve, and the intensity-loudness power law. This feature extraction technique is widely used in speech recognition systems.

The author in [73] proposes an approach to feature extraction which utilizes two-dimensional spectro-temporal modulation filters (Gabor functions). The use of two-dimensional Gabor filters is motivated by their similarity to the spectro-temporal patterns of neurons in the auditory cortex of mammals, reported in physiological and psychoacoustic studies. The resultant feature vector size is relatively large (more than 2000).

In [43] the authors propose to use the matching pursuit algorithm to obtain time-frequency features. The matching pursuit based method utilizes a dictionary of spectro-temporal Gabor atoms to select features. The results showed that this approach is promising in capturing unstructured environmental sounds, while traditionally used MFCC features may fail to effectively model them.

In [74] the audio signal is passed through a pole-zero filter cascade (a time-varying filterbank) to simulate the output of the inner hair cells along the length of the cochlea. In [75] author proposes to first modify the signal spectrum by means of an A-weighting filter, which models the spectral response the human auditory system. In that work the use Gammatone filterbank is also explored, which provides a good approximation to experimentally determined spectral responses of the basilar membrane in cochlea.

Instead of using the Fourier transform which results in a constant time-frequency resolution, some works on feature extraction employed the wavelet transform [47, 75]. It provides better frequency resolution at low frequencies and better time localization of the transient phenomena in the time domain. In that respect it resembles the first stage of human auditory perception and to basilar membrane excitation which exhibits similar time-frequency resolution characteristics.

Because of the large number of possible features some studies focused on elaborating feature selection techniques [76]. The feature selection is traditionally motivated by three main reasons [68]: performance improvement; general data reduction, to limit storage requirements and increase algorithm speed; and feature understanding, to gain knowledge about the process that generated the data. There are a lot of different feature selection algorithms reported in the literature. In [77] significant feature space reduction was obtained by applying Principal Component Analysis (PCA) in each frequency sub-band. In [68] a fast one-pass-training technique was introduced in the context of a multimodal AED, resulting in an optimal feature subset for each acoustic event class.

2.5 Classification algorithms

Any recognition task requires a classification step, on which the features are fed to a classifier that provides a label for an unseen input pattern.

One of the very first works on audio classification used a minimum distance classification model, i.e. a simple distance-based classifier with the Euclidean distance between extracted features [78]. The minimum distance classifiers choose a class according to the closest training sample. A bit more complex algorithms determine k-nearest neighbors to an unknown input, and then they choose the class that is most represented by that neighbors [43]. Yet, classification becomes very complex when a lot of training data is used as one must measure a distance to all the training samples. By using clustering and storing only the centres of the clusters (class prototypes) the computational efficiency can be improved. The mentioned algorithms and other related optimization steps for audio classification have been reviewed in [49, 51].

A rule-based classification algorithm, that is based on a good task domain knowledge, has been used in [58]. In that work, several simple task-specific features were put to work with a set of heuristic classification rules to ensure the feasibility of real-time processing. In [54] the detection algorithm applies decision rules to zero-crossing rate of the autocorrelation of the signal envelope for detecting alarms in industrial environments, where the real-time detection with low latency is crucial for notifying the user of a dangerous situation.

Among other classification paradigms, a way to classify audio data consists in using already developed and well-tested speech recognition algorithms. In ASR, usually Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) are used [79]. They are well suited to work with time series data, and to deal with the information included in the temporal evolution of the audio signal. A lot of audio recognition works have exploited the above-mentioned techniques. For instance, GMMs have been used in [57,65], and HMMs in [21,41,63,70]. Instead of using generative classification models, discriminative ones, like artificial neural networks in [46,48] or Support Vector Machiness (SVMs) in [45,59,67], have been used in a number of works.

Recently, deep learning has been drawing a lot of interest in the pattern recognition field [80], dramatically improving the state-of-the-art in image, audio, and speech processing areas [61, 81, 82]. For instance, the combinations of Deep Neural Networks (DNNs) and HMMs (the so-called hybrid approach) applied to speech recognition tasks are able to achieve significantly higher accuracies than the conventional GMM-HMM classifiers in [83, 84]. Among the main factors responsible for the recent emergence of DNNs are the possibility of initializing the weights sensibly, which results in a very efficient learning algorithm [85], and the dramatic improvement of computing power, which makes it feasible to train DNNs effectively. Commonly, the network weights are initialized by unsupervised generative training, then, by adding a top label layer and using a standard backpropagation algorithm, the generative network is converted to a discriminative one and, therefore, becomes appropriate for classification tasks [83] (see Section 2.5.1 for more details).

A large variety of DNN architectures have been proposed in the literature. Typically DNNs are designed as feedforward networks, but other variants like recurrent neural networks (and, in particular, Long Short-Term Memory networks (LSTMs)) [86] and Convolutional Neural Networks (CNNs) [87,88] have been successfully applied to audio and speech processing tasks. In [89] the three types of networks (namely, DNNs, LSTMs and CNNs) are combined in a unified architecture and trained jointly, thus, the complementarity in their modelling capabilities is exploited.

Deep learning requires very little engineering by hand and can easily take advantage of increases in the amount of available computational power and data [80]. The powerful learning procedures allow DNNs to handle correlations between input features, and it has been shown that for the ASR task they work significantly better on filterbank outputs than on standard MFCC features [90]. Moreover, attempts on directly modelling a raw waveform have been reported in the literature [91], although such systems have not yet outperformed the ones employing filterbank features.

In [92] the feature extraction is integrated in a DNN, and mel-scale filterbank is replaced by the filterbank layer, which is trained jointly with the rest of the network. It is argued that the standard perceptually motivated filterbank may not be particularly optimal for the ASR task, and letting the network learn appropriate feature extraction and discrimination is much more powerful. Both feature extraction and acoustic modelling are performed jointly by a DNN in [93] for the voice activity detection task, where the standard PLP features are approximated by a neural network.

Unlike the binary classification algorithms, where the decision is taken between two classes, the multiclass classifiers can deal with several classes. Some classification algorithms, like neural networks, k-nearest neighbors, HMMs, naturally permit the use of more than two classes. Yet, a combination of several binary classifiers can be used to solve a given multiclass problem [94]. Among the commonly used strategies are: one-against-all, where for each class a classifier which distinguishes that class from all other classes is trained; one-against-one, where the classifier is trained for every pair of classes and the decision is taken by voting. In [95] a hierarchical architecture for a group of binary classifiers was proposed for audio segmentation task, where each classifier is responsible for detecting the class of interest and posterior classifiers benefit from the previous classifier decision.

A post-processing (smoothing) is commonly applied to yield longer contiguous segments of classes of interest, which corresponds to de-noising the classification output. In [57] for the speech / non-speech segmentation system a post-processing of GMMs and maximum entropy classifier outputs is performed by using an ergodic HMM, which has two states (“speech” and “non-speech”) and equal transition cost in either direction. There is no cost for remaining in the same state, what creates a smoothing effect by discouraging state changes. In [96] the authors employed convolutive NMF to force grouping of SVMs outputs into events. The proposed approach is reported to yield better results compared to the traditional “winner-takes-all” strategy, where the whole audio segment is assigned to a class to which the majority of the classifier outputs in that segment belong. In [67] the hypothesized sequence is smoothed by assigning to the current decision point the label that is the most frequent in a string of five decision points around the current one.

2.5.1 Neural network based pre-training

Deep Belief Networks (DBNs) are originally probabilistic generative models with multiple layers of stochastic hidden units above a layer of visible variables which represent an input vector (e.g., see Figure 2.1). There is an efficient greedy layer-wise algorithm for learning DBNs [85]. The algorithm treats every two adjacent layers as a Restricted Boltzmann Machine (RBM) network, which is constructed from a layer of binary stochastic hidden units and a layer of stochastic visible units. The output of each RBM is considered as the input to the next RBM.

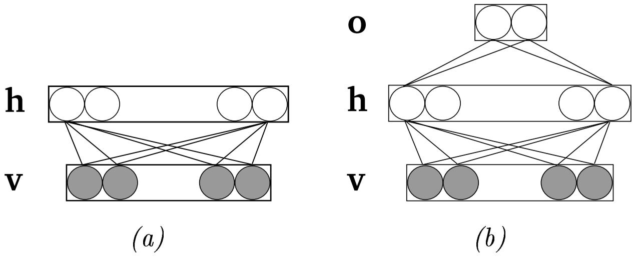


Figure 2.1: (a) Generative one-layer DBN structure and (b) whole NN structure used in the experiments.

Training an RBM is based on an approximated version of the contrastive divergence algorithm [82, 85] which consists of three steps. At first, hidden states ($\mathbf{h}$) are computed given visible states ($\mathbf{v}$); then, using those $\mathbf{h}$, $\mathbf{v}$ is reconstructed; and, on the third step, $\mathbf{h}$ is updated, given the reconstructed $\mathbf{v}$. Finally, the change of connection weights is given as follows,

$$\Delta w_{ij} \approx -\alpha \left(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon} \right), \tag{2.1}$$

where α is the learning rate, w_{ij} represents the connection weight between the visible unit i and the hidden unit j, $\langle . \rangle_{data}$ and $\langle . \rangle_{recon}$ denote the expectations when the hidden state values are driven respectively from the input visible data and the reconstructed data.

In fact, the training process tries to minimize the reconstruction error between the actual input data and the reconstructed one. The parameter updating process is iterated until the algorithm converges. Each iteration is called an epoch. It is possible to perform the parameter update after processing each training example, but it is often more efficient to divide the whole input data (batch) into smaller size batches (minibatch) and to do the parameter update after each minibatch. More theoretical and practical details can be found in [82, 85, 97].

When the unsupervised learning is finished, it can be converted to a discriminative model by adding a label layer ($\mathbf{o}$) on top of the network and doing a supervised backpropagation training. In other words, the unsupervised learning can be considered as a pre-training for the supervised stage. It has been shown [85] that this unsupervised pre-training can set the weights of the network to be closer to a good solution than random initialization and, therefore, avoids local minima when using supervised gradient descent.

In this work, a Gaussian-Bernoulli RBM [97, 98] is employed as the first RBM. The input vectors are mean-variance normalized before being fed to the network; the mean and variance values calculated on the training data are also applied to the testing data. The training data is balanced with regards to classes by randomly selecting samples of the predominant class.

2.6 Audio enhancement and source separation

Like in other audio processing areas, acoustic event detection or classification in mismatched conditions is a very challenging task, and the recognition systems are known to deteriorate in the presence of background noise. Audio enhancement techniques deal with this problem by denoising the signal so that the extracted features are closer to the training conditions. The typical techniques include spectral subtraction (see Section 2.6.1) and Wiener filtering [99].

In rich uncontrolled multisource environments AED systems undergo the problem of temporal overlappings between events, which makes the recognition problem more challenging. One of the possible solutions to the overlapping problem is employing source separation at the front end of the system, which allows to segregate the desired source from the other interfering sources or noise. Source separation techniques can be broadly classified into two categories:

1. Array processing based separation that exploits the information about the positions and orientations of the sources and sensors [100, 101]. In [101] although the partial source separation is achieved by employing null-stearing beamformers, the signals obtained after it are used for joined classification and localization of acoustic events.
2. Blind Source Separation (BSS), which separate the signal with or without the aid of the information about the source signals or mixing process [102].

It should be mentioned that in a multi-microphone setup a channel selection technique may be applied for selecting the signal with the highest quality, which may be either the intrinsic quality of a given signal or how well it fits the acoustic models of the recognition system. Several such signal-based and decoder-based techniques have been proposed in [103] in the context of an ASR task.

Many different approaches of source separation have been attempted by numerous researchers using artificial learning, higher order statistics, minimum mutual information, beamforming based adaptive signal separation and noise cancellations, each claiming various degrees of success. Among the most well-known BSS techniques are PCA, Independent Component Analysis (ICA), NMF, deflation approach and singular value decomposition.

In particular, PCA is a second order statistical method that decorrelates the data and reduces the dimensionality of the problem, but does not achieve full independence [104]. Moreover, it generally results in variables that are hard to interpret. On the other hand, ICA [105] has the capacity to make the signals fully independent, since it is a higher order statistical method. The combination of both PCA and ICA is used for BSS in [106], where initially the dimensionality of the problem is reduced using PCA, thus making it easier for the ICA algorithm to solve it. In [107] the deflation method was proposed, which consists in iteratively extracting the sources one after the other. The main advantage of this method lies in the fact that the contrast function does not present any spurious local maximum.

Among the popular techniques applied in the noisy or overlapped audio recognition, NMF-based are the ones that have recently received the most interest in the literature. NMF (see Section 2.6.2) in its basic form was first presented in [108], where it was applied for learning parts-based representation of face images and for automatic semantic indexing of encyclopedia articles by topic. Since then it has proven to be useful in many pattern recognition areas [109–111], and in particular in audio processing, and various algorithm enhancements were prosoposed [112]. Both single-microphone [113] and multi-microphone versions of NMF have been widely investigated [114].

In [115] the convolutive NMF version is introduced, which takes into account the dependencies between successive columns of the input matrix. In audio spectrum analysis such dependencies across columns correspond to the time-frequency representations or patterns. Further inclusion of the sparseness constraint on the activation patterns [116] enables the discovery of the over-complete representations.

In [113] an AED system for the natural multisource environments which uses a sound source separation front-end is described. The audio is pre-processed using an unsupervised NMF-based algorithm, which is capable of separating up to four individual streams from the overlapping events audio, where each of the streams corresponds to a combination of the physical sources present in the original signal. This work was further continued in [117], where the problem of finding the separated stream which contains the targeted source is solved using two iterative approaches based on the expectation-maximization algorithm.

In [118] the exemplar-based sparse representations are used for the noise robust ASR. The noisy speech is modeled by a sparse linear combination of speech and noise exemplars, which are, correspondingly, the examples of clean speech and noise segments spanning multiple time frames. In this work, the exemplar-based approach is used as a source separation technique in order to do missing data mask estimation and feature enhancement. The hybrid exemplar-based/HMM method called sparce classification is employed, where the calculation of the likelihoods of HMM states is based on the activations of exemplars.

Probabilistic latent component analysis (PLCA), which is numerically equivalent to NMF, is presented in [119] for the separation of multiple speakers from mixed single-channel recordings. Unlike NMF, the probabilistic decomposition naturally extends from matrices to tensors of arbitrary dimensions [120]. Due to the possibility of statistical interpretation, it can be easily extended or generalized, to allow, for example, overcomplete sparse representations, invariance to transformations, etc. The use of PLCA in several audio-related applications such as feature extraction, source recognition, source separation and denoising is described in [121].

2.6.1 Spectral subtraction and minima-controlled recursive-averaging algorithm

Spectral Subtraction (SS) algorithm is the classical tool used for audio denoising where an additive model is assumed, i.e. the noise-corrupted input signal $y(t)$ is composed of the clean signal $x(t)$ and the additive noise signal $d(t)$, that is $y(t) = x(t) + d(t)$. Then, the clean signal spectrum $\hat{X}(t,k)$ can be estimated by subtracting an estimate of the noise spectrum $\hat{D}(t,k)$ from the noisy signal spectrum $\hat{Y}(t,k)$ as follow [122]:

$$|\hat{X}(t,k)|^\gamma = \begin{cases} |\hat{Y}(t,k)|^\gamma - \alpha|\hat{D}(t,k)|^\gamma, & \text{if } |\hat{Y}(t,k)|^\gamma > (\alpha+\beta)|\hat{D}(t,k)|^\gamma \\ \beta|\hat{D}(t,k)|^\gamma, & \text{otherwise} \end{cases} \tag{2.2}$$

where t and k are, correspondingly, the frame and the frequency bin index, $\gamma = 1$ yields magnitude and $\gamma = 2$ yields power spectrum subtraction, α is the subtraction factor, which controls the amount of noise to be subtracted, and $0 < \beta \ll 1$ is the spectral floor parameter, which controls the amount of the residual and perceived musical noise. This approach is referred to as SS using oversubtraction (because usually $\alpha \geq 1$) [99].

The use of a proper noise estimate $\hat{D}(t,k)$ is crucial for the quality of the enhanced signal. Often, it is obtained once from the first frames of the input audio. Alternatively, the noise estimate can be obtained and updated along the input signal, taking into account the probability of speech presence. In the Minima-Controlled Recursive-Averaging (MCRA) algorithm [99], employed in this work, an estimate of the noise power spectrum is obtained recursively as follows:

$$|\hat{D}(t,k)|^{\gamma} = \alpha_d(t,k)|\hat{D}(t-1,k)|^{\gamma} + (1-\alpha_d(t,k))|\hat{Y}(t,k)|^{\gamma}, \tag{2.3}$$

where $\alpha_d(t,k)$ is a smoothing factor defined as

$$\alpha_d(t,k) = \alpha + (1-\alpha)p(t,k). \tag{2.4}$$

Here, $p(t,k)$ is the speech-presence probability which is calculated using the ratio of the smoothed noisy signal spectrum to its local minimum. The smoothed noisy spectrum $S(t,k)$ is obtained as follows:

$$S(t,k) = \alpha_s S(t-1,k) + (1-\alpha_s)\sum_{i=-L_w}^{L_w} w(i)|\hat{Y}(t,k-i)|^{\gamma}, \tag{2.5}$$

where α_s is a time smoothing factor, and the second term represents smoothing over frequency with the Hamming windowing function $w(i)$ and a window of length $2L_w + 1$. The local minimum $S_{min}(t,k)$ is found via samplewise comparison of $S(t,k)$ in the previous D frames. The ratio $S(t,k)/S_{min}(t,k)$ is compared to a threshold δ yielding a binary speech-presence probability estimate $\bar{p}(t,k)$, which is further smoothed over time with a smoothing factor α_p as

$$p(t,k) = \alpha_p \bar{p}(t-1,k) + (1-\alpha_p)\bar{p}(t,k). \tag{2.6}$$

After computing the smoothed speech-presence probability $p(t,k)$, the time-frequency-dependent factor $\alpha_d(t,k)$ is calculated using Equation 2.4, and the noise spectrum estimate is updated using Equation 2.3.

2.6.2 Non-negative matrix factorization

Non-negative Matrix Factorization (NMF) is a linear decomposition technique that attempts to approximate an input non-negative matrix V as a product of two non-negative matrices, i.e.

$$V_{F\times T} \approx W_{F\times R} \cdot H_{R\times T}, \tag{2.7}$$

where $R \leq F$ controls the rank of the approximation. In audio signal processing, NMF is typically applied to the magnitude spectrogram of the signal, and F and T correspond to the number of frequency bins and number of frames, respectively. The columns of W are usually referred to as bases, and the rows of H are their corresponding weights or activations in time.

The problem of minimizing the divergence between the input matrix and its approximation needs to be solved:

$$\arg\min_{W,H} D(V||WH) + \lambda|H|_1 \quad W,H \geq 0 \tag{2.8}$$

where D is a cost function (in this work, the Kullback-Leibler divergence), and the parameter $\lambda \geq 0$ is used to impose a sparsity constraint on the activations, thus favouring solutions with fewer bases activated at a given time.

In the supervised NMF approach the bases matrix is trained beforehand on the training data. In the general case, when S sound sources are considered, a bases matrix is trained for each source separately and a global bases matrix is constructed via concatenation $W_{train} = [W_1; ...; W_S]$. At the source separation step the bases matrix is fixed and only the activations matrix is estimated $H = [H_1; ...; H_S]$.

2.7 Chapter summary

In this chapter, the task of AED in real-world environments was discussed, and the state of the art, in particular of feature extraction, classification algorithms, source separation techniques, in the wider area of audio recognition was reviewed. Also, the work already done in analysis of the acoustic environment of a preterm in a NICU was presented.

Chapter 3

Database acquisition and annotation

3.1 Chapter overview

This chapter contains the description of how the database used in this thesis work was produced. In Section 3.2 we provide details of the recording setup, specify the selected acoustic scenarios and describe the recording sessions carried out. In Section 3.3 the protocol used for audio data annotation is defined and in Section 3.4 information about the produced labelling is given. Since several independent annotators were involved in the annotation campaign, in Section 3.4 we also provide a rough estimation of consistency of the acquired labels.

3.2 Audio data acquisition: recording setup, sessions and scenarios

The acoustic analysis and the experimental evaluations presented in this thesis were performed using the real-world audio recordings made in the Neonatal Intensive Care Unit (NICU) of Hospital Sant Joan de Déu. It is a level III NICU that accommodates 42 places. In particular, the recordings were carried out in a room designated for intensive care of very preterm newborns, i.e. the ones born before 32 weeks of gestation. It is a rectangular room equipped with 4 incubators (see Figure 3.1), which covers an area of 35 m^2 and is attended by two nurses 24 hours a day. The room is limited on top by a corridor for the medical staff passage with the door depicted in Figure 3.1 at the top left corner; the bottom wall isolates from the hall; the left wall is the border with another NICU room; and the right part of the room is open, adjoining the nursing station and the door for the family members entrance.

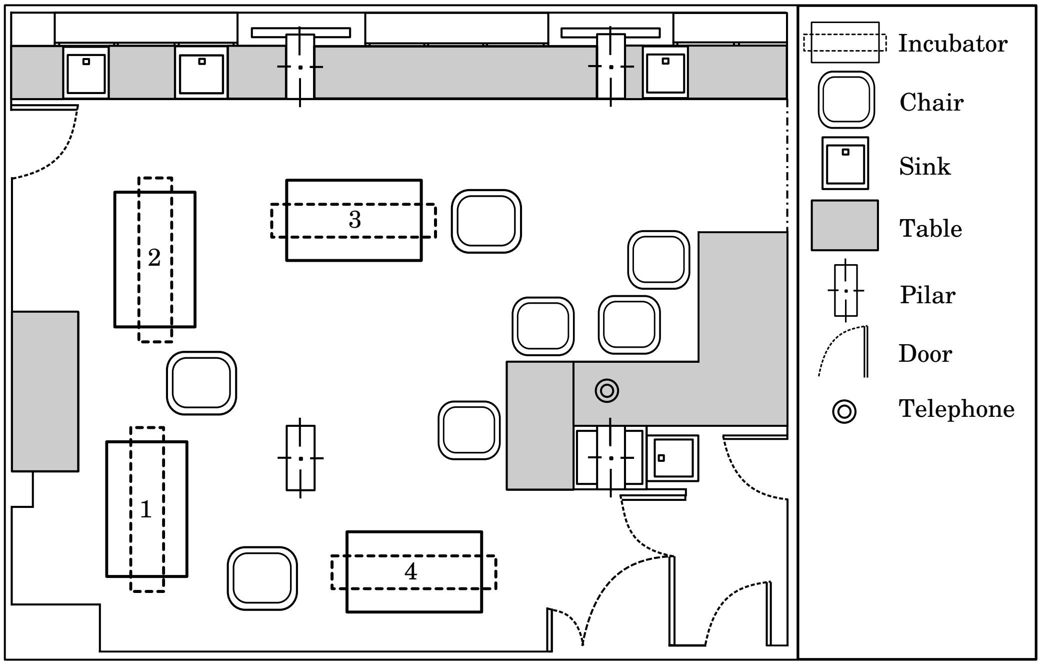


Figure 3.1: The layout of the NICU room with the four positions of the incubators and other equipment

The whole set of recordings includes 18 recording sessions that capture the NICU environment in different scenarios. Only 10 sessions were included in the database for consistency reasons; the other 8 sessions had either exploratory character or did not follow the recording protocol completely. The information about all the carried out recording sessions can be found in Appendix A. It should be mentioned that another database of continuous acoustic environment recordings (that also includes video recordings of the cardiorespiratory monitor screen) was recorded, but not annotated. The details of that database are given in Appendix B.

Table 3.1 provides information about the ten recording sessions included in the database used in this work, which were carried out both in the morning and in the afternoon. The overall duration of the acquired audio data is 108.7 minutes. In each recording session the incubator position was chosen under the following conditions:

1) The preterm infant did not have any malformations at the time of birth and was at least one week old.
2) The preterm infant was clinically stable (without infections or cardiorespiratory instability), did not need mechanical ventilation and was not given medications depressing the nervous system.
3) The consent of the preterm's parents for audio acquisition was obtained.

Table 3.1: Information about the recording sessions included in the database

Part of the day	Session code	Time	Incubator	Duration (s)
Morning	RS4	13.00	1	481.28
	RS11	09.25	4	1030.9
	RS12	13.00	3	581.87
	RS15	09.05	4	659.76
	RS16	09.15	4	804.03
Afternoon	RS3	15.30	3	587.16
	RS13	17.10	2	683.58
	RS14	17.25	4	918.19
	RS17	17.00	3	658.63
	RS18	17.20	2	429.03

Two electret unidirectional microphones connected to the Olympus LS-5 Linear PCM Recorder were used to make recordings. One microphone was placed inside the incubator, close to the infant's ear, and the other one outside the incubator, at approximatively 50 cm distance above it, as shown in Figure 3.2. The recording protocol was created and shared with the medical staff taking part in the recording sessions, where the positions of the microphones, recording device settings and guidelines for the recordings were specified.

Obviously, the amount of activities that take place in the NICU can be very large. The list of the most common of them is given below:

- *Nursery care*
- Visit of parents
- Kangaroo care
- Visit of a specialist (*paediatrician*, cardiologist, etc.)
- Examinations (X-ray, ultrasound, etc.)
- Preterm's entry or relocation
- Preterm's death
- Surgical interventions
- Music therapy
- Cleaning

A set of acoustic scenarios, which mostly correspond to the daily nursery care related activities (marked in italic in the above presented list), was selected for recording. Approximately every 3-4 hours every preterm in the NICU receives a nursery care. There is a standard list of interventions that

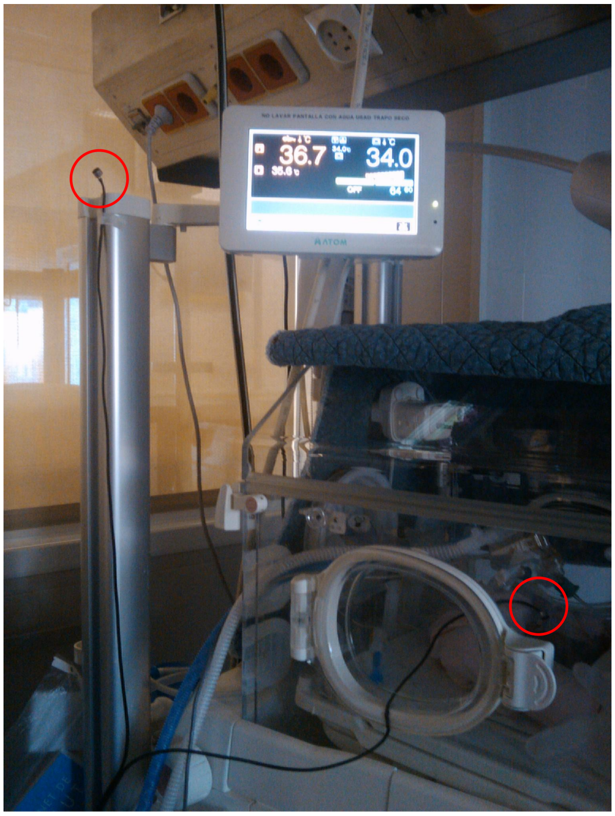


Figure 3.2: The conventional microphone positions used throughout the recording sessions

should be done to a preterm infant, which may vary depending on the infant's needs, and includes: changing a diaper, measuring blood pressure, changing an oxygen sensor, cleaning respiratory secretions, measuring temperature, changing temperature sensors, weighting a newborn, paediatric observation, changing medications. Each of these nursery care operations is considered as a scenario. Also, a *neutral* scenario was defined for including the time periods when no nursery care scenario takes place, the baby is untouched and the doors of the incubator are closed. It should be mentioned that, except for nursery care, the activities from the list above can take place during this generic *neutral* scenario.

Every recording session contains a subset of the defined scenarios. Each scenario was recorded to a separate audio file, and in average 8.6 recordings per scenario were acquired. Except for *neutral*, the possibility to make a recording of the concrete scenario during the recording session depended on preterm's individual needs. Besides, the weighting of a newborn was usually performed in the afternoon. The duration of each scenario varied from session to session, which depended on ad-hoc variations of the procedure and on the work style of a nurse performing it. The duration of the *neutral* scenario was controlled, and usually a 1-2 minute recording was acquired. The detailed information about the scenarios recorded in each session can be found in Table 3.3.

3.3 Labelling protocol: list of acoustic events and criteria

The audio data was manually annotated using the ELAN tool [123]. For each relevant audio event it's time boundaries and identity (label) were specified. For each scenario, a single annotation was produced, which corresponds to the audio recorded both inside and outside the incubator.

All the labelled acoustic events were grouped into several tiers of annotations:

- *Alarms* for alarm sounds generated by various biomedical equipment.
- *Vocalizations* for all the sounds produced through a vocal tract, either by infant or adult.
- *Events* for the rest of the relevant acoustic events, which are mainly the events specific to the considered scenarios.
- *CPAP*, a separate tier for the Continuous Positive Airway Pressure (CPAP) noise.
- *Noises* for the acoustic events that were not specified in the labelling protocol, but which annotators considered relevant and perhaps recognized. All the events in this tier have a special label *nn* followed by (if recognized) a label in square brackets.

Table 3.2 contains a list of acoustic events considered for annotation, their corresponding labels and the criteria used for annotation. The annotation campaign was performed in two stages, where the first portion of labels was obtained in an exploratory work during summer 2013 and later augmented during summer 2014, and the two annotators worked independently. The list of considered acoustic events defined for the first stage of the annotation campaign differs from the definitive list presented in Table 3.2. These differences are outlined in Appendix C. At the second stage, the annotation was limited to the two most relevant types of sounds (i.e. equipment alarms and vocalizations), for which the automatic detection systems were developed, and some resembling events (buttons). This was done in order to speed up the annotation process and to obtain as much annotated acoustic event samples as possible.

The general labelling guidelines shared with annotators were the following:

1. Only the defined labels (see Table 3.2) must be used for annotating the relevant acoustic events.
2. The annotation should be primarily based on the audio acquired using the microphone placed outside the incubator, since it is closer to the sound sources. The audio obtained inside the incubator should be consulted in case of ambiguity or doubt, as well as used for verification of the complete annotation.
3. The spectrogram must be checked during the annotation process.
4. It is better to label more than less, i.e. the timestamps should not be too narrow, to avoid possible cutting of an acoustic event.
5. In case of several simultaneous events belonging to one tier, an additional tier should be created with the name [{TierName}{N}], where *TierName* is one of the defined tier names, and N – is the index number of the tier (e.g., tiers Voices1, Voices2, Voices3, etc).
6. In case of doubts about the specific class of an acoustic event, a more generic or a more common

label should be used. I.e., confusing vocalization events should be labelled as “background voices” (*bv*), and for alarms and other sounds generic *al* and *nn* labels used, respectively.

7. For the periodic events each period should be labelled separately. For instance, in case of alarms every alarm signal interval should be annotated.

The criteria used during the annotation process, which are provided in Table 3.2, were initially defined for some of the confusing events and complemented by the annotators during the work. The labelling was produced by hearing audio samples and guided by spectrogram observations. Still, in some particularly difficult cases (i.e. some blurred sounds) it was not possible to clearly follow the defined criteria.

Note that for alarms a source device name is specified as an acoustic event name in Table 3.2 and it may be not unique. To assist the annotation of alarms, an exemplary sample of each alarm class was extracted and provided to annotators. The annotators were also instructed to obtain such sample for every new alarm class found. Also, the document containing the basic information about the alarm classes was shared, in which the tonal structure (i.e. melody), duration of signal and silence intervals in an alarm period and major frequencies were described.

Table 3.2: List of labelled acoustic events and corresponding annotation criteria

Tier	Label	Acoustic event	Criteria
Alarms	a1	Monitor Philips IntelliVue MP30	• The timestamps are easier to be decided at 1.465 kHz. If not possible, based on any other harmonic visible on a spectrogram, paying more attention to listening.
	a2	Ventilator Infinity C500 Dräger	• The entire alarm signal interval (4 higher beeps and two lower beeps) should be labelled. • If possible, the beginning is decided at 1 kHz and the end around 0.8 kHz.
	a3	Incubator Atom	• The beginning is decided at 0.665 kHz and the end at 0.54 kHz.
	a4	Respirator of non-invasive ventilation	
	a5	Incubator Atom	
	a6	Respirator Babylog Dräger	• The timestamps can usually be clearly decided at 2.4 kHz.

Alarms	a7	Monitor Philips Agillent V24C	• The beginning timestamp is clear at 2.9 kHz. • The silence interval is very short, but is approximately 20-30 ms. • If not followed by another period of the same alarm, the end timestamp is decided based on listening and spectrum.
	a8	Monitor Philips Agillent V24C	• If possible, the beginning is decided at 3.5 kHz and the end at 4.5 kHz.
	a9	Thermometer	• Conventionally, the device produces the alarm for 10 periods, but may be stopped before.
	a10	Infusion pump Alaris GH Plus	• It is easier to decide about the timestamps at 1.14 kHz.
	a11	Infusion pump Alaris GH Plus	
	a12	Respirator	• The timestamps can be based on the fundamental frequency 2.3 kHz. • Each signal interval consist of 5 tones.
	a13	Incubator Kaleo	
	a14	Incubator Atom	
	a15	Infusion pump Alaris GH Plus	
	a16	Monitor Philips IntelliVue MP70	
Vocalizations	fv	Foreground voices	• Any kind of speech (except shouts) that is close to the microphone. • A distinct voice over background voices or babble having a high intensity.
	bv	Background voices	• Speech that is far from the microphone, but the content is understandable. • Babble, possibly coming from several people talking at a time.
	bc	Baby crying/voice	• All sounds coming from preterm infants. • Depending on whether the vocalization was produced inside or outside the incubator, the decision about the timestamps should made based on the audio acquired, correspondingly, inside or outside of it. • A special label *bci* should be used for the infant vocalizations heard only inside the incubator.

	co	Cough	
	sh	Shout	• Short loud speech produced either close or far from the microphone.
	lg	Laughter	
Events	bi	Buttons of the infusion pump	• Each sound should be labelled separately.
	bw	Buttons of weights	• The timestamps can be decided at approximately 0.45 kHz.
	nn[bm]	Buttons of the monitor	
	nn[bp]	Buttons of the infusion pump	• The timestamps are easier to be decided at 3.55 kHz.

Information about the amount of labelled alarm and vocalization samples can be found in Appendix D. Note that some of the alarm classes (namely, a4, a5, a13, a14 and a15) were found in the NICU environment, but were not present in the annotated files.

The resultant annotation file, which has an XML-like structure, was further converted to a CSV format for each considered detection task.

3.4 Produced annotations

As mentioned before, the labelling production was performed in two stages, and later a revision stage was needed in order to correct the labels of equipment alarm and vocalization sounds and to make them follow a unified protocol (see Appendix C for more details on the protocol changes). In particular, at the revision stage the labelling was reviewed for adding the annotations of the classes not considered in the labelling protocol before, for changing the labels that were assigned to a wrong class, and for removing labels of some acoustic events with very low signal-to-noise ratio. A label timestamp was modified only in case it was clearly inaccurate.

During the two annotation stages the audio database was labelled only partially. Table 3.3 provides an overview of the acquired and annotated acoustic scenario samples, where each cell contains the duration (in s) of the corresponding audio file. Table 3.4 provides information about the amount and type of labelling produced by each annotator. In total, for the two major types of sounds, the amount of annotated data is the following: alarm sounds are annotated in 47 files (which is 54.7% of total files in database or 54.3 minutes) and vocalizations are annotated in 35 files (40.7% of files or 40.2 minutes).

Table 3.3: Information about the acquired scenario samples, their duration (in s) and the produced labelling

Scenario code	Scenario	Recording session									
		RS3	RS4	RS11	RS12	RS13	RS14	RS15	RS16	RS17	RS18
AS1	Changing a diaper	🟩🟦 35.67	⬛ 81.08	118.56	114.73	99.13	168.44	76.35	95.16	🟨 96.06	–
AS2	Measuring blood pressure	🟩 59.56	45.14	140.64	65.48	🟦 98.01	81.92	🟦 112.64	170.81	75.79	–
AS3	Changing an oxygen sensor	🟩 34.34	🟨 64.44	83.94	–	–	71.54	89.03	🟦 60.81	48.13	–
AS4	Cleaning respiratory secretions	🟦 36.71	77.46	🟨 119.26	🟦 43.47	150.81	–	–	–	–	🟦 145.8
AS5	Measuring temperature	87.14	🟦 61.16	🟨 80.39	🟨 57.89	60.12	135.77	🟦 59.77	161.19	105.12	65.76
AS6	Changing temperature sensors	44.09	🟦 31.21	61.09	🟨 83.66	–	🟨🟦 153.88	131.31	–	118.14	–
AS7	Weighting a newborn	⬛ 47.58	–	95.78	–	🟨 26.61	🟨 35.32	–	–	🟦 73.77	–
AS8	Paediatric observations	🟦 190.03	🟦 22.36	155.55	60.46	🟨 81.29	118.35	🟨 93.27	132.01	–	62.62
AS9	Changing medications	52.04	31.42	53.29	🟦 35.39	44.03	27.79	🟦🟨 35.25	🟨 60.4	🟦 13.31	🟦 33.37
AS10	Neutral	–	🟩 67.01	🟩 122.39	🟩 120.79	🟩 62.07	🟩 61.58	🟩 62.14	🟩 61.51	🟩 65.41	🟩 60.81
		–	–	–	–	🟩 61.51	🟩 63.6	–	🟩 62.14	🟩 62.9	🟩 60.67

Table 3.4: Information about the labelling produced by each annotator

Annotator	Annotated sounds	Time (s)	Number of files
I	🟩 All sounds	1124.1	17
I	🟨 Alarms	987.72	13
II	🟦 Alarms, vocalizations	1242.61	18
S	⬛ Alarms	47.58	1
S	⬛ Alarms, vocalizations	81.08	1

Note that the annotator names I and II correspond to the first and the second annotation stages, respectively, and Annotator S (the author of this thesis work) was supervising their work and carried out the corrections. In general, Annotator I was working standalone guided by the labelling protocol and the auxiliary information provided, and the subsequent corrections primarily affected the labelling obtained at this stage. At the second stage the labelling was produced in closer collaboration between the Annotator II and the Annotator S, and it was included in the database with slight changes. Alarm annotations were to a larger extent produced by Annotator I (57.91% of total time annotated for alarms). Vocalization annotations were produced more or less in equal proportion by both annotators (Annotator II labelled 51.52% of total time annotated for vocalizations).

As can be seen from Table 3.3, some randomly chosen audio files (namely, *RS3_AS1*, *RS14_AS6*, *RS15_AS9*) were intentionally given for labelling to both Annotators I and II. The aim was to estimate the labelling consistency (or, in other words, to measure an interobserver agreement) for alarm and vocalization sounds, which can give a notion about the upper bound of the detection systems performance. From this set of files, vocalization sounds were labelled by both annotators only in *RS3_AS1* recording, which is 1.48% of total time with vocalizations annotated. Concretely, in this file there is only one vocalization event, a background voice *bv*, which in the final version of annotations spans an interval of 2.43 s. On the other hand both annotators labelled alarms in all the abovementioned files, and it corresponds to 6.74% of total time with alarm sounds annotated. Only several alarm classes were present in these recordings, namely, a1, a3, a7, a10 and a16. It should be noted that a16 alarm was labelled as belonging to a1 class by the Annotator I. The final database contains the labelling of these three files made by Annotator II.

The labelling consistency was measured using the frame-level Missing Rate (MR) and the False Alarm Rate (FAR) metrics that were used during the development of the detection systems. These metrics are defined as

$$MR = \frac{N_M}{N_T}, \qquad FAR = \frac{N_{FA}}{N_{NT}}, \tag{3.1}$$

where N_T and N_{NT} are the total number of frames for target and non-target class (e.g. alarm and non-alarm), respectively; and N_M and N_{FA} are the number of misclassified frames for target and non-target class, respectively. Note that the frame and frame shift durations were different for the two types of sounds: for alarms these values were set to 85.3 ms and 42.6 ms, respectively, and for vocalizations they were equal to 30 ms and 10 ms, respectively.

Table 3.5 provides the consistency measures in terms of metric scores for the files labelled by both annotators. These scores are obtained by evaluating the labels of Annotator II with regards to the labels of Annotator I, which serve as a reference. Since the amount of data used for this estimation is small, the obtained conclusions are not significant. Also, the labels from Annotator I were changed during the correction stage, so in fact the presented metric scores don't reflect the consistency of the final database. Still they can be viewed as a rough estimation and, hopefully, the consistency was improved after corrections.

Table 3.5: Estimation of the produced labelling consistency in terms of metric scores (in %) for the two types of sounds annotated

Recording code	Alarms		Vocalizations	
	MR	FAR	MR	FAR
RS3_AS1	10.36	2.91	0.41	0.03
RS14_AS6	17.26	1.69	–	–
RS15_AS9	20.34	1.30	–	–
Average	15.99	1.97	0.41	0.03

In fact, the labelling of Annotator II doesn't contain 9 out of 50 alarms for *RS14_AS6* and 7 out of 11 alarms for *RS15_AS9*, with respect to the labelling of Annotator I. These differences are attributable to two factors. First, Annotator II did not include the labels due to a too low energy of the sound. And, second, almost all the alarm sounds labelled by Annotator I in file *RS15_AS9* as belonging to class a11 were the buttons of the infusion pump (*bi*). Therefore, for metric scores calculation these alarms were removed from consideration in labelling of Annotator I (as was done, in general, during the correction stage), and the scores reported in Table 3.5 are obtained after these modifications. Without removing such alarm labels, the MR metric scores calculated for the recordings *RS14_AS6* and *RS15_AS9* are equal to 37.81% and 55.05%, correspondingly.

The consistency measures for alarms were also calculated in terms of the event-level metric, that was defined to present the performance of detection systems in a way more meaningful for the medical application. The details of this metric, which is called Period-Based ERror Rate (PB-ERR), are provided in Section 5.3. In terms of this metric, for all the alarm classes the error scores were equal to **0%**.

3.5 Chapter summary

In this chapter, all the steps of the database production, from acoustic scenarios definition and audio acquisition to annotation production, were reviewed in detail. During the work on the database production, key specifications (e.g. recording setup and guidelines, labelling protocol) were designed and the whole framework of the audio database production for the NICU environment was set up. Due to the pioneering character of the work in that acoustic environment, the abovementioned specifications had to be designed from scratch in close collaboration with medical and engineering staff. Several rounds were required, which allowed refinement of these specifications based on the obtained experience, but also implied more effort for making possible that the produced database follows a unified protocol.

A number of recording sessions were carried out in the NICU following the designed guidelines. A part of these recordings formed the produced database. In total, it includes more than 1.5 hours of audio data, which corresponds to 86 samples of the defined acoustic scenarios. Not taking into account the pilot recording sessions, the audio acquisition process lasted for about 4 months.

The annotations cover roughly half of the database. They were obtained in two stages, and posteriorly revised. The first stage had an exploratory character, where both initial experience and more specific knowledge about the NICU acoustic environment were gained. The second stage used a refined labelling protocol and led to better defined criteria for acoustic events labelling. Eventually, the produced database contains labelling mainly for the two major types of sounds (namely, equipment alarms and vocalizations). It should be noted that database annotation required a lot of effort. It was the most time-consuming part of the database production and took about 6 person-months of work in total, with last revisions taking place in October 2015.

Chapter 4

Acoustic description

4.1 Chapter overview

A typical Neonatal Intensive Care Unit (NICU) environment is acoustically very rich and may be characterized by a large diversity of sounds coming from numerous sources. This chapter presents the results of the acoustic description of that environment carried out from a set of audio recordings. In particular, the first exploratory recording sessions RS1 and RS2 (see Appendix A) as well as recordings from the acquired database were used, and, in case of alarm analysis, also recordings made in a quiet room.

This chapter is organized as follows. Section 4.2 presents the list of acoustic events found in the NICU environment, whereas Section 4.3 provides an analysis of the acoustic scenarios from the acquired database. In Section 4.4 a sound taxonomy for NICU sounds is proposed, which is followed by the description of the most typical sounds from the NICU environment in Section 4.5.

4.2 List of acoustic events

During our study more than 60 different acoustic events happening in the NICU environment were found. It should be mentioned that this number was surprisingly high even for the medical staff working in the unit. Table 4.1 provides an extensive but not exhaustive list of the acoustic events found.

Table 4.1: List of NICU acoustic events

N	Acoustic event	Label
1-16	Equipment alarms	a1-a16
17,18	Buttons of infusion pump	bi, nn[bp]
19	Buttons of the weights	bw
20	Buttons of the monitor	nn[bm]
21	Foreground voice	fv
22	Background voice (babble)	bv
23	Baby crying/voice	bc
24	Cough	co
25	Shout	sh
26	Laughter	lg
27	Incubator door opening/closing	od/cd
28	Doors slam	nn[do]
29	Telephone ring	nn[te]
30	Mobile phone	nn[ma]
31	Chair moving	nn[ch]
32	Knock	nn[kn]
33	Step	nn[st]
34	Respiration noise	nn[tr]
35	Continuous Positive Airway Pressure (CPAP)	nc
36	Paper work	pw
37	Squeak	nn[xi]
38	Secretions cleaning	cs
39	Spray	nn[sp]
40	Glass jingle	gc
41	Plastic wrapping	pl
42	Drawer	nn[dr]
43	Diaper	dp
44	Click of infusion pump	cb

45	Hissing	hi
46	Jingle	jg
47	Metallic tray stroke	ms
48	Plastic hit	ph
49	Folding paper bags	fb
50	Keyboard typing	kt
51	Dragging object	so
52	Shifting with wheels	sw
53	Putting on/off sphygmomanometer	ab
54	Water running	wr
55	Cleaning hands	nn[clh]
56	Using paper towels	pt
57	Click	cl
58	Clap	cp
59	Putting on rubber gloves	rg
60	Windows OS error sound	nn[xp]
61	Music	mu
62	Taking sensor off	ts
63	Object on incubator	ob
64	Cloth rustling	ct

The acoustic events listed in Table 4.1 are roughly the sounds produced by the human body, by objects handled by humans or by equipment. Some of the acoustic event types, like mobile phone (nn[ma]), music (mu) or knocks (nn[kn]), are acoustically very broad and may contain various realisations.

4.3 Description of acoustic scenarios

This section provides the description of a set of scenarios from the audio database used in this thesis work (see Chapter 3). Apart from the *neutral* scenario (AS10), which denotes the periods of time when the preterm infant is untouched and the doors the incubator are closed, that set comprises of nursery care related scenarios, namely: changing a diaper (AS1), measuring blood pressure (AS2), changing an oxygen sensor (AS3), cleaning respiratory secretions (AS4), measuring temperature (AS5), changing temperature sensors (AS6), weighting a newborn (AS7), paediatric observation (AS8), changing medications (AS9). The respective codes of scenarios are provided in round brackets.

Regarding Sound Pressure Level (SPL) measurements, for all the scenarios the average SPLs obtained from recordings [1] acquired inside the incubator are higher than the ones outside the incubator.

[1] The calculation of SPL values from the recorded audio is discussed in the Technical report HSJD-UPC-5-2015.

Depending on the scenario, this difference may be from 1.65 to 3.61 dBA, where the boundary values of that range correspond to measurements obtained, respectively, for the *neutral* and the paediatric observation scenarios. It might be related to sound reverberation inside the incubator (note that similar results were obtained in [17]). On the other hand, no significant differences in SPL measurements were observed between sessions recorded in the morning and in the afternoon.

Most sounds (like steps, door slam, vocalizations) from Table 4.1 are common to all the scenarios and, in principle, can happen at any period of time. But some of the acoustic events are specific to the scenario in which they occur, namely, they are happening under certain conditions. Table 4.2 shows the specific sounds for some of the scenarios from the recorded database.

Table 4.2: Scenario-specific acoustic events

Scenario	Specific events
AS1 Changing a diaper	Diaper (dp)
AS2 Measuring blood pressure	Putting on/off sphygmomanometer (ab), rhythmic hissing (hi), squeak (nn[xi])
AS3 Changing an oxygen sensor	Taking sensor off (ts)
AS4 Cleaning respiratory secretions	Plastic wrapping (pl), secretions cleaning (cs)
AS5 Measuring temperature	Alarm of a thermometer (a9)
AS7 Weighting a newborn	Buttons of the weights (bw), diaper (dp)
AS9 Changing medications	Buttons (bi), click of infusion pump (cb)

As was observed, scenario-specific events are present in some scenarios but neither are they present in the same scenarios along different recording sessions nor their presence is significant enough in comparison to other acoustic events.

The spectral characteristics of the different scenarios were analysed by observing the distribution of the average energy along frequency sub-bands. In total 24 sub-bands were applied to the original 44.1 kHz recordings. The audio data was pre-processed: the DC component was removed and an energy normalization was performed. Figure 4.1 shows the distribution of energy along sub-bands for different scenarios in a given recording session, and Figure 4.2 shows this information the other way round, namely, the distribution of energy along sub-bands for a given scenario in different recording sessions.

It can be seen that the considered scenarios are very similar acoustically and are really hard to distinguish. Figures 4.1 and 4.2 indicate that the difference between recording sessions is more significant than the difference between acoustic scenarios. The main reason for that strong similarity of the acoustic scenarios from the same recording session lies in the presence of a distinct session-specific equipment noise (see Section 4.5.3), which is present throughout the audio recordings acquired in a recording session.

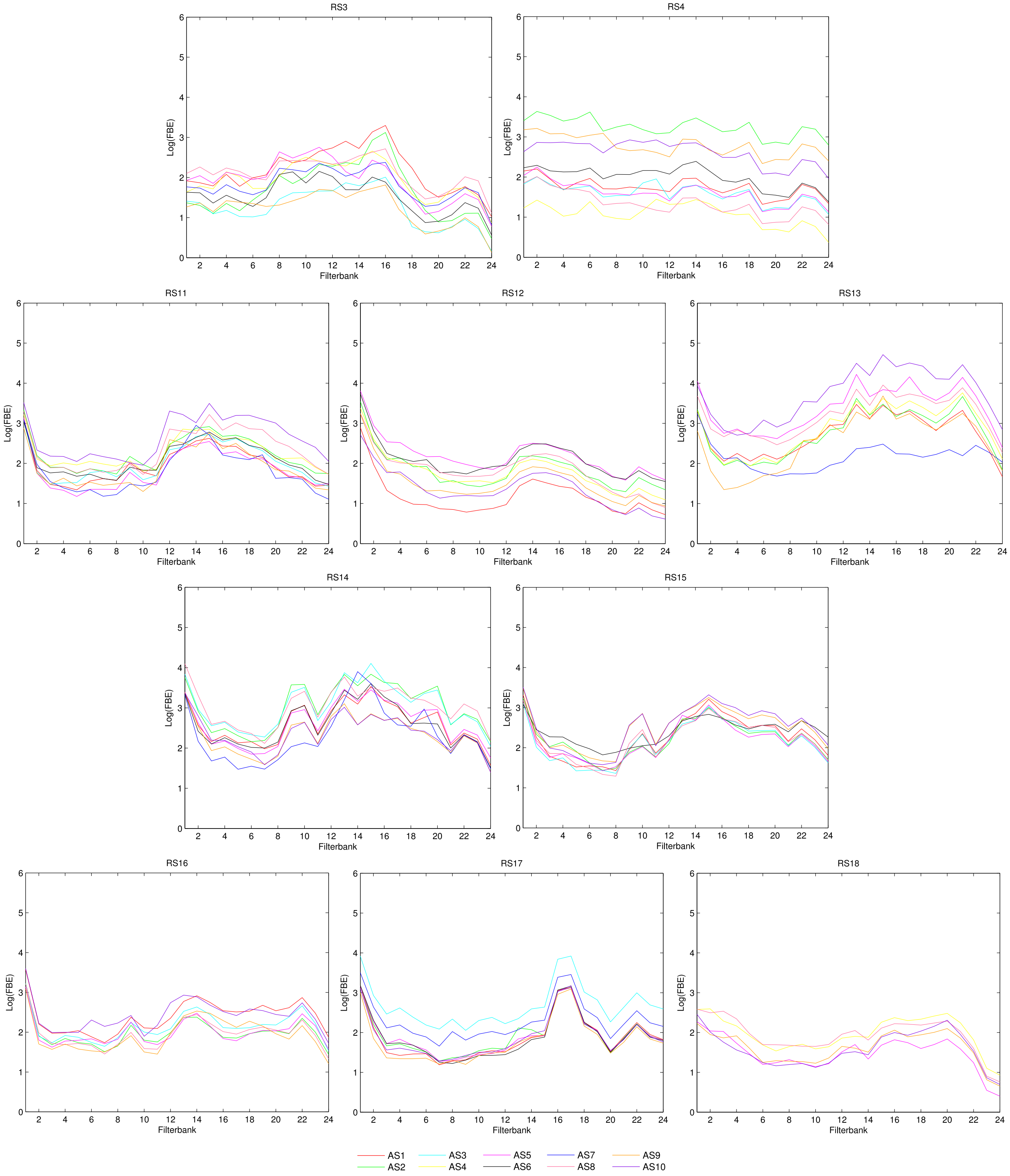


Figure 4.1: Average logarithm filterbank energies for the scenarios in a given session.

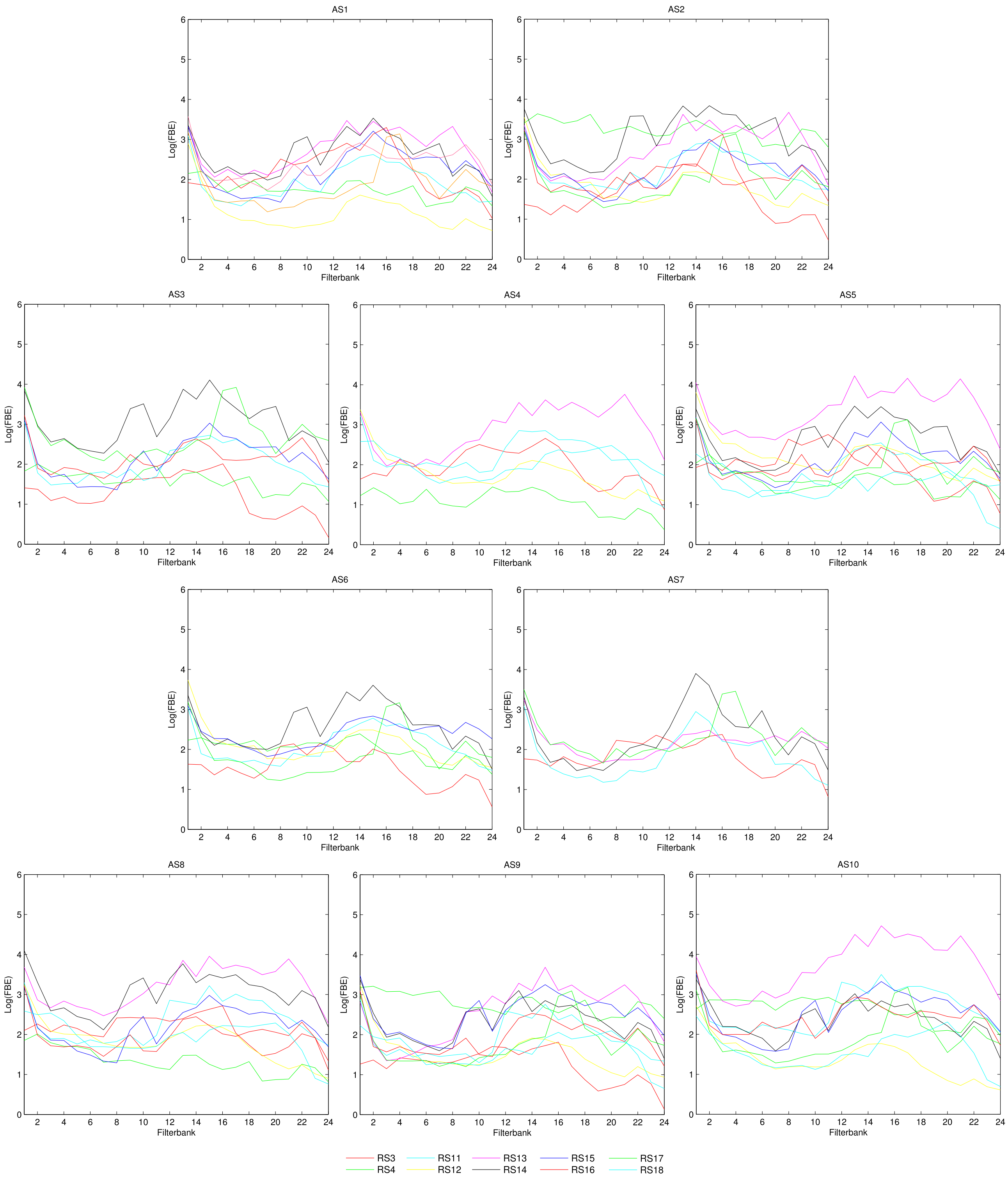


Figure 4.2: Average logarithm filterbank energies for a given scenario in different recording sessions.

Apart from the presence of equipment noise, there are other factors that may contribute to the variability of a scenario acquired in different sessions. These factors are outlined below, grouped by major sources of variability:

I. Recording device and microphones:
   a. Variations in microphones position and direction
   b. Recording device setup (mostly concerns recording volume, which was controlled by a dial)

II. Equipment and its position:
   a. Incubator type
   b. Incubator position (if close to other incubators, to working desks, drawers, sink, doors, etc.)
   c. Accompanying equipment (i.e. monitor, type of ventilation equipment, infusion pumps); also, the neighbouring incubators, as the sounding alarms may be different.

III. Time period of the recording session:
   a. Part of the day
   b. The activities it coincides with (i.e. attending hours, parents visit, surgical intervention, etc.), which also influence the number of people present in the unit room and their behaviour.

IV. Nursing work, which influences the sequence of actions in each sample of the scenario:
   a. Ad-hoc variations based on the particular preterm infant needs
   b. Organization of the working place
   c. Style of work
   d. Number of nurses attending a preterm (i.e. in recording session RS17 there were two nurses working together)

4.4 Sound taxonomy

The sounds found in the NICU environment have diverse spectro-temporal characteristics. For instance, regarding the time dimension, we observe sounds which are continuous (like chair moving or drawer) or impulsive (like knocks, steps or door slam); sounds which are periodic (like alarms or CPAP noise) or aperiodic (like spray or plastic wrapping). By building a taxonomy we try to structure the whole diversity of observed sounds into homogeneous groups.

On the other hand, there is a necessity to limit the number of acoustic event classes considered for automatic detection, and one way of doing so is by providing a sound taxonomy. The development of the sound taxonomy helps to better understand the data domain [124], increase the accuracy and speed of classification [125]. Obviously, a sound taxonomy is subjective and very dependent on the chosen classification domain.

Figure 4.3 presents the sound taxonomy proposed for NICU sounds. Both acoustical and semantical criteria are used, and sounds are joined in acoustical groups and semantical categories, where the semantical categories are denoted in italic. The leaves of the taxonomy tree show examples of sounds belonging to a particular group or category.

In general, the set of all the sounds can be split into three major groups, which possess different acoustical properties: tone, vocalization and other.

Tone (in music, note) denotes a sound of distinct fundamental frequency and duration. The *tone* group mostly comprises equipment alarms, and can be divided into two subgroups regarding the presence or absence of a long-term periodicity in time, i.e. by whether there are repetitions of a basic sound over identical time periods or not. From a semantical point of view this group contains informative sounds produced by devices.

The *vocalization* group includes all the sounds that are produced through the vocal tract, either by infant or adult. In this group, three semantical categories are distinguished: adult speech (i.e. foreground and background voices), infant vocalizations (mostly cries) and non-speech adult vocalizations (like cough or laughter).

The sounds which cannot be referred to the two previous groups are assigned to the *other* group. Acoustically this group is even more diverse, and with regards to the spectral domain we divide it into different subgroups: lower- ($\leq$ 3 kHz) and higher-frequency sounds, and sounds with their content spread over a wide frequency range.

As was commented in Chapter 3, some audio recordings were labelled completely during the first exploratory stage of database annotation (in Table 3.3 these scenarios are marked in green). Based on the analysis of these recordings, we observed that tones, vocalizations and other sounds are present in 20.59%, 63.55% and 70.14% of time, respectively.

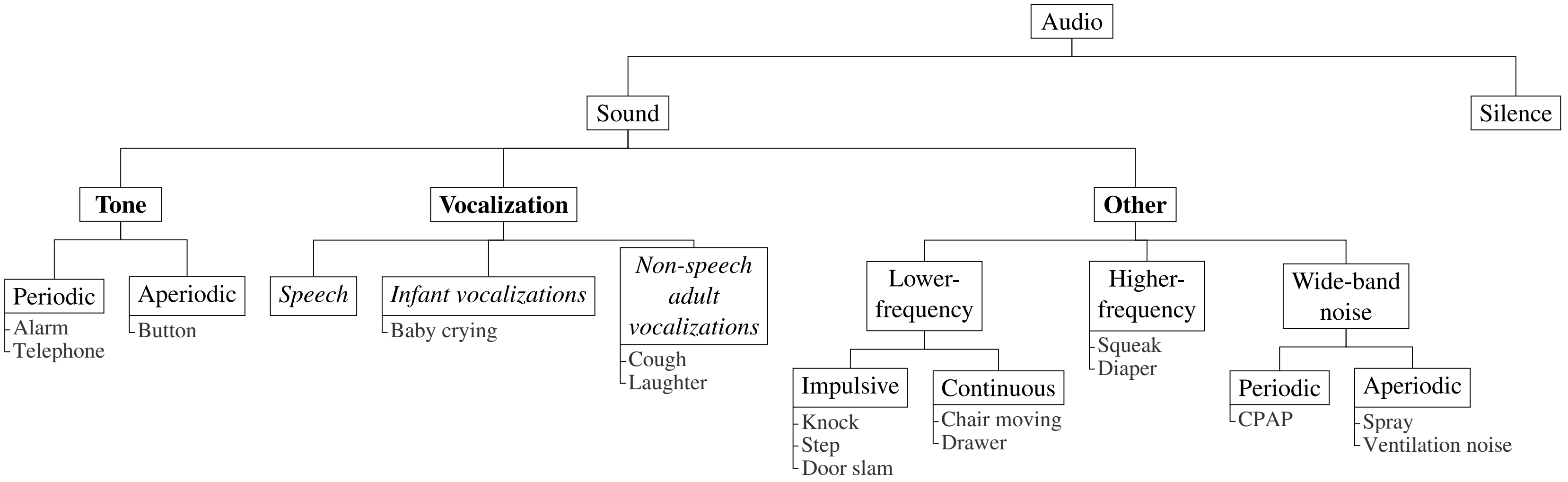


Figure 4.3: A general sound taxonomy of a typical NICU.

4.5 Analysis of typical types of sounds

4.5.1 Equipment alarms

Equipment alarms are extensively present in a NICU environment and are used in monitoring or supporting equipment to alert of situations requiring medical attention. By observing the recorded audio data we have found 16 different types of alarms happening in our NICU and coming from:

1) cardiorespiratory monitors – 4 types of alarms (65.7% of samples),
2) incubators – 4 (9.1%),
3) ventilators – 4 (15.6%),
4) infusion pumps – 3 (7.5%),
5) thermometer – 1 (2.1%).

It should be noted that although the set of observed alarms is quite representative of the NICU environment, it is not exhaustive and more alarm classes can be found. It can be seen that most alarms are generated by monitors and ventilation devices.

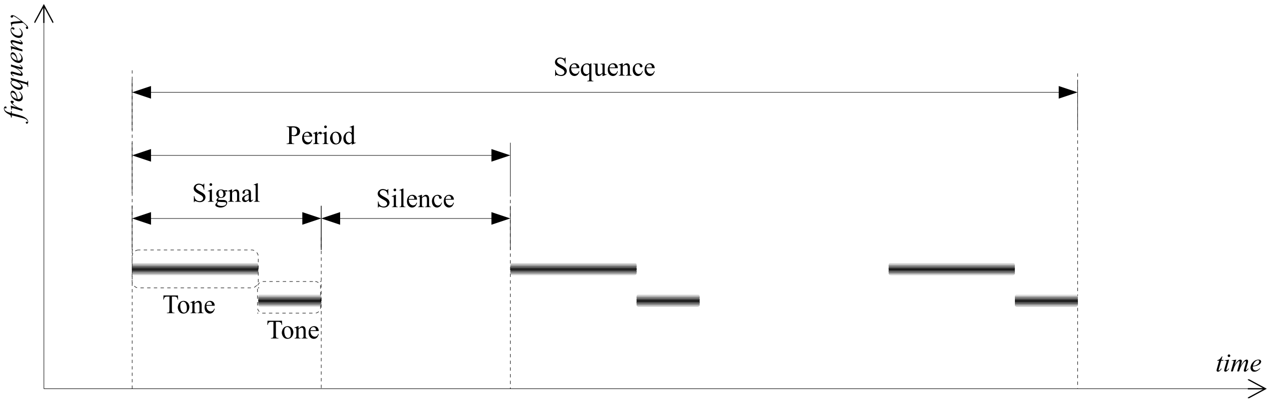


Figure 4.4: Graphical description of terms used to denote particular alarm properties. Only the fundamental frequency is depicted for clarity of presentation.

Generally, the acoustic properties of the observed alarms can be described as:

1. They reveal periodicity in time. Each alarm period consists of signal and silence intervals of established durations (see Figure 4.4). The period duration of 13 alarm types is from 0.45 to 4.25 s, while other types are up to 15.3 s long.
2. The signal interval may consist of one or several consecutive stationary signals (tones): only one tone (7 alarm types); several repetitions of the same tone, possibly of different duration (5 types); or several different consecutive tones (4 types).
3. Each tone contains one or several simultaneous frequency components, which may or may not be harmonically related.

The particular characteristics of each of the alarm classes are presented in Table 4.3. The alarm-specific frequencies, and the signal and silence interval duration parameters were obtained from the recorded database and, in part, from the recordings made in a quiet room. The alarm-specific frequency values (with a resolution of 1 Hz) and period durations were obtained by visual inspection of alarm samples. The reported signal interval durations are an average over the annotated samples. Note that signal (and, therefore, silence) duration measurements are affected by the reverberation inherent to a room environment and may differ from the factory setup values.

Five of the alarm classes (namely, a1, a3, a7, a9 and a10) show some variation in the frequency and duration values among different device units of the same model. Since for the medical staff such alarms are perceived alike, for our purposes they belong to the same alarm class and are referred to as different versions of the alarm.

In total, there are 1431 alarm samples in the annotated data, which corresponds to 19.28% of annotated time. Only the alarm signal interval was labelled (see Figure 4.4 for notation). The acoustic environment of a NICU represents a complex scenario where numerous acoustic events happen spontaneously and simultaneously. Concerning the alarm sounds, most of the time they are overlapped with other sounds, but the overlaps between alarms of different classes or even between those belonging to the same class are not rare either. For example, for the annotated data, the statistics of time when several alarms occur simultaneously is the following: 2 alarms – 6.81%, 3 alarms - 0.70%, 4 alarms - 0.07% of total time labelled as alarm signal.

The average increment of SPLs in intervals with sounding alarms with respect to the intervals with no alarms is 0.7 dBA, which suggests that equipment alarms are not contributing much to the noise levels in the NICU environment. However, it is still of medical concern that the specific spectro-temporal structure of alarms (beats, tones and specially high frequencies) might adversely affect preterm infants.

Table 4.3: Detailed characteristics of the equipment alarm sounds

Class	Source device	Signal description	Frequencies (kHz)	Signal duration (s)	Silence duration (s)	Notification
a1	Monitor	1 tone	i) 0.495, 1.465, 2.435 [1,2] ii) 0.515, 2.455, 3.445, 4.415	0.698±0.170	i) 1.351 ii) 1.548	One of the physiological variables is out of range.
a2	Ventilator	4 higher tones and 1 lower. There is a longer pause between each two alarm periods.	1 / 0.830 [3]	1.232±0.078	0.803	
a3	Incubator	1 higher and 1 lower tone	i) 0.665, 1.330, 1.990, 2.660 / 0.540, 1.600, 3.150 ii) 0.520 / 0.420	0.634±0.142	14.666	Problem with the incubator setup.
a4	Ventilator	3 shorter and 1 longer tone, short pause, 1 longer tone	2.350, 4.700	2.675	0.785	
a5	Incubator	3 tones	0.530, 1.060, 1.590, 2.120	0.970	3.280	Humidity or temperature value is out of range.
a6	Ventilator	1 tone	2.410	0.374±0.069	0.073	
a7	Monitor	1 tone	i) 0.980, 2.935 ii) 2.880	0.836±0.188	0.179	Desaturation.
a8	Monitor	1 tone	0.490, 1.480, 2.460, 3.440, 4.420	0.280±0.060	1.965	

[1] Comma-separated frequencies are simultaneous in time.

[2] Information in each item corresponds to a different version of the alarm.

[3] Information separated with slash corresponds to tones consecutive in time.

Table 4.3: Detailed characteristics of the equipment alarm sounds

Class	Source device	Signal description	Frequencies (kHz)	Signal duration (s)	Silence duration (s)	Notification
a9	Thermometer	1 tone	i) 5.320 ii) 5.190 iii)6.030	0.525±0.161	0.545	Switching on/off or end of measurements.
a10	Infusion pump	1 tone	i) 1.140, 2.280, 3.425 ii) 0.880	0.675±0.112	0.325	Medication finished.
a11	Infusion pump	3 tones	0.880, 1.740	0.376±0.073	2.574	Waiting mode.
a12	Ventilator	3 tones, short pause, 2 tones	2.305, 4.610, 6.915	1.820±0.077	0.680	
a13	Incubator	2 lower tones and 1 higher	0.475, 1.335, 3.100, 3.985, 5.750 / 0.540, 1.590, 2.650, 3.680, 5.770	1.105	10.905	
a14	Incubator	3 tones, short pause, 2 tones. There is a longer pause between each two alarm periods.	3.075, 6.115, 9.195	1.730	0.370	The incubator setup is incorrect.
a15	Infusion pump	2 higer tones and 1 lower	1.270, 3.810, 6.330, 8.870 / 1.015, 3.015, 5.015, 7.020	0.56	0.3	
a16	Monitor	1 tone	0.495	0.307±0.055	1.746	Temperature sensor is not connected.

4.5.2 Vocalizations

In fact, the considered vocalization sounds are generic and, except perhaps of baby crying, may occur in many acoustic environments. The properties of these sounds have already been described in reported works; therefore, their analysis is based on the literature review of these works. In general, the content of vocalizations can be observed up to 8 kHz. In our data, speech-related sounds (i.e. foreground and background voices) are predominant and occur 89.68% of time annotated as vocalizations.

Speech properties were analysed thoroughly in the scope of development of automatic speech recognition systems [126]. Speech is a slowly time varying signal (fairly stationary up to 100 ms), which contains aperiodic unvoiced and quasi-periodic voiced intervals. In spectral representation, voiced intervals usually consist of individual spectral harmonics corresponding to the pitch of the speech waveform, and intervals of unvoiced speech mainly correspond to high-frequency content. The distinctive frequency components of the speech signal, called formants, are frequency bands that carry most of the acoustic energy. Typically, there are three significant formants below about 3.5 kHz. Basically, shouts have the same acoustic properties as normal speech, except that some of their formants occur at higher frequencies [127], they are associated with higher amplitude values [128], and show a shorter duration [129].

The acoustic properties of both cough and infant vocalization sounds were extensively studied for diagnostics of various diseases [130,131]. The fundamental frequency of cries ranges from 0.2 to 0.6 kHz, and usually six formants can be observed up to around 7.2 kHz. The duration of an infant cry is quite short, and in our data was up to 600 ms. A typical cough sound is generated by a sudden air expulsion from the airways [130], and has a higher degree of irregularity compared to speech. It is usually described as having three phases (namely, explosive, intermediate and voiced), where the first and the last phases are sometimes called “bursts” [132]. Cough has a wide distribution of energy across the frequencies, in contrast to the voiced speech, which exhibits harmonic content. The typical duration of cough is about 350 ms.

Laughter is described as a highly diverse signal with various subtypes, which in some aspects resembles speech [133]. The voiced laughter (which is the type we mostly observed in our data) usually has a vowel-like structure and is harmonically rich, although the formant structure of laughter is less prominent than that of speech vowel sounds. Also, the fundamental frequency is much higher in laughter than in speech.

4.5.3 Equipment noise

During the analysis of the audio recordings we found two specific types of noise produced by equipment. In most of the cases these noises are simultaneous, but may also happen individually, and are present throughout recordings obtained both inside and outside the incubator. Figure 4.5 shows the spectrogram of a *neutral* scenario from the recording session RS17 with the observed equipment noises marked.

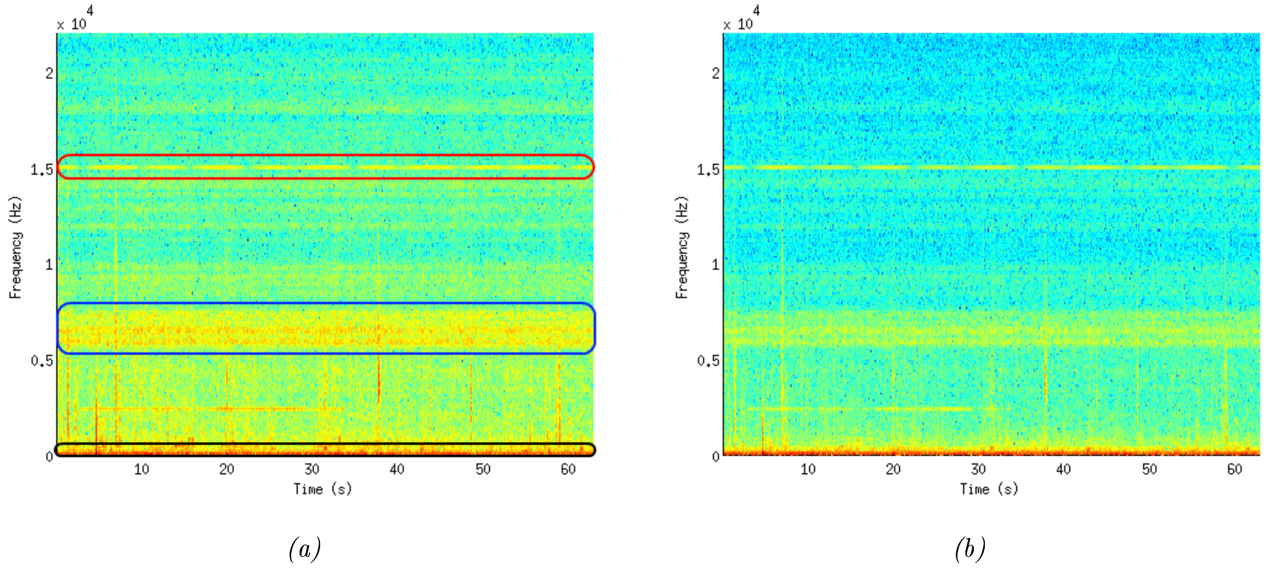
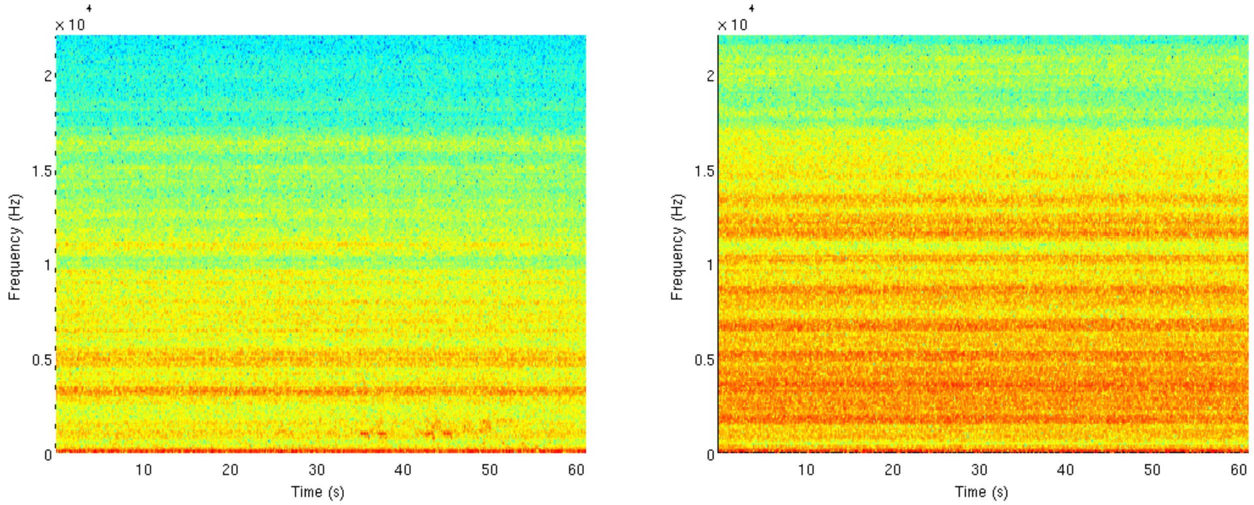


Figure 4.5: Spectrogram of a neutral scenario acquired (a) outside and (b) inside the incubator with the equipment noise marked: the narrow-band noise in red, the ventilation noise in blue and the noise studied in the reported work in black.

The first type of noise is a narrow-band noise at 15 kHz frequency. That noise has short temporal interruptions that can be considered almost periodic: 5.5 s of noise are followed by 1-1.5 s of pause. Most probably this noise is generated by ventilators, biomedical devices for the supporting breathing function of infants. In the vast majority of recordings the noise is stronger outside than inside the incubator.

The second type of noise is the ventilation noise, which is a stationary noise usually spread over a wide frequency range. Depending on the recording session (i.e. equipment used) the noise is stronger either inside or outside the incubator; in particular, for the session showed in Figure 4.5 it is stronger outside the incubator. There are several different types of ventilation equipment in the NICU, having noises with different spectral characteristics. Depending on the particular needs of a preterm infant an appropriate type of ventilation is used, and this fact introduces a lot of variability to the data. Note that the ventilation noise sample depicted in Figure 4.5 is one of the weakest in our database and was chosen for the clarity of presentation. To give a better notion, Figure 4.6 shows spectrograms of the *neutral* scenario (recorded with the microphone inside the incubator) from other sessions, which represent more typical samples of the ventilation noise.

A specific type of ventilation noise is the CPAP noise, which was only observed in the recording session RS3. This noise is also wideband, but has periodic intervals with higher amplitude. Basically, the noise periodicity corresponds to the required breathing pattern, and during that periodic intervals the air under pressure is pumped into the airway of lungs.



(a)

(b)

Figure 4.6: Spectrograms of the neutral scenario from recording sessions (a) RS11 and (b) RS13.

The low-frequency noise marked in Figure 4.5 in black was studied in [17]. In that work, only the frequencies up to 2048 Hz were considered. Some noise at 200, 400, and 600 Hz was regarded to be generated by the incubator fan, and some ventilator noise was reported to contribute to higher SPLs in the low-frequency band of 0 to 100 Hz.

4.6 Chapter summary

In this chapter, the results of exploring the NICU acoustic environment from the set of acquired audio recordings were reported. The whole content of the audio signal was analysed, and besides the usual measurements of SPLs, the identity of sounds and their spectro-temporal properties were described.

First of all, an extensive list of more than 60 acoustic events found in the NICU environment was presented, which generally contains sounds produced by human body, by objects handled by humans or by equipment. A general sound taxonomy of a NICU environment was proposed to structure the whole diversity of sounds into acoustically homogeneous groups. Three major acoustic groups were defined, so any NICU sound can be attributed to *tone*, *vocalization* or *other*. These groups were then divided into more specific acoustic subgroups and semantical categories.

Further, a detailed acoustic analysis of the most represented types of sounds was provided, namely, equipment alarms, vocalization sounds and equipment noise. A thorough description of 16 classes of acoustic alarms found in the NICU was carried out, where their sources, types of notification, spectro-temporal properties and information about occurrence were specified. Since vocalizations are generic sounds that are well-studied, the description was based on the review of reported works. Finally, the two types of equipment noise found were analysed, and it has been observed that the stationary ventilation noise is predominant in most audio recordings.

Apart from that, a set of acoustic scenarios was defined and characterised. The scenarios, which were used in the recordings, were described in terms of their general spectral properties and specific acoustic events. Also, average SPL measurements from scenario recordings were compared with regards to the position of the microphone (inside or outside the incubator) and the period of the day (morning or afternoon). It has been observed that the difference between the recording sessions is more significant than the difference between the acoustic scenarios itself, and the factors contributing to a strong inter-session variability were outlined.

In summary, the provided description of the acoustic environment of a NICU showed its strong diversity.

Chapter 5

Acoustic alarm detection

5.1 Chapter overview

A number of alarm sounds triggered by biomedical equipment occur frequently in the noisy environment of a Neonatal Intensive Care Unit (NICU) and play a key role in providing healthcare. This chapter presents our work on the automatic detection of acoustic alarms in that difficult environment. In particular, two detection problems were considered in this thesis work. First, a binary detection problem (alarm vs. non-alarm) with the aim to automatically label temporal regions within the input audio where an alarm is sounding, the work on which was reported in [134]. Second, a more challenging detection problem, where not only the timestamps, but also the particular type of alarm sound is detected.

For the latter problem, we propose several detection systems, which are based on different approaches: 1) a relatively simple signal processing based approach; 2) a knowledge-based machine learning approach that takes advantage of the peculiar spectral and temporal properties of alarms; 3) an approach based on neural networks, in which all stages of the detection system are machine learning based.

This chapter is organized as follows. Section 5.2 reviews the research on the topic of automatic detection of alarm sounds done so far. The general evaluation setup used to assess the detection systems performance and some considerations about their development are given in Sections 5.3 and 5.4, respectively, whereas Sections 5.5, 5.6 and 5.7 describe the three alternative approaches proposed.

5.2 Related work

To our knowledge, research on the topic of automatic alarm sounds detection was first reported in [46], where general characteristics of alarms are described and the conventional approach based on techniques and representations from speech recognition is compared to the signal-separation approach based on sinusoidal modelling.

Posterior works investigated acoustic alarm detection for the purposes of hearing impaired assistance in traffic [135] or hearing support in very noisy conditions [136]. The proposed methods usually try to make use of peculiar properties of alarms in one form or another. The algorithm presented in [136] is based on detection of amplitude periodicity in a specified frequency bandwidth and applies a set of rules to the zero-crossing rate of the autocorrelation function. In [137] a real-time siren detection system is proposed that employs pitch detection in the predefined frequency range and makes a decision by comparing the presence probability to a fixed threshold. A rule-based approach is proposed in [138], where spectral and time-domain morphological features, which estimate various parameters of the considered alarms, and are used together with template-based distance computation. In [139] the acoustic siren detection problem is tackled from the image processing perspective. In that work the spectrogram is treated as an image and part-based models, which consist of spectro-temporal patches in relative and flexible time-frequency configurations, are learnt.

5.3 Evaluation setup

The experiments were carried out with the part of the recorded database that was annotated. The total amount of data used is around 54.3 minutes, and 19.74% of this time is labelled as alarm (note, only the alarm signal interval was labelled). In total there were 47 files from different recording sessions (the concrete scenario files used can be found in Table 3.3). Only recordings made with the microphone placed outside the incubator were used to keep homogeneous experimental conditions, and also because this microphone is closer to the alarm sources. The original 44.1 kHz recordings were downsampled to 24 kHz.

As the dataset is relatively small, in order to obtain more statistically relevant results, a 10-fold cross-validation scheme was applied, i.e., on each fold, 9 sessions of data were used for training and 1 session for testing. Further, the results were aggregated over all 10 folds and the overall metric scores were obtained. The reported results correspond to the average of class-based metric scores.

Apart from the frame-level metrics used during system development, we evaluate the detection systems using the event-level metric that can present system performance in a way more meaningful for the medical application.

5.3.1 Frame level evaluation

For the frame-based or frame-level evaluations, the Missing Rate (MR) and the False Alarm Rate (FAR) metrics are used. These are defined as

$$MR = \frac{N_M}{N_A}, \quad FAR = \frac{N_{FA}}{N_{NA}}, \tag{5.1}$$

where N_M and N_{FA} is the number of misclassified frames for alarm and non-alarm class, respectively, and N_A and N_{NA} is the total number of alarm and non-alarm frames, respectively.

In our initial work reported in [134] a different frame-based metric was used, which is defined as one minus the relative system error:

$$\mathit{FB\text{-}ACC} = 1 - \frac{N_M + N_{FA}}{N_{total}}, \tag{5.2}$$

where $N_{total} = N_A + N_{NA}$ is the total number of frames evaluated. This metric reflects the overall system accuracy and equally treats both types of errors, namely misses and false alarms, which may not be particularly suitable when there is a tangible unbalance in the number of alarm and non-alarm frames. Therefore, only MR and FAR metrics were used for the frame-level evaluation as they provide more adequate information about the detection performance.

5.3.2 Event level evaluation

The period of the alarm is chosen as event, since it is a natural alarm-specific unit. The period-based error rate, denoted as PB-ERR, is defined as the reformulated F-score as follows

$$\mathit{PB\text{-}ERR} = 1 - \frac{2 \cdot N_C}{2 \cdot N_C + N_{FA} + N_M}, \tag{5.3}$$

where N_C is the number of correctly detected reference alarm periods, N_M and N_{FA} is the number of missed and falsely inserted periods. Each reference period is regarded as correctly detected if there exists a detected alarm period in the tolerance interval $[T_{ref} - T_{tol}; T_{ref} + T_{tol}]$, where T_{ref} is the reference period timestamp and T_{tol} is the tolerance interval duration. Note that T_{tol} should be less than half the alarm period duration, otherwise one detected period may be associated with two reference periods making both correctly detected. In this work, the T_{tol} was set to 49% of the alarm period duration, and in fact it is the largest value T_{tol} can take on. In this case, the system is expected to detect an alarm in the tolerance interval that has the duration of almost one alarm period, which is acceptable for the medical application, taking into account that the duration of most of the alarm classes is quite short.

Another event-level metric, that with regards to a time span is a trade-off between an alarm period and an alarm sequence (see Figure 4.4 for notation), was proposed in our earlier work [134] for the binary detection problem. Inspired by [113], we call it a block-based metric. To compute that block-based metric, the input audio stream is divided into consecutive non-overlapping blocks of 5 s length.

For each of them a label (alarm or non-alarm) is assigned using the following criterion: the block is labelled as alarm in case it has more than one alarm signal; otherwise it is labelled as non-alarm. The basic idea that is being pursued is that neither the staff nor the preterm baby respond to only one alarm signal, but there should occur several of them (we believe from 2 to 4 periods of signal-silence, are enough) in order that the sound is perceived as alarm. The defined block-based metric is based on the detection cost function (C_{Det}) used in NIST evaluations [140], and is computed using the formula:

$$\begin{aligned} \mathit{BB\text{-}ACC} = 1 - \alpha \cdot ((C_M \cdot P_{M|Target} \cdot P_{Target}) + \\ (C_{FA} \cdot P_{FA|NonTarget} \cdot P_{NonTarget})), \end{aligned} \tag{5.4}$$

where $P_{Target} = 0.59$ and $P_{NonTarget} = 0.41$ are, respectively, the fractions of alarm and non-alarm blocks calculated over all the database; $C_M = 0.3$ and $C_{FA} = 0.7$ are estimated application-specific costs of misses and false alarm errors; α is a normalization factor equal to a fraction of 1 by a score of the system, that is always wrong; and, finally, $P_{FA|NonTarget}$ and $P_{M|Target}$ basically correspond to the FAR and MR metrics, defined above in (5.1), calculated at the block level.

In fact, the inclusion of the application-specific costs of miss and false alarm errors in the block-based metric was motivated by the fact that the vast majority of sounding alarms are non-actionable [141]. Although this may suit well the purposes of medical staff notifications, the alarms that are missed by the system due to the low cost of miss errors may still be perceived by the preterm baby.

Only the period-based metric (PB-ERR) is used in our event-level evaluations. Note that this metric could also be used for the binary detection problem evaluation, in which case the tolerance value T_{tol} should be fixed.

5.4 Development of detection systems

Only 7 alarm classes out of 16 described in Section 4.5.1 were chosen in our tests under the criteria of having sufficient number of samples in the database and being relevant from the medical point of view. These are classes a1, a3, a6, a7, a8, a10 and a16.

Each developed detection system consists of a set of binary detectors (alarm class vs. non-alarm class), where the total number of detectors corresponds to the number of considered alarm classes, i.e. is 7 in our case. An individual binary detector is designed to deal with a particular alarm class, and is trained following the one-against-all strategy. The input audio is processed by each binary detector independently. Although this solution may be complicated when the number of considered classes is large, there are several reasons behind using it:

1) A set of considered alarm classes is not definitive, e.g. new alarm types can appear in the NICU environment if new equipment is installed. Also, due to the lack of data, not all the alarm types already found were considered for detection. Building individual binary detectors provides flexibility as the detection system can be easily extended to new alarm classes.

2) A detector designed for a particular alarm class can better exploit its specific properties, which are useful for discriminating that class (e.g. by employing a specific feature set).
3) In case of temporal overlaps between alarm sounds, multiple labels can be provided by the detection system for the audio region with overlapping.
4) A set of alarm classes in another NICU room may be different, and already developed individual detectors could be reused for coinciding alarm classes.

Due to the specific characteristics of the proposed systems, they perform detection either at the frame (neural network based systems) or at the period (signal processing based system) level. The knowledge-based system initially provides decisions at the frame level, but includes a specific post-processing scheme (with temporal modelling, see Section 5.6.3) for obtaining the period level decisions. For systems operating at the frame level, the decision threshold is chosen based on the Equal Error Rate (EER) criterion, so assuming that both miss and false alarm errors are equally important at the frame level. In these systems, for the period-level evaluation, the beginning of a sequence of consecutive frames detected as belonging to the alarm class is regarded as the timestamp of the alarm period label. For the systems providing period-level decisions, to obtain the frame-level decisions, L_{sig} frames after each of the detected alarm periods are assigned to the alarm class. In all cases, a constraint of minimal distance between the detected periods is applied, where the minimal distance is taken equal to 75% of the alarm period duration.

5.5 Signal processing based approach

The proposed system[1] that is based on matched filter and morphological tools consists of 4 different stages (see Figure 5.1). The first stage is an Energy Overload Protection (EOP), which performs a prior enhancement of the input signal $s[n]$. The second stage, a Matched Filter (MF), is used to obtain a signal proportional to the detection output, where $a_i[n]$ is the reference sample of alarm of class i. At the third stage that signal is processed with morphological tools to obtain an envelope $e_i[n]$. Finally, at the last stage the decision about whether the alarm is detected or not is taken. The output of the system is the detection signal $d_i[n]$, which is equal to 1 if alarm i is detected and to 0 otherwise.

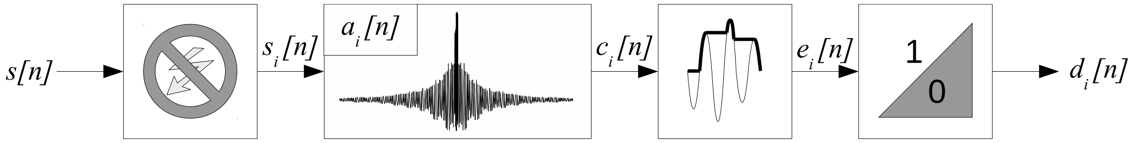


Figure 5.1: General scheme of the proposed detection system with four stages depicted (from left to right): energy overload protection, matched filter, morphological envelope, decision.

[1] The detection system described in this section was developed by Sergi Gómez Quintana in his Final Project work, which was supervised by the author of this PhD thesis.

In the following sections the stages of the system are described in details. Although the EOP is the very first stage of the proposed system, it was developed the last as an enhancement and will be described in the end.

5.5.1 Matched filter stage

From the system point of view, the Matched Filter (MF) can be expressed as a linear and time-invariant system (see the second block in Figure 5.1). The output $c_i[n]$ is computed as:

$$c_i[n] = xc\{s_i[n], a_i[n]\} = \frac{\sum_{k=0}^{L-1} s_i[n+k]a_i[k]}{\sum_{k=0}^{L-1} a_i^2[k]}, \tag{5.5}$$

where xc is the cross-correlation function, $s_i[n]$ and $a_i[n]$ are the input and the reference alarm signals, respectively. Note that for a particular alarm class the detection system consists of a bank of MFs, each dealing with a different version of that alarm class.

Since the recorded alarm reference may have certain noise floor at non-alarm frequencies, filtering is performed to obtain a “clean” reference. The designed filter is based on relevant harmonics, which are estimated from the noisy alarm reference. Relevant harmonics are defined as frequencies corresponding to local maxima of power spectral density such that their power with regards to the power of the strongest frequency (e.g. fundamental) is under Relevant Harmonics Ratio (RHR) value (see Figure 5.2). Note that RHR has to be less than the Signal-to-Noise Ratio (SNR) of the noisy reference, and in this work it is set to 90 dB. For alarms with several tones, the relevant harmonics are defined in frames (200 ms, half-overlapped), where the power spectral density of each frame is normalized with respect to the maximum among all frames.

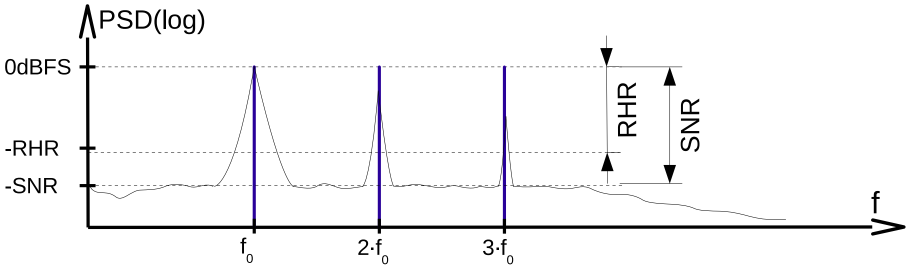


Figure 5.2: Graphical example of relevant harmonics definition (shown in blue).

5.5.2 Morphological envelope stage

The Morphological Envelope (ME) stage can be represented as a concatenation of two non-linear and time-invariant systems: full-wave rectifier and morphological closing. Closing is a morphological operator defined as:

$$\varphi\{x[n]\} = \varepsilon_b\{\delta_b\{x[n]\}\}. \tag{5.6}$$

In this equation, $\delta\cdot$ and $\varepsilon\cdot$ are the other two simple morphological operators called dilation and erosion, respectively, both of which use a binary sequence $b[n]$ called structuring element. In particular, a flat structuring element of size S is used:

$$b[n] = \begin{cases} 0, & \text{if } 0 \leq n \leq S \\ -\infty, & \text{otherwise} \end{cases} \tag{5.7}$$

where the value of S is based on the fundamental frequency of alarm f_0 and is equal to the integer closest to $1/f_0$. Since closing only takes into account positive peaks, a full-wave rectifier is used to obtain absolute values.

5.5.3 Decision stage

At this stage, the decision about whether a certain peak at the output of the ME stage corresponds to an alarm detection is taken. First, a low-pass Finite Impulse Response (FIR) filter is used to smooth the envelope signal $e_i[n]$. To construct the filter, an expected response is obtained as the output of the ME stage when the reference alarm signal is introduced at the input of the system. Further, since the envelope can contain some peaks that do not correspond to the alarm being detected, the values of the envelope below a predefined threshold U are assigned to zero. The particular value of the threshold is based on the maximum value of that non-alarm peaks. Finally, the peak detector outputs the binary decision signal $d_i[n]$ which equals to 1 at the local maxima (i.e. peak) locations and to 0 otherwise.

5.5.4 Energy overload protection stage

The recordings often have some strong knocks and glitches, i.e. signal intervals with the energy which is very high in comparison to the usual energy of the input signal, which may affect the response of the MF. In order to deal with that the Energy Overload Protection (EOP) stage is proposed, which consists of a filter followed by a dynamic compressor. The filter designed for obtaining the clean reference alarm signal is employed (see Section 5.5.1).

A compressor is defined by a threshold T (in dB), after which the signal starts to be compressed, and by a compression ratio $R:1$ (in our case, $R = 10$). In this work, the threshold T is estimated statistically from the training data as 90th percentile (thus, only 10% of the input signal are compressed). The input signal is smoothed before compression by convolving its absolute value with a window. That window is defined by 3 time parameters: attack (also known as a look ahead time, 5 ms), sustain (10 ms) and release (50 ms) times, which are defined based on the length of glitches.

5.5.5 Experimental results

Two different setups were used for the system evaluation. In the first setup, which we call *"oracle"*, the knowledge about the temporal location of non-alarm intervals (i.e. labelling) in each scenario sample at the input of the system is used to estimate the threshold U. This setup was used during the system development and gives a notion about the upper bound of its performance.

In the second setup, the threshold is estimated from the training data. For each scenario from the training sessions a threshold is computed using labelling, as was done in the "oracle" setup. The testing threshold is computed based on training thresholds as an average of their minimum and maximum values. Note that the testing threshold will have the same value for all scenarios of the testing session. Since the amplitude of the smoothed envelope may vary from session to session, for each scenario it is normalized by its mode value.

Table 5.1 presents results for the two evaluation setups, and also for the initial version of the developed system (*Without EOP stage*). It can be seen that once the EOP stage is added as the first stage of the system, the amount of false alarm errors is reduced drastically, which leads to a very low PB-ERR metric score of 12.15%.

The results for the evaluation setup *with training* are not as good as the ones obtained for *"oracle"* setup as the threshold U estimate is less precise due to mismatched conditions. The system produces more miss and false alarm errors, where a relative deterioration of 57.27% and 184.24% is obtained in terms of MR and FAR metric scores, respectively. Still, this relatively simple detection system is able to perform quite well in terms of PB-ERR metric.

Table 5.1: Alarm detection performance obtained by the signal processing based system

System setup	Evaluation metrics (%)		
	MR	FAR	PB-ERR
Without EOP stage	**18.23**	67.99	92.46
"Oracle"	27.50	**0.19**	**12.15**
With training	43.25	0.54	34.26

Analysing the detection results for each alarm class separately, classes a1, a8 and a16 reveal the highest scores in terms of MR metric. This can be explained by the fact that these alarms have similar spectro-temporal properties (see Table 4.3), which is confirmed by calculating the normalized maximum of cross-correlation between the alarm reference signals (see Table 5.2, the most similar alarms are marked with greyer background). Basically, a large number of miss errors is caused by a high threshold U estimate, which is due to high non-alarm peaks obtained from similar alarms present in the training data.

5.6 Knowledge-based approach

This section describes a machine-learning detection system where the knowledge about the particular spectral and temporal characteristics of each alarm class is integrated at different stages. The feature

Table 5.2: Normalized alarm cross-correlation values

	a1_v1	a1_v2	a3_v1	a3_v2	a6	a7_v1	a7_v2	a8	a10_v1	a10_v2	a16
a1_v1		0.005	0.013	0.006	0.003	0.003	0.001	0.007	0.005	0.002	0.005
a1_v2	0.005		0.017	0.048	0.010	0.025	0.012	0.589	0.025	0.011	0.080
a3_v1	0.013	0.017		0.067	0.001	0.002	0.000	0.021	0.003	0.001	0.025
a3_v2	0.006	0.048	0.067		0.000	0.004	0.000	0.061	0.002	0.001	0.067
a6	0.003	0.010	0.001	0.000		0.002	0.005	0.014	0.007	0.000	0.002
a7_v1	0.003	0.025	0.002	0.004	0.002		0.004	0.014	0.012	0.001	0.002
a7_v2	0.001	0.012	0.000	0.000	0.005	0.004		0.002	0.003	0.000	0.012
a8	0.007	0.589	0.021	0.061	0.014	0.014	0.002		0.017	0.003	0.231
a10_v1	0.005	0.025	0.003	0.002	0.007	0.012	0.003	0.017		0.006	0.002
a10_v2	0.002	0.011	0.001	0.001	0.000	0.001	0.000	0.003	0.006		0.002
a16	0.005	0.080	0.025	0.067	0.002	0.002	0.012	0.231	0.002	0.002	

extraction is performed around the alarm-specific frequencies and is based on applying either a method for detection of sinusoidal signals (previously published in [142]) or the Non-negative Matrix Factorization (NMF) algorithm. The temporal structure of alarms, in terms of duration of signal and silence intervals in every alarm period, is incorporated by aggregating the frame-level posterior probabilities. The system uses a set of Gaussian Mixture Model (GMM) based or pre-trained Neural Network (NN) based detectors, each designed to deal with a specific alarm.

5.6.1 Modelling of the alarm spectral structure

5.6.1.1 Feature extraction

In all the feature extraction schemes, the acoustic signal is split into frames each frame containing $N = 2048$ samples of the signal and the shift between frames being $L = 1024$ samples. The Discrete Fourier Transform (DFT) of each frame is calculated.

Baseline

The baseline feature extraction scheme consists of obtaining 18 Frequency Filtered Logarithm Filter Bank Energies (FF-LFBE) [66] along with their first temporal derivatives, for each frame. Therefore, the dimension of the feature vector is 36. The FF-LFBEs are generic audio features used in speech and audio processing that cover the entire frequency bandwidth.

Sinusoidal detection based

This feature extraction scheme is based on the fact that the alarms consist only of sinusoidal components, and therefore employs detection of sinusoids. A variety of methods for sinusoid detection have been proposed, e.g., see [143] for a review of methods used in audio processing. In this work, we have employed a method for detection of sinusoidal signals introduced in [142] which tackles the detection of sinusoidal components as a pattern recognition problem. This method was employed in recent works on analysis of bird vocalisations in [144, 145], and in its earlier version for speech recognition [146].

Sinusoid detection is performed independently for each frame. Let us denote by $S_t(k)$ the short-time spectrum of the t^{th} frame of the acoustic signal and by k_p the frequency index of a spectral peak found in the short-time magnitude spectrum. A given spectral peak k_p is characterised by a feature vector $\mathbf{y}{=}(\mathbf{y}^1, \mathbf{y}^2)$, where $\mathbf{y}^1$ and $\mathbf{y}^2$ are formed using M points of the short-time magnitude and phase spectrum around the peak, respectively, to capture the spectral magnitude shape and phase continuity information around the peak. The magnitude shape feature vector $\mathbf{y}^1$ is obtained by using spectral magnitudes normalised by the magnitude value at the peak, i.e., $\mathbf{y}^1{=}(|S_t(k_p{-}M)|/|S_t(k_p)|, \ldots, |S_t(k_p{+}M)|/|S_t(k_p)|)$. The phase continuity feature vector $\mathbf{y}^2$ is obtained by using the spectral phase difference between the current and previous frame, i.e., $\mathbf{y}^2{=}(\Delta\phi_t(k_p{-}M), \ldots, \Delta\phi_t(k_p{+}M))$. The phase difference is defined as $\Delta\phi_t(k) = \phi_t(k) - \phi_{t-1}(k) - 2\pi k_p L/N$, where $\phi_t(k)$ and $\phi_{t-1}(k)$ denote the phase of the frequency point k at frame-time t and $t-1$, respectively.

The distribution of the multivariate feature vector $\mathbf{y}$ is modelled using a multi-component Gaussian mixture. A model is obtained for spectral peaks corresponding to noise, denoted by λ_n, and for sinusoidal signals, denoted by λ_s, at various Signal-to-Noise Ratios (SNRs). For a given spectral peak represented by the feature vector $\mathbf{y}$, the likelihood is obtained on the sinusoidal model, denoted by $p(\mathbf{y}|\lambda_s)$, and on the noise model, $p(\mathbf{y}|\lambda_n)$. The log-likelihood corresponding to non-peak spectral points is randomly drawn from a uniform distribution in the interval $[-710; 690]$.

The above provides an information about the detected sinusoidal components at each signal frame. We performed the following steps to refine this result. Firstly, only the peaks above 40 dB in relation to neighbouring spectral points were considered. Further, all segments of a very short length, specifically those of less than 4 frames, were discarded assuming that these were detected by error.

An example of a spectrogram of an audio recording and the detected sinusoidal components is depicted in Figure 5.3. Note that the binary decision about each peak based on the difference $p(\mathbf{y}|\lambda_s) - p(\mathbf{y}|\lambda_n)$ is shown. It can be seen that even weak sinusoidal components are detected well.

We form a feature vector characterising each alarm frame by selecting the log-likelihood values $\log p(\mathbf{y}|\lambda_s)$ and $\log p(\mathbf{y}|\lambda_n)$ obtained from the sinusoidal detection in the frequency intervals around each alarm-specific frequency (from the Table 4.3) with the tolerance $\delta = \pm 20$ Hz. Only one value for each likelihood is chosen in each interval, and in this work it corresponds to the spectral point that has the maximum sinusoidal model likelihood in that interval.

We also incorporate the amplitude structure of the alarms by including in the feature vector the magnitude values at individual alarm-specific frequency regions. In order to disregard the effect of volume, these magnitudes are normalised by the sum of the magnitudes of all the alarm-specific frequencies.

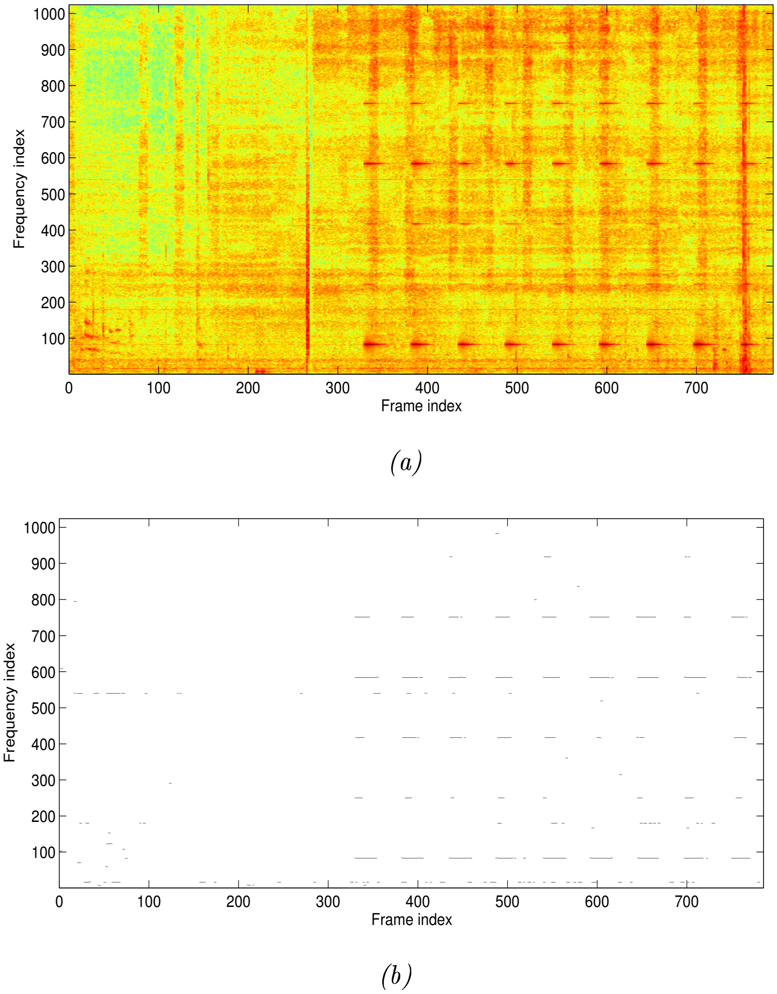


Figure 5.3: An example of a spectrogram (a) of audio recording and the detected sinusoidal components (b).

The following parameter setup is used. Each frame is rectangularly windowed and padded with 2048 zeros to obtain a finer sampled DFT spectrum. The parameter M is set to 6 frequency bins and the sinusoidal/noise models consist of 32 Gaussian mixture components.

NMF based

Similar to works reported in [64,147], the feature representation employed in this work is based on the activations obtained after NMF separation (see Section 2.6.2). In our experiments, we consider $S = 2$ sources corresponding to alarm and non-alarm classes, and the global bases matrix $W_{train} = [W_A; W_{NA}]$ consists of the bases trained for each class, respectively. The alarm bases W_A are trained using the alarm signal intervals only and the non-alarm bases W_{NA} are trained using the data segments that do not contain any alarms. In this case, the number of alarm bases accounts for the variability that may be present in the alarm signal interval, i.e. different alarm versions, distinct alarm tones, variation of the tone amplitude, etc. The whole set of activations H is normalized in each frame such that it sums to 1 and only activations corresponding to alarm bases H_A are used as features.

In our work, the implementation of NMF described in [116] is used, with the following parameter setup: the input matrix V is a magnitude spectrogram computed on Hann-windowed frames. As in the sinusoidal detection method, only the spectral points within frequency regions around each alarm-specific frequency with a tolerance δ are used for NMF processing. We train $R = 4$ and 15 bases per alarm and non-alarm classes, respectively, where each base corresponds to a vector of dimension $F \times 1$. The sparsity parameter λ is set to 1. At the training and testing time we use up to 20 iterations. Note that a cross-validation scheme was also applied for NMF-based feature extraction, where 9 sessions were used for training the bases, which were applied to process 1 testing session.

5.6.1.2 Statistical modelling

A variety of pattern recognition techniques can be employed to construct class-specific detectors modelling the spectral features described in the previous subsection. In this work, we used Gaussian Mixture Modelling (GMM) and Neural Networks (NN).

For each alarm class, a GMM-based detector consists of a model for alarm and a model for non-alarm. Generally each model is a single Gaussian probability density function with diagonal covariance matrix as, in our experiments, this resulted in better recognition performance than using more mixture components.

The unsupervised pre-training of NN is performed using Deep Belief Networks (DBNs) as described in Section 2.5.1. Due to the scarcity of data, only one hidden layer networks are explored (as shown in Figure 2.1). The hidden layer has 32 units. Experimentally, the size of each minibatch is set to 10 and the inputs are randomly distributed among minibatches. The learning rate (α), the number of epochs (NofE), and the momentum in the unsupervised stage are set, respectively, to 0.001, 80, and 0.9. The supervised learning is then carried out with $\alpha = 0.001$, NofE $= 50$, and a fixed momentum of 0.9. The weight decay for unsupervised and supervised stages is set, correspondingly, to 2×10^{-7} and 1.2×10^{-4}.

5.6.2 Modelling of the alarm temporal structure

The log-posteriors of the alarm and the non-alarm class are calculated for each frame based on the probabilities obtained from the statistical models described in Section 5.6.1. The information about the longer-term temporal structure of alarms is incorporated by aggregating these frame-level log-posteriors over the intervals corresponding to durations of signal and silence segments in every alarm period. At each frame t, the probability of it being the first frame of the alarm period is calculated as

$$P_{period}(t) = \sum_{i=t}^{t+L_{sig}+L_{sil}-1} \alpha(i) \cdot (P_A(i) - P_{NA}(i)) \tag{5.8}$$

where P_A and P_{NA} are log-posteriors of the alarm and non-alarm class, L_{sig} and L_{sil} are, correspondingly, the duration of signal and silence intervals in an alarm period, and $\alpha(i)$ is set to 1 for

$i \in [t, t+L_{sig}-1]$ and to -1 for $i \in [t+L_{sig}, t+L_{sig}+L_{sil}-1]$. The alarm period probability estimates correspond to the peaks of the curve resulting from computing that aggregated probability along the frame time index. An illustration is given in Figure 5.4.

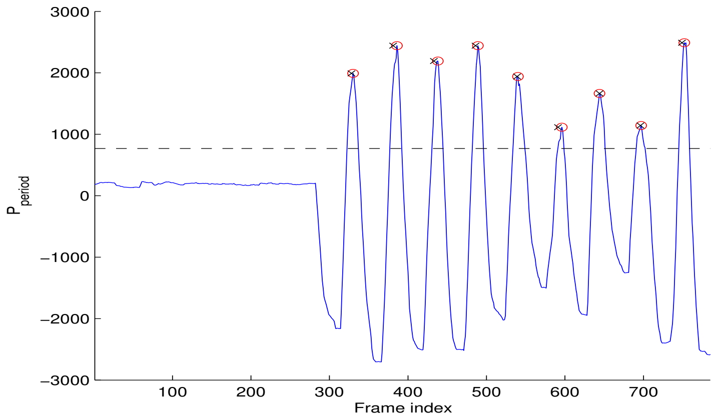


Figure 5.4: The output of the period probability estimation. Circles correspond to the estimated period timestamps after applying a threshold and crosses are the reference period timestamps.

5.6.3 Post-processing and decision

Several alternative decision and post-processing schemes were applied after the modelling.

First, with the likelihoods obtained from the models, each frame was classified either as alarm or non-alarm. The resulting sequence of labels was smoothed by means of the majority voting. The length of the smoothing window was set to be the minimum of the signal and silence interval length in an alarm period.

Second, the period probability $P_{period}(t)$ was subjected to a class-specific thresholding and the peaks of this probability curve above the threshold were chosen as the detected alarm periods and were directly evaluated at the period level (circles in Figure 5.4). Note that the class-specific threshold was chosen so as to provide the best period-level performance.

Third, a parallel combination of the previous two schemes was applied as follows. By an alarm event we denote a sequence of consecutive frames belonging to the alarm class. If none of the detected alarm periods obtained from the second scheme coincides with an alarm event from the first scheme or is around it with a tolerance of $\pm L_{sig}/2$, the frames of that event are assigned to the non-alarm class.

5.6.4 Experimental results

5.6.4.1 Comparison of feature extraction schemes

First, in Table 5.3 we present experimental results obtained when only the modelling of the spectral structure of the alarms is incorporated. In this setup, the GMM-based detectors are used and the post-processing steps are omitted. The EER, which corresponds to both MR and FAR metrics having the same value, is reported.

It can be seen that the baseline features (row 1) are not performing well as they do not take into account the specific properties of the alarms we are dealing with. Both features based on the sinusoidal detection (SD) and on the non-negative matrix factorization (NMF), which exploit the knowledge of alarm properties, can significantly outperform the conventional features. The relative improvement obtained in both cases is equal to, correspondingly, 62.12% and 45.01%.

Table 5.3: Alarm detection performance obtained by a system modelling the spectral structure only

Features	Evaluation metrics (%)
	MR = FAR
Baseline	35.30
SD LLH ratio	32.16
SD LLH	14.52
SD LLH & Amp	**13.37**
NMF	19.41

The second part of the Table 5.3 (rows 2-4) shows results when SD is applied for feature extraction. In this case the feature vector can be formed using the log-likelihood ratio of the sinusoidal and noise models (LLH ratio) or using these log-likelihoods as separate features (LLH). The performance of the detection system employing the latter features is clearly better as more information is provided to classifiers. These features are further combined with the normalized magnitude values to model the alarm amplitude structure (row 4), which brings an additional relative improvement of 7.92%. In fact, the information about the amplitude structure may be helpful for distinguishing between alarms that share very similar frequency components as well as between different alarm versions.

The last part of the table presents the results for the NMF-based features and it can be seen that they do not outperform the SD-based features. Actually, their performance is 45.18% relatively worse, which may be explained by the fact that the spectral information captured by NMF-based features is less accurate. In fact, the NMF framework is based on an approximation, which is performed both at the training and the source separation (i.e. feature extraction) steps. While the SD algorithm treats each spectral point independently, in NMF processing, the spectral structure of alarms is captured as a whole by the trained bases. Also, unlike the SD-based features, the activations obtained from NMF processing can be sensitive to the amplitude of the signal.

5.6.4.2 Assessing the performance of the system according to the quality of alarm samples

The alarm occurrences which are most difficult to detect are likely those associated with low Signal-to-Noise Ratio (SNR) values. In this case, the effect of the alarm stimuli on the preterm infant is very small, so a more adequate measurement of the detection error may be obtained by discarding the alarm occurrences with low SNR values. In this section we explore the performance of the system for the best performing feature setup (i.e. *SD LLH & Amp*) considering the quality of the labelled alarm, which is assessed by calculating the local SNR value. The idea is that the effect on the preterm infant of the auditory stimulus due to an alarm is noticeable only if its SNR is sufficiently high. The SNR value is calculated using the recordings made with the microphone placed inside the incubator, so it measures what the preterm infant was receiving. As in the previous section, the GMM-based detectors are employed and no post-processing steps are performed.

For each labelled alarm, the local SNR is calculated around alarm-specific frequency bins f_b with a margin $\pm\delta$ and both the signal and noise powers are estimated by averaging the spectrum both in time and frequency. The signal power for a bin f_b is estimated as

$$P_{s_b} = \frac{1}{T \cdot (2\delta + 1)} \left(\sum_{t=1}^{T} \left(\sum_{k=f_b-\delta}^{f_b+\delta} S_{k,t}^2 \right) \right), \tag{5.9}$$

where $S_{k,t}^2$ is the spectral power at bin k and frame t, T is the number of frames of the current alarm signal. And taking K spectral power values around the alarm-specific bin, we estimate the noise power as,

$$P_{n_b} = \frac{1}{T \cdot K} \left(\sum_{t=1}^{T} \left(\sum_{k=f_b-\Delta}^{f_b-\delta-1} S_{k,t}^2 + \sum_{k=f_b+\delta+1}^{f_b+\Delta} S_{k,t}^2 \right) \right), \tag{5.10}$$

where $K = 2(\Delta - \delta)$ and Δ corresponds to 100 Hz. The noise margin value Δ is chosen so as to avoid overlapping with alarm-specific bins while keeping enough samples for estimation. Then the SNR value for an alarm sample is obtained as an average over the corresponding alarm-specific bins as follows

$$SNR_{dB} = 10 \cdot \log_{10} \left(\frac{1}{B} \sum_{b=1}^{B} \frac{P_{s_b}}{P_{n_b}} \right), \tag{5.11}$$

where B is the total number of alarm-specific bins.

Figure 5.5 shows the distribution of the alarm samples as a function of their local SNR value. It can be seen that this distribution is rather exponentially modified Gaussian with the exponential decay towards higher SNR values.

The whole range of SNR values over the entire labelled database was further divided in 5 dB intervals and all alarm samples were grouped according to these intervals. These groups were evaluated independently and the evaluation results are presented in Table 5.4 as an average over the considered alarm classes. It can be clearly seen that the system performance improves as the SNR becomes higher and so the quality of the evaluated alarm samples increases. Note that the models used for this

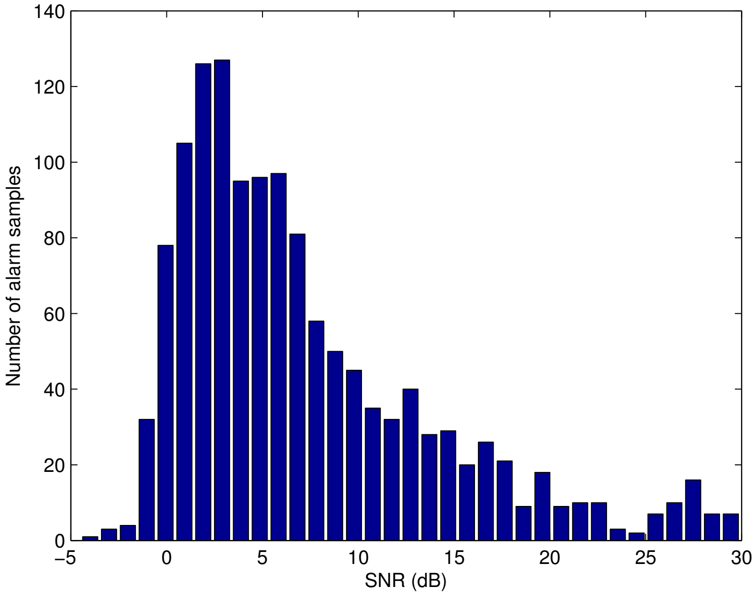


Figure 5.5: Global SNR histogram over all labelled alarm samples from the database.

evaluation were trained using the whole set of alarms from the database, which means that the models were trained on multiple conditions.

Table 5.4: Alarm detection performance obtained over the SNR intervals

SNR range (dB)	< 0	0 - 5	5 - 10	10 - 15	15 - 20	20-25	> 25	All
MR = FAR (%)	19.79	15.84	13.66	13.19	7.40	9.01	**7.27**	13.37
Alarms evaluated	77	555	354	122	96	36	47	1347

We further explored how the performance of the detection system changes in case the lowest quality alarm samples are discarded from the evaluation. Table 5.5 shows the evolution of the detection error with regards to the threshold placed on the SNR values, where alarms with SNR below this threshold are not included in the evaluation. Notice that there is a drop in the detection error when alarm samples with SNR value below 5 dB are discarded, and in that case the detection error (MR = FAR) becomes 10.55%.

5.6.4.3 Comparison of statistical models

The extracted spectral features are further modelled by the individual class-specific detectors, and in this work we explore two different statistical models, described in Section 5.6.1. As in the previous subsection, for this comparison no post-processing schemes are applied, and the best-performing feature extraction setup, namely *SD LLH* & *Amp*, is employed.

Table 5.5: Alarm detection performance obtained by discarding the alarm samples below the SNR threshold

SNR threshold (dB)	None	0	5
MR = FAR (%)	13.37	13.09	10.55
Alarms discarded	0	77	632

The Detection Error Tradeoff (DET) graphs for the GMM-based and NN-based statistical models are shown on Figure 5.6. The curves were obtained by varying a threshold on the log-likelihood ratio and averaged over the considered alarm classes. It can be seen that the GMM-based models outperform the NN-based ones at almost all the operating points of the curve, even though the NN-based models are discriminatively trained. This behaviour may be explained by the fact that a very limited amount of data is available for model training, which reduces the generalization capability of the networks and may cause overfitting.

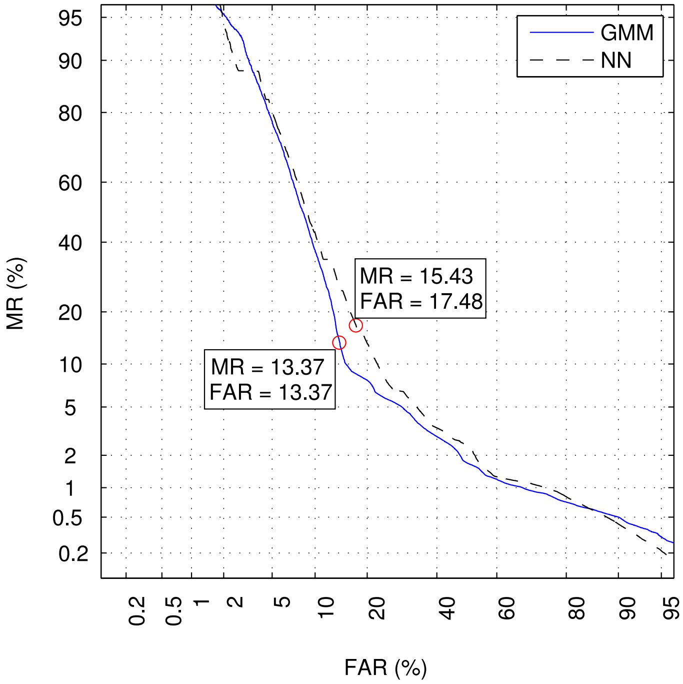


Figure 5.6: The Detection Error Tradeoff (DET) graphs for different statistical models. Circles correspond to points closest to EER.

5.6.4.4 Comparison of post-processing schemes with application-specific evaluation

Table 5.6 shows the results when temporal modelling and smoothing are incorporated, as described in Section 5.6.3. It can be seen that none post-processing scheme improves MR scores compared to not performing any post-processing at all, but all schemes improve the FAR metric scores to a large extent (up to 87.8% relative improvement in the best case). Moreover, all the post-processing schemes are able to improve the PB-ERR scores.

Table 5.6: Alarm detection performance obtained from different ways of post-processing

Post-processing	Evaluation metrics (%)		
	MR	FAR	PB-ERR
None	**13.37**	13.37	68.96
Smoothing (S)	13.70	9.61	53.62
Temporal modelling (TM)	33.56	2.36	36.27
Combination (S & TM)	32.49	**1.57**	**33.09**

In general, we could say that smoothing provides better results at the frame level, while temporal modelling performs better at the period level. This fact should be mainly attributed to the way the results are obtained for these post-processing schemes, as described in Section 5.6.3. It can be seen that smoothing slightly increases MR, but is able to significantly improve results in terms of FAR (which corresponds to -2.64% and 28.12% relative improvement). Temporal modelling, on the other hand, reduces even stronger the FAR error (by 82.35%, relatively) and is not performing well in terms of MR metric, but gives better period-level score, which is more important for the medical application. Although there is a big difference between frame-level metrics, in terms of the absolute number of frame errors the deterioration of MR results is smaller than the improvement of FAR results.

The best PB-ERR metric score corresponds to the combination of both smoothing and temporal modelling (*S & TM*), which is 52.02% relatively better than the baseline. Moreover, it should be noted that the combination of both schemes outperforms the temporal modelling not only in terms of PB-ERR, but also at the frame level.

5.7 Neural network based approach

The systems[1] proposed in this section are based entirely on the use of Neural Networks (NNs), and no specific feature extraction schemes or signal processing techniques are employed at the input. The idea is to let NN learn by itself the specific spectro-temporal structure of alarms and their particular discriminative properties. Following this idea, two detection systems are proposed, in which the topology of the net is focalized to either a *generic* or a *particular* type of alarm sounds. Moreover, partially connected hidden layers with limited weight sharing are explored for weighting the input information in time and in frequency and, perhaps, thus emphasizing the alarm-specific properties. Due to the

[1] The detection systems described in this section are the result of fruitful discussions in the scope of the master thesis work carried out by Alex Peiró Lilja, which was supervised by the author of this PhD thesis.

limited amount of available annotated data, the employed network structures have small scale, thereby the number of network parameters to train is constrained.

5.7.1 Generic system

The network structure employed by this system is designed for a *generic* type of alarms, i.e. is used for all the alarm classes. This means that no specific knowledge or assumptions about the properties of alarm classes are exploited by this system.

The acoustic signal is split into frames. Each frame contains 2048 samples of the signal, and the frames are half-overlapped. The logarithmic spectral amplitude of each frame is used at the input of the network, thus, the input size is 1024 units. The input features are mean-variance normalized, and the mean and variance values calculated on the training data are also applied to the testing data.

The hidden and output units have the sigmoid and the softmax activation functions, correspondingly. Stochastic gradient descent is used for network optimization, and the binary cross-entropy objective function is employed. The number of epochs is 70 and the minibatch size is set to 10. The learning rate and momentum parameters are set to 0.01 and 0.9, respectively. The training data is balanced with regards to classes by randomly selecting samples of non-alarm class. No unsupervised pre-training of NN is performed. For simplification purposes, the described network configuration is used in all the experiments.

In the baseline setup the whole spectral frame is introduced at the input, and it has been observed that even using only 8 hidden units the NN is very prone to overfitting. For that reason, max pooling is used at the input to compress the spectral representation and reduce the number of parameters to be trained. The max pooling strategy is borrowed from convolutional neural networks, where it is used to reduce spectral variance [148]. It seems to fit well our task as high spectral peaks corresponding to alarms are supposed to be preserved. In this work, we use either uniform (basically as the one depicted in Figure 5.7) or mel-scale filterbank based (with 60 filterbanks) distribution of pooling filters, and the input layer size is reduced to either 256 or 60 units, respectively.

Further, we explore the inclusion of partially connected hidden layers, which apart from reducing the number of network parameters, perhaps, could also exploit the information about alarms in frequency and in time. In fact, these layers correspond to simple unidimensional convolutional layers of one filter with limited weight sharing and no overlapping.

Figure 5.7 shows partially connected hidden layer for frequency weighting, where the weighting filters are uniformly distributed and are non-overlapping. Since alarms occupy narrow frequency regions, the width of the filter is relatively small and only spreads 4 input units. Note that no information about the alarm-specific frequencies is provided to the layer.

Figure 5.8 shows partially connected hidden layer for weighting the spectral information in time, as used in our experiments. The temporal context of several frames is exploited by this layer, and the smoothed representation of the spectral frame is obtained.

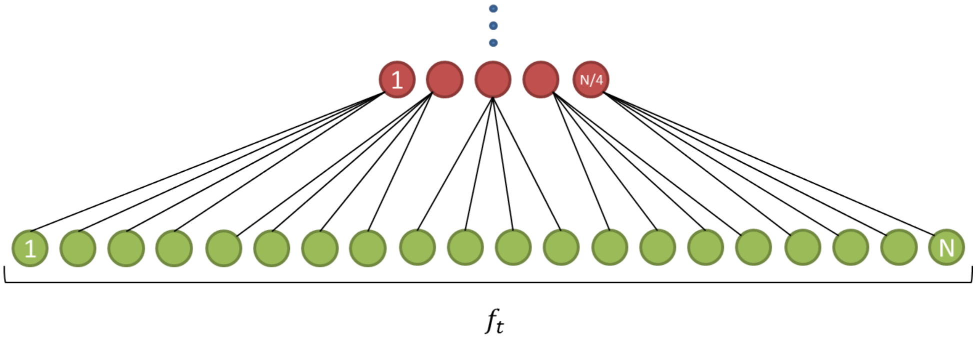


Figure 5.7: Partially connected hidden layer for weighting in frequency.

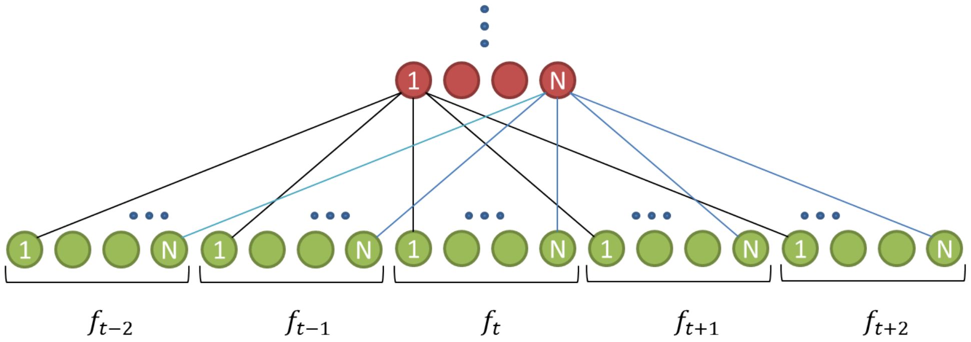


Figure 5.8: Partially connected hidden layer for weighting in time.

5.7.2 Particular system

As in the knowledge-based approach described in Section 5.6, in this system it is assumed that the particular spectral and temporal properties of alarms are known. In fact, only the spectral information is used, since the inclusion of the temporal information (i.e. signal and silence interval duration) would require much more network weights to be trained and is not feasible (in our experiments, the results were clearly worse). The spectral information is exploited at the input of the network, where the input features are the logarithmic spectral amplitudes at the alarm-specific frequency bins and bins around them. Obviously, no pooling strategies and partially connected layers for frequency weighting are used by this system.

5.7.3 Experimental results

The development of the systems was first carried out for the alarm class a8, which is the class that has only one version and the largest number of samples in our database (see Table D.1). Nevertheless, the proposed detection systems perform in the same manner when being extended to the other considered classes, with some minor exceptions. Therefore, we first present the results for the generic and particular systems for alarm class a8, and then the best setups obtained are extended to all the considered alarm classes. No post-processing schemes are employed for presenting the results, and the EER value is reported.

Table 5.7 contains the baseline results as well as the results obtained by the generic system employing different pooling strategies for class a8. We also perform the comparison of the uniform max pooling with average pooling. For each experiment, the number of trained NN parameters is provided. Note that depending on the alarm class, the balanced training data contains from 1744 frames (for class a16) to 7083 frames (for class a1), and, in particular, 5320 frames for class a8. It can be seen that none pooling strategy improves the baseline results, although the uniform max pooling provides comparable results with roughly 4 times less number of parameters. The results using mel-scale max pooling are worse (by 6.66% relatively compared to the baseline results), most probably due to the stronger information reduction it performs. As supposed, max pooling is able to better empasize the alarm-specific frequencies rather than simple averaging.

Table 5.7: Alarm detection performance obtained by the generic system with different pooling strategies

Pooling	Evaluation metrics (%) MR = FAR	Number of parameters
None (baseline)	**23.44**	8208
Average	30.24	2064
Uniform max	23.55	2064
Mel-scale max	25.00	496

Table 5.8 shows the generic system performance when partially connected hidden layers are included in NN structure for frequency weighting (FW) and temporal weighting (TW). In order to keep the number of trained parameters small, max pooling is applied at the input of the NN. Uniform max pooling is used before the partially connected hidden layer for weighting in frequency, while mel-scale max pooling is used before the partially connected hidden layer for weighting in time, and in both cases the detection results are improved. In particular, weighting in frequency and in time yield 22.06% and 40.83% relative improvement over baseline, respectively, which suggests that the temporal context is more important. Note that no extra hidden layers are used in these experiments.

Table 5.8: Alarm detection performance obtained by the generic system with partially connected hidden layers

Layer type	Evaluation metrics (%) MR = FAR	Number of parameters
FW	18.27	384
TW	**13.87**	420

Table 5.9 presents the results for the particular system, and it can be seen that improvements over baseline are obtained in all the setups due to the specific knowledge about alarms being used. We first try the input that contains spectral bins corresponding to alarm-specific frequencies and ± 1 or ± 2 neighbouring bins. Thus, for the alarm class a8, the input size is equal to 15 or 25 units, respectively. Also, a fully connected (FC) hidden layer with 8 units is used. There is no significant difference in detection results between both types of input, still the input with ± 2 bins provides better performance and may be more suitable when extending the results to other classes. It is further used with the partially connected hidden layer for weighting in time (TW), which yields 19.18% relative improvement compared to the network structure with a fully connected hidden layer. Moreover, when an extra fully connected hidden layer (FC) is introduced, the detection error drops to only 10.69%.

Table 5.9: Alarm detection performance obtained by the particular system

Input, layers	Evaluation metrics (%)	Number of
	MR = FAR	parameters
± 1 bin, FC	16.44	146
± 2 bins, FC	16.37	226
± 2 bins, TW	13.23	202
± 2 bins, TW + FC	**10.69**	376

Table 5.10 summarizes the results for best-performing system setups, which are higlighted in bold in Tables 5.7, 5.8 and 5.9, over all the considered alarm classes. Although the average results for the generic and the particular systems are a bit worse that the ones obtained for alarm class a8, the same general conclusions about the performance of the various systems hold true.

Table 5.10: Average alarm detection performance obtained by the neural network based systems over all alarm classes

System	Evaluation metrics (%)
	MR = FAR
Baseline	23.42
Generic	17.76
Particular	11.13

5.8 Chapter summary

In this chapter, our work on the problem of automatic detection of acoustic alarms in a NICU environment was reported. Several detection systems were proposed for the detection of particular types of alarm sounds, which are based on the approaches that deal with the problem from different perspectives.

The detection system based on a signal processing approach was presented first, where matched filter and morphological tools were employed. The system requires that a sample of the alarm class to be detected, including all its versions, is available. That sample is used as a reference signal at the matched filter stage, and should be an "ideal" representation of the alarm class, with which the matching will be performed. In order to enhance the reference signal, the system includes a prior filtering step, which suppresses the content at all frequencies not relevant to the alarm class. The system performance depends greatly on the proper choice of the decision threshold U. According to experimental results, the inclusion of the EOP stage played a crucial role in improving the system performance.

Basically, the detection system following the signal processing approach is deterministic and only employs several training samples for each alarm class, where the exact number of samples corresponds to the number of alarm versions. Note that the decision threshold choice is based on the non-alarm training data. The other two detection systems employ statistical modelling of the training data, where a multitude of alarm samples is used. In fact, for these two systems, and especially for the system based on neutral networks, the amount of training data available is an important factor.

The knowledge-based detection system strongly relies on the feature extraction process, and is based on exploiting the knowledge about the particular spectro-temporal properties of alarms. First, the spectral information about alarms is captured at the feature level. The best-performing features are based on the output of sinusoidal detector complemented by the amplitude structure information at the alarm-specific frequency regions. Second, after the statistical modelling of that features, the temporal information is included at the post-processing step. In particular, the period probability estimate is obtained at each frame by aggregating the log-posterior probabilities from statistical models along the signal and silence intervals in the alarm period. It has been shown that the detection system benefits largely from the introduction of both spectral and temporal information, and both were important to improve the detection performance.

At the statistical modelling step, due to the scarcity of data, simple detectors based on GMMs outperformed pre-trained NNs. Also, the experimental results showed that, as can be expected, the system is able to better detect alarm samples that are associated with higher SNR values, and this matches well the fact that the effect of the auditory stimulus on a preterm infant is more noticeable for those alarm samples, since they likely show a higher amplitude.

Last but not least, a neural network based approach was explored. Following this approach, two detection systems were developed: 1) a generic system where no specific knowledge about the alarm properties is used to construct the network; 2) a particular system where, similarly to the knowledge-based approach, the information about particular spectral properties of alarms is incorporated. It has been shown that the particular detection system clearly outperforms the generic system. On the other hand, while the particular system has to be adapted to each alarm class, the generic system has the advantage of using the same network structure for all the classes, so it can be easily extended to new alarm classes.

Due to the limited amount of data available, the employed network structures must have small scale, thereby the number of network parameters to train is constrained. Taking into account this consideration, two types of partially connected layers with limited weight sharing are explored for weighting the input information in time and in frequency. Apart from reducing the number of network parameters, these layers also reduce the time complexity of network training. It has been shown that including such layers in the neural network provides better detection results than employing only fully connected layers. Moreover, for both generic and particular systems, the layer exploiting temporal context improved the results to a larger extent. It should be noted that, according to experimental results, there is no clear dependency between the number of network parameters and the detection performance.

The detection errors obtained by the proposed systems are rather high, which can be attributed to both the fact that the real-world NICU environment is noisy and to the scarcity of available data. In general, all the detection systems obtained the worst metric scores for the alarm classes that share similar spectro-temporal properties and the discrimination between these classes seems to be difficult. E.g. for the signal processing based system these alarms were associated with high cross-correlation values at the matched filter stage.

Table 5.11 summarizes the results obtained by the proposed detection systems, where either smoothing (S) or temporal modelling (TM) post-processing is applied. Note that, for comparison purposes, smoothing, which uses the information about the durations of signal and silence intervals in alarm period, is also applied to the generic NN based system.

Table 5.11: Alarm detection performance obtained by systems following the three alternative approaches

Approach	EER	Evaluation metrics (%)		
		MR	FAR	PB-ERR
Signal processing	–	43.25	**0.54**	**34.26**
Knowledge + TM	13.37	33.56	2.36	36.27
Knowledge + S	13.37	**13.70**	9.61	53.62
Generic NN + S	17.76	33.52	7.94	54.80
Particular NN + S	**11.13**	17.72	5.22	53.75

Overall, as can be expected, the systems performing detection at the frame level provide better performance in terms of the frame-level metrics, and the same applies to the systems operating at the period level, which provide better results in terms of the period-level metric. Also, it can be seen that the inclusion of the knowledge about alarm properties is beneficial for the detection results.

Apart from presenting the detection systems and their respective experimental results, some considerations about the importance of proper metrics design were outlined in this chapter. In our case, several rounds were required to define metrics that are adequate for the considered detection problem and medical application.

Chapter 6

Vocalization detection

6.1 Chapter overview

This chapter presents our work on the automatic system for detection of vocalization sounds, which encompass all sounds produced through a vocal tract either by infant or adult (i.e. speech, cries, laugher, cough, etc.). Due to the rich multisource nature of the Neonatal Intensive Care Unit (NICU) acoustic environment, various sound events are usually taking place simultaneously. Considering vocalizations, the temporal overlaps with other sounds are even more probable due to their extensive presence. Moreover, a stationary ventilation noise which spreads over a wide frequency range is strongly present in the recordings. These factors make a vocalization detection in a NICU rather challenging.

In order to have a more robust detection, the proposed system includes a pre-processing enhancement step that reduces the presence of irrelevant sounds prior to detection. Several techniques are investigated for vocalizations enhancement, which are based on either the widely used Spectral Subtraction (SS) algorithm or the Non-negative Matrix Factorization (NMF) algorithm, or a combination of both. In this case, SS is used to attenuate the stationary ventilation noise, while NMF, which is more suitable for audio enhancement in the presence of non-stationary noises, segregates vocalizations from the other interfering sounds and noise. The vocalization sounds are further detected from the enhanced audio signal.

The chapter is organized as follows. In Section 6.2 the evaluation setup is explained. Section 6.3 provides details on how the pre-processing step of the detection system is implemented, and Section 6.4 contains the description of the detection system itself. The experimental results are presented Section 6.5.

6.2 Evaluation setup

The evaluation setup for the vocalization detection is very similar to the one used for the alarm detection task. The experiments were carried out with the part of the recorded database that was annotated. The total amount of data used is around 40.2 minutes, and 56.7% of this time is labelled as vocalizations. In total there were 35 files from different recording sessions (the concrete scenario files used can be found in Table 3.3). Only recordings made with the microphone placed outside the incubator were used to keep homogeneous experimental conditions, and also because this microphone is closer to the vocalization sources. The original 44.1 kHz recordings were downsampled to 16 kHz. As for the alarm detection task, a 10-fold cross-validation scheme was applied, including for NMF processing.

In our preliminary work reported in [149] an event-level metric (based on F-score) was used to assess the system performance. That metric was defined mainly for presenting the first results to the medical staff, and in that case the temporal precision of the results was not required. However, since the main application of the vocalization detection system is automatic annotation of the audio data, time precision is of concern. On the other hand, since the set of vocalization subclasses is quite diverse and includes events related to speech that may span over long periods of time, the definition of a common event unit is not much meaningful. Therefore, for vocalization detection task the recognition performance was only evaluated at the frame level and, as for alarm detection, the Missing Rate (MR) and the False Alarm Rate (FAR) metrics were defined as follows:

$$MR = \frac{N_M}{N_V}, \qquad FAR = \frac{N_{FA}}{N_{NV}}, \tag{6.1}$$

where N_M and N_{FA} are the number of misclassified frames for vocalization and non-vocalization class, respectively, and N_V and N_{NV} are the total number of vocalization and non-vocalization frames, respectively.

6.3 Enhancement techniques

6.3.1 Spectral subtraction

In the standard SS, the details of which are given in Section 2.6.1, the noise estimate is obtained once from the first frames of the input audio. But since the annotation data is not available, it is not guaranteed that there are no vocalization sounds present in that beginning segment. On the other hand, since ventilation noise is stationary and is present throughout the recording, we propose to use as noise estimate the average spectrum of the whole input signal. Alternatively, the noise estimate can be obtained and updated along the input signal, and such approach is able to deal better with highly nonstationary noise environments. In this work, we employ the Minima-Controlled Recursive-Averaging (MCRA) algorithm, which is described in Section 2.6.1.

The following parameter setup is used: the processing is performed on Hann-windowed half-overlapped 64 ms frames with $\gamma = 2$. For standard SS, $\alpha = 0.01$ $\{0..3\}$[1], $\beta = 0$ $\{0..1\}$ and the noise estimate is obtained from the first 7 frames of the audio recording (which roughly corresponds to 200 ms); for SS with the average spectrum noise estimate, $\alpha = 0.2$ $\{0..1\}$ and $\beta = 0$ $\{0..1\}$; for SS with MCRA, $\alpha, \beta, \alpha_d, \alpha_s, \alpha_p$ are equal to, correspondingly, 1 $\{0..1\}$, 0.01 $\{0..0.1\}$, 0.2 $\{0.2..0.95\}$, 0.9 $\{0.7..0.95\}$ and 0.1 $\{0.01..0.7\}$.

6.3.2 Non-negative matrix factorization

Following the general NMF framework, described is Section 2.6.2, the bases matrix was trained beforehand on the training data. In the case of binary vocalization detection, we consider $S = 2$ sources corresponding to vocalization and non-vocalization classes. The global bases matrix $W_{train} = [W_V; W_{NV}]$ consists of the bases trained for each class, respectively. The enhanced audio signal is then reconstructed using only the vocalization spectra $\hat{V}_V$ and the phase of the original input audio.

In the basic case, the spectrum of each source can be obtained by multiplication of the source bases by the corresponding activations, i.e.

$$\hat{V}_i = W_i H_i, \quad i \in [1..S]. \tag{6.2}$$

Commonly, an approach similar to Wiener filtering is applied to reconstruct each source:

$$\hat{V}_i = \frac{W_i H_i}{\sum_i W_i H_i} \otimes V, \tag{6.3}$$

where multiplication $\otimes$ and division operations are element-wise [112].

The implementation of NMF described in [116] is used, with the following parameter setup: the input matrix V is a magnitude spectrogram computed on Hann-windowed frames of 32 ms length with 16 ms shift. We train $R = 25$ $\{25..100\}$ bases per class, where each base corresponds to a vector of dimension $F \times 1$. The sparsity parameter λ is set to 0.01 $\{0..2\}$. At the training and testing time we use up to 25 iterations.

6.3.3 Combined approach

We want to exploit the complementarity that may exist between the SS and NMF algorithms, by investigating several combinations of the techniques.

Firstly, we try the combination in which SS and NMF are applied consecutively. In this case, the audio data is previously processed by SS in order to attenuate the ventilation noise, and then this enhanced audio is used as training data for NMF. Alternatively, NMF is applied prior to SS processing.

[1] The range of values on which each parameter was optimized is shown in curly brackets. Note that the parameter tuning was not exhaustive and there may be more optimal parameter configurations, but, as observed during tuning, no large improvement should be expected and the general relation between the technique performance will hold.

Secondly, we employ NMF to obtain the noise spectrum estimate $|\hat{D}(n,k)|^{\gamma}$ for SS technique. Contrary to NMF based pre-processing, where the spectrum is reconstructed for vocalizations, here we obtain the reconstructed spectrum $\hat{V}_{NV}$ for non-vocalizations or, in other words, the irrelevant sounds. Each column n of the reconstructed non-vocalization spectral matrix $\hat{V}_{NV}$, which corresponds to the time frame n, is assigned to the vector $|\hat{D}(n,k)|^{\gamma}$ in (2.2). The advantage of this approach is that the noise estimate is supposed to be more accurate.

6.4 Detection system

The input signal is split into frames using a Hamming window with the frame length of 30 ms and the frame shift of 10 ms. 16 Frequency Filtered Logarithm Filter Bank Energies (FF-LFBE) features [66] along with their 16 first temporal derivatives were extracted from each frame. Therefore, the dimension of the feature vector is 32.

A Gaussian Mixture Model (GMM) based detector was used, consisting of a model for vocalization and a model for non-vocalization. Each model is a single Gaussian pdf with diagonal covariance matrix as, in our experiments, this resulted in better recognition performance than using more mixture components. With the likelihoods obtained from the two models, each frame is classified either as vocalization or non-vocalization. The decision threshold is chosen based on the Equal Error Rate (EER) criterion, so assuming that both types of errors are equally important at the frame level.

In contrast to GMMs, which is a generative classification model, we also perform experiments employing a discriminative Support Vector Machines (SVM) based classifier. SVMs aim at maximizing the margin between the classes and have an advantage of using only the training samples that are the closest to the decision surface, which can be beneficial when a limited amount of training data is available. In this work, both linear and Radial Basis Function (RBF) kernels are employed. Before being fed to the classifier, the input features are mean-variance normalized; the mean and variance values calculated on the training data are also applied to the testing data.

Optionally, smoothing (via majority voting) is applied to the string of output labels. The length of the smoothing window was optimized with regards to the recognition performance and is equal to 31 frames.

6.5 Experimental results

The baseline system performance is presented in Table 6.1 as a function of the number of Gaussian components used. The EER, which corresponds to both MR and FAR metrics having the same value, is reported when no post-processing (i.e. smoothing) is applied. It can be seen that the increase of the number of Gaussians seems to be detrimental to the detection performance, therefore only one Gaussian is used in subsequent experiments. Furthermore, in all cases smoothing the classifier output improves the detection results in terms of both metrics, yielding up to 12% relative improvement in the best case. The system performance will be further compared to the smoothed baseline.

Table 6.1: Vocalization detection performance obtained by the baseline system

Number of Gaussians	No post-processing	Smoothing	
	Evaluation metrics (%)		
	MR = FAR	MR	FAR
1	**32.90**	**29.64**	**29.68**
2	35.40	31.19	31.23
4	36.55	32.17	32.49
8	39.33	34.15	37.23

Table 6.2 shows the detection performance of the system when different pre-processing schemes are applied prior to detection. Several of the proposed schemes are able to improve the baseline results.

First of all, the results for the SS and NMF techniques applied separately are presented. It can be seen that applying the standard SS leads to the performance loss (by 3.74% and 5.93% relatively in terms of MR and FAR, respectively). This may be explained by the fact that some of the recordings contain vocalization sounds at the beginning and the obtained noise estimate is not accurate, which may cause the distortion of vocalizations. This explanation is also justified by the optimal parameter values obtained ($\alpha = 0.01, \beta = 0$) which basically corresponds to not doing almost any subtraction.

On the other hand, SS using the average spectrum noise estimate (*SS average*) and SS with the MCRA algorithm for the noise estimation (*SS + MCRA*) are both able to improve the baseline result due to the better noise estimate obtained. In the case of *SS average* the relative improvement is of 7.35% and 6.27% in terms of MR and FAR, respectively, showing that the average noise estimate is able to represent the ventilation noise. It is also reflected in the higher optimal value of $\alpha = 0.2$. And as *SS + MCRA* pre-processing results in a more accurate noise estimate, it yields even higher relative improvement: 11.13% in terms of MR and 13.58% in terms of FAR metric scores.

As for NMF-based pre-processing the gain is not so obvious. Employing the basic technique for vocalizations reconstruction (*NMF basic*) doesn't bring any improvement to the baseline result; conversely, a relative loss of 1.18% in terms of MR and of 0.98% in terms of FAR is obtained. On the other hand, NMF with Wiener-like reconstruction (*NMF Wiener*) improves the results, but to a small extent: by 1.79% and 4.78% relatively in terms of MR and FAR. The reason for NMF not performing so well may be the fact that the strong ventilation noise and other sounds are present in the training data of both vocalizations and non-vocalizations, thus reducing the discriminative power of the trained bases.

The last part of the table contains the detection results for different technique combinations: when

Table 6.2: Vocalization detection performance obtained by the GMM-based system with different pre-processing schemes

Pre-processing	No post-processing	Smoothing	
	Evaluation metrics (%)		
	MR = FAR	MR	FAR
None	32.90	29.64	29.68
SS standard	33.92	30.75	31.44
SS average	31.20	27.46	27.82
SS + MCRA	29.39	26.34	25.65
NMF basic	33.56	29.99	29.97
NMF Wiener	32.12	29.11	28.26
SS → NMF	31.99	27.44	27.96
NMF → SS	**28.31**	**24.44**	**25.26**
SS + NMF	33.80	30.54	30.19

SS and NMF are applied consecutively to the audio signal (*SS → NMF* and *NMF → SS*) and, also, when NMF is used to obtain the noise estimate for SS (*SS + NMF*). Note that the best setups of SS and NMF techniques are used for the audio-based combinations, namely, *SS + MCRA* and *NMF Wiener*. For *SS + NMF* pre-processing, *NMF Wiener* is used for the noise estimation, the parameters of SS are set to $\alpha = 0.2, \beta = 0$ and the frame length is set to 64 ms. In the rest of cases the optimal parameter setups obtained for each technique separately are kept.

The best detection results are obtained when SS is applied to the audio signal pre-processed with NMF (*NMF → SS*) and in this case the relative improvement achieved is 17.54% in terms of MR and 14.89% in terms of FAR. The detection results for the alternative pre-processing sequence (*SS → NMF*) are worse than using SS alone (only 7.42% and 5.80% relative improvement in terms of MR and FAR, respectively, compared to the baseline results). This may be due to the fact that SS processing introduces a musical noise to the output audio which, like it occurs with the ventilation noise, is not beneficial for bases training. It can also be seen that *SS + NMF* combination is not outperforming the baseline setup. At least partially, this can be attributed to the fact that the processing window length used in SS is not optimal for NMF.

The Detection Error Tradeoff (DET) graphs are shown at Figure 6.1 for the best performing setups of SS, NMF and their combination when no post-processing is applied. It can be seen that the combination of both techniques outperforms each one of them at all the operational points of the curve except for the ones where FAR is very low.

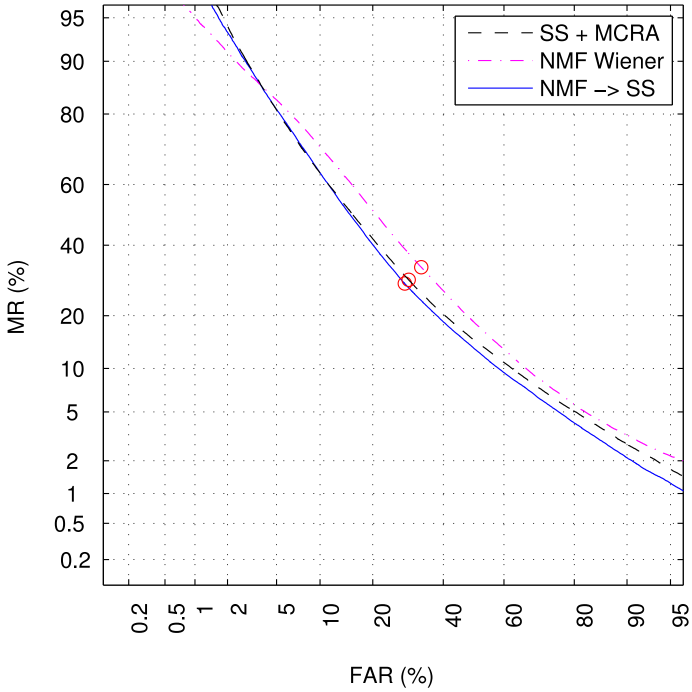


Figure 6.1: The Detection Error Tradeoff (DET) graphs for the three best performing setups. Circles correspond to EER points.

In Table 6.3 we provide the detection results for the SVM-based classification, both with linear and RBF kernels. Either no pre-processing or the *NMF* → *SS* pre-processing, which gave the best results for GMM-based classifier, is applied. For linear SVM, the parameter C, which controls the tradeoff between the training error and the margin, is set to 1e−4 {1e−5..1}. For SVM with RBF kernel, this parameter equals to $C = 0.05$ {1e−4..1}, and the parameter γ of RBF is set to 0.001 {0.0001..0.25}.

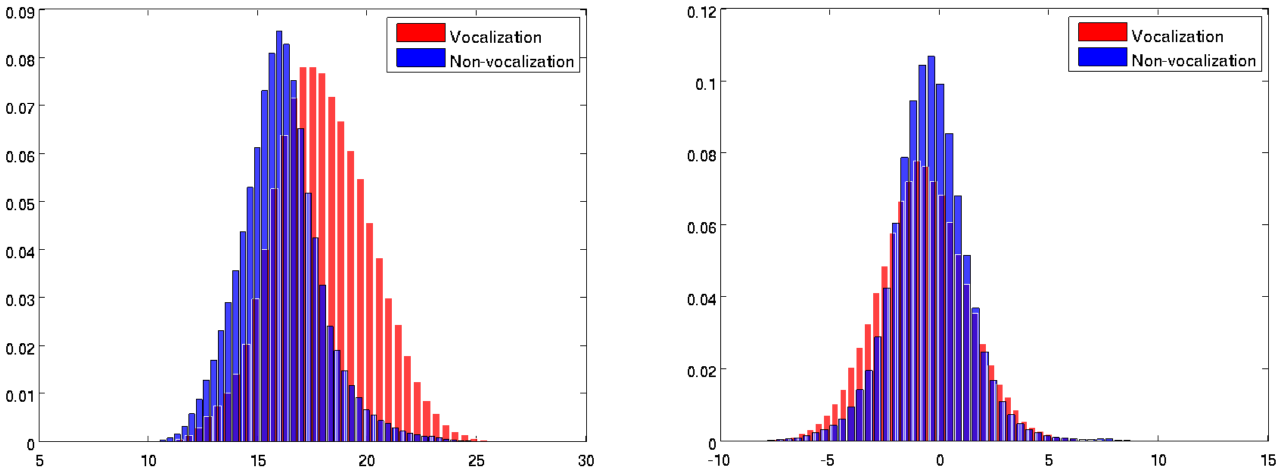
Table 6.3: Vocalization detection performance obtained by the SVM-based system

Pre-processing, kernel	No post-processing		Smoothing	
	Evaluation metrics (%)			
	MR	FAR	MR	FAR
None, linear	30.65	37.60	27.83	36.59
None, RBF	30.67	37.24	27.93	36.08
NMF → SS, RBF	**25.09**	**35.2**	**22.07**	**33.94**

It can be seen that there is no significant difference in the detection results for the two types of SVM kernel functions on our data, and the RBF kernel only slightly outperforms the linear one. I.e. with smoothing post-processing, the total error (MR+FAR) for linear kernel equals to 64.42%, while for

RBF kernel it is equal to 64.01%. Similarly to the GMM-based system, these results are improved when the pre-processing step is added, although the overall improvement is somewhat smaller. In particular, the relative improvement in terms of MR and FAR is equal to 20.98% and 5.93%, respectively.

Comparing the results for the two types of classification models, a generative GMM and a discriminative SVM, it can be seen that SVM-based system is not able to outperform the GMM-based one. The total error for the GMM-based and for the SVM-based systems is equal to 49.7% and to 56.01%, respectively. Perhaps, this is due to a strong overlap between the vocalization and non-vocalization classes. Figure 6.2 provides an illustration of it, where the distribution of 1st and 4th FF-LFBE features over all the recording sessions is provided for both classes. Most of the features behave like the 4th one, so indeed there is a strong overlap between the classes, and the 1st feature is rather an exception and seems to be the most discriminative.



(a) *(b)*

Figure 6.2: Distribution of (a) 1st and (b) 4th FF-LFBE features for vocalization and non-vocalization classes.

6.6 Chapter summary

In this chapter, we presented a system for automatic detection of vocalization sounds. This type of sounds is the most represented in the NICU environment recordings, and includes various subclasses like voices (background and foreground), baby crying, laughter and cough. In this work, the binary vocalization detection problem is considered, where the decision is taken between the generic vocalization and non-vocalization classes.

The proposed system includes a first step of non-vocalization sounds reduction, based either on NMF or SS or their combination. In general, the detection system benefits from introducing the enhancement step, which in the best setup leads to 17.54% relative improvement over the baseline. It has been shown that for our data NMF alone as the pre-processing step is not performing as well as applying SS alone, most likely due to the predominant presence of the ventilation noise. Anyhow, when NMF is applied before SS denoising for pre-processing, the best detection performance is obtained. Two types of classification models were explored, namely, a generative GMM based and a discriminative SVM based, and in our experiments the GMM-based system outperformed the SVM-based one.

The obtained detection error is still quite high due to the complexity of the detection problem in a real-world hospital environment and the scarcity of data. But it should be taken into account that in this work the focus was mainly put on the pre-processing step, and, apparently, better results could be achieved by improving other steps, like feature extraction or post-processing.

Chapter 7

Conclusions and future work

7.1 Summary of conclusions

This thesis work presents a rather new and challenging research area of automatic analysis: the acoustic environment of a preterm infant in a Neonatal Intensive Care Unit (NICU). The work carried out can be briefly summarized in the following lines of work: 1) audio database production, from acoustic scenarios definition and audio acquisition to annotation production; 2) overall description of the NICU acoustic environment from the audio recordings; 3) development of automatic detection systems for some relevant acoustic events. In the following, several contributions of the thesis work along these lines are summarized.

To our knowledge, the database produced in this thesis work is the first annotated audio database acquired in a NICU environment. During the work on the database production, key specifications (e.g. recording setup and guidelines, labelling protocol) were designed and the whole framework of the audio database production for the NICU environment was set up. Due to the pioneering character of the work in that acoustic environment, the abovementioned specifications had to be designed from scratch in close collaboration with medical and engineering staff from the Hospital Sant Joan de Déu Barcelona. Several rounds were required, which allowed refinement of these specifications based on the obtained experience, but also implied more effort for making possible that the produced database follows a unified protocol. A number of recording sessions were carried out in the NICU following the designed guidelines. In total, the produced database includes more than 1.5 hours of audio data. The annotations cover roughly half of it; they were obtained via manual annotation in two stages, and posteriorly revised.

Using the acquired audio recordings, an exploratory acoustic description of a NICU environment was performed, which shows its strong acoustic diversity. Unlike most works previously reported in the literature, the whole content of the audio signal was analysed, and, besides the usual measurements of sound pressure levels, the identity and the spectro-temporal properties of sounds were described. An extensive list of more than 60 acoustic events found in the NICU environment was presented. To structure the whole diversity of sounds, a general sound taxonomy of a NICU environment was proposed, where the three major acoustic groups defined are *tone*, *vocalization* and *other*. Further, a detailed description of the types of sounds most represented in recordings (namely, equipment alarms, vocalization sounds and equipment noise) was carried out. Besides, the set of considered acoustic scenarios was analysed. The differences between the scenarios appeared to be not significant in comparison to the inter-session variability, and the factors of that were outlined.

The main contribution of this thesis lies in the work on development of acoustic event detection systems for some relevant types of sounds from the NICU acoustic environment. In particular, the detection of equipment alarms and vocalization sounds was targeted, since, according to the results of acoustic description, these types of sounds are the most common.

Regarding the task of automatic detection of acoustic alarms, several detection systems were proposed. These systems are based on approaches that deal with the task from different perspectives:

1) A signal processing based approach that employs matched filter and morphological tools. The detection system following this approach is deterministic, and its performance depends greatly on the proper choice of the decision threshold. Nevertheless, the performance of this system is comparable to the performance of the detection systems that employ machine learning.
2) A knowledge-based machine learning approach, which integrates the knowledge about the peculiar spectral and temporal properties at different stages of the system. The spectral information is captured in a feature vector, which in the best-performing case includes the output of the sinusoidal detection along with the amplitude structure information, both obtained around alarm-specific frequencies. The temporal information is incorporated at the post-processing step by aggregating the frame-level posterior probabilities, obtained from statistical modelling, along the duration of signal and silence intervals in every alarm period. According to experimental results, the detection system benefits largely from the introduction of both spectral and temporal information.
3) An approach based on neural networks in which all stages of the detection system are machine learning based. Two detection systems were developed following this approach: a *generic* system, where no knowledge about the alarm properties is used, and a *particular* system, where the information about spectral properties of alarms is incorporated, similarly to the abovementioned knowledge-based approach. Also, two types of partially connected layers with limited weight sharing were explored for weighting the input information in time and in frequency, and the temporal weighting improved the detection performance to a larger extent. In our experiments, the particular system clearly outperformed the generic one.

Note that the developed systems consist of several individual detectors, each one dealing with a particular alarm class, and the specific class of each detected alarm occurrence was specified. It has been shown that the discrimination between some of the alarm classes is difficult due to the fact that these alarms share similar specto-temporal properties.

Additionally, the development of the detection systems required a design of proper evaluation metrics and the targeted medical application has been considered for that purpose. In particular, for the alarm detection task an event-level metric was proposed that is based on the alarm period unit.

Regarding the task of automatic detection of vocalization sounds, the work on the detection system focused on the pre-processing step of non-vocalization sounds reduction. In particular, several techniques were investigated for vocalization enhancement, which are based either on non-negative matrix factorization or spectral subtraction or their combination. It has been shown that the detection system clearly benefits from introducing the enhancement step, and the combination of non-negative matrix factorization followed by spectral subtraction at the pre-processing step provided the best detection performance. Also, the two types of classification models, namely, a generative Gaussian mixture model based and a discriminative support vector machines based, were assessed for this task. The binary detection problem (vocalization vs. non-vocalization) was considered.

The detection errors obtained by all the developed detection systems are still rather high, which could be attributed to the rich multisource, noisy nature of the real-world hospital environment and to the scarcity of available annotated data. However, these results encourage further advances and sophisticated solutions for the challenging problem of acoustic event detection in a NICU environment.

7.2 Future work

The list provided below contains the most important points requiring improvements as well as several directions for future work that seem promising.

7.2.1 Collection of a larger database

The collection of a larger database is essential to advance in all the directions of future work. Apart from acquiring more data, the work on a larger database production could include completing the labelling of already recorded audio data and extending the labelling protocol to more acoustic event types. In particular, the following research lines will benefit from larger data:

1) A large database recorded to account for diverse conditions of the NICU is useful for performing a more comprehensive acoustic analysis of its environment. Detailed annotations, where the labels possibly cover all the acoustic events present in the recordings, will be needed for this task.
2) The incorporation of more data will allow further improvements in detection results of the developed systems. As greater amount of samples of various acoustic events will be obtained, detection systems could be developed for the types of sounds not considered before, e.g. other acoustic alarms, finer vocalization classes, other relevant acoustic events (like telephone, steps, door slam, etc.).
3) Recordings with more newborns would be necessary for investigating the influence of the auditory stimuli from the NICU environment on a preterm infant (see Section 7.2.2).

7.2.2 Analysis of sound impacts on a preterm infant

As mentioned before, it is medically relevant to correlate the presence of particular sounds with the preterm physiological variables in order to investigate how a preterm infant reacts in a short term to an auditory stimuli from the NICU environment. Note that the criteria of inclusion of a preterm baby in the study would be the same as described in Section 3.2 for database acquisition.

Various parameters and influencing factors have to be taken into account and related in such investigation:

1) Clinical parameters collected at the outset. These in general include gestational age, weight at birth, sex, anthropometry, patient perinatal history (e.g. mother's age, type of delivery, need and type of resuscitation at birth, etc.), co-morbidities, treatment received, type of mechanical ventilation.
2) Physiological variables, which include haemoglobin saturation, heart rate and respiratory rate. It should be noted that, according to our experience, the extraction of these parameters from the monitoring equipment may be not a trivial task as it requires either agreement of collaboration between the hospital and equipment manufacturers or bearing substantial financial costs of purchasing a specific software.

3) Environmental exposure parameters, which include:
    a) Acoustic parameters, i.e. sound pressure levels, spectro-temporal properties and identity of sounds. At the initial stage of the study, when detection systems for automatic sound identity annotations may not be available yet, Sound Pressure Levels (SPLs) and spectro-temporal features alone can constitute acoustic parameters representation.
    b) Micro-environmental parameters like luminosity, temperature, humidity.
4) State of consciousness, which could be estimated from video recordings of a preterm infant. This data could also be used for investigation of sound impacts on preterm's sleep patterns.

In various works reported in the literature, the analysis of newborn infant responses to sound stimuli has been usually addressed from a statistical point of view, where ANalysis Of VAriance (ANOVA), ANalysis of COVAriance (ANCOVA) and Student's t-test are among the frequently used techniques. For example, the sensitivity of auditory cortex of newborns to the temporal structure of sounds was investigated in [150]; the study of hemodynamic responses in newborn infants to speech and music was reported in [151]; the electromyographic and behavioural reactivity of newborns to various sound intensity was assessed in [152]; the mean arterial blood pressure and heart rate in preterm infants were correlated with SPLs inside the incubator in [153], to list a few.

In the study [19] similar to the proposed research line, the physiological variables, cerebral and behavioural data of preterm infants were analysed in relation to SPL peaks from 5 to 15 dBA above the background level. In that study, each physiological parameter was compared using ANOVA, and post-hoc analyses using Newman-Keuls test were performed when appropriate.

While statistical techniques provide a descriptive analysis of the data, machine-learning based data mining algorithms may be useful to extract meaningful patterns and provide predictions on that data (e.g. [154]). For instance, the deep learning approach may be used to construct predictive models, linking environmental exposure, clinical and state of consciousness parameters to the physiological variables of a preterm infant.

7.2.3 Extended acoustic description

A more accurate sound taxonomy could be built automatically by employing an event-based clustering (e.g. using k-means algorithm) or by building a decision tree (e.g. based on a measure of entropy between acoustic events). Both types of analysis require a set of features that provide a comprehensive spectro-temporal description of sounds. Features like zero-crossing rate, energy, frequency sub-band log energies, fundamental frequency, spectral centroid, spectral roll-off, spectral bandwidth, spectral flux, and their evolution in time could be employed.

A possible direction of further research could be to look for typical chains of acoustic events and relationships between them. For instance, interrelationships between different equipment alarm categories (i.e. provoked by desaturation, bradycardia or apnea) and patterns in alarming in a NICU have been investigated based on heuristic techniques in [141]. Extending such research to more sound classes would help to gain a better insight about the processes in a NICU acoustic environment and to better characterise the types of sound sources present (i.e. provoked/spontaneous, evitable/non-evitable).

Besides, a detailed analysis could be provided for other relevant sound classes (like telephone, steps, chair moving, door slam, etc.), which could be further considered for automatic detection.

7.2.4 Acoustic alarm detection

The developed detection systems consist of individual detectors, each one dealing with a particular alarm class. In each system, to obtain the final decision, the independent outputs of the detectors should be combined (e.g. superimposed), and at this step the knowledge about confusion between the alarms that share similar spectro-temporal properties could be integrated. Alternatively, in order to improve the detection performance, some detection hierarchy could be considered. E.g. the alarm classes that have similar spectral structure could be detected consecutively, starting first with those having more frequency components.

Due to the complementarity that may exist between the detection systems following different approaches, the fusion of their output scores could be performed, e.g. using weighted arithmetical mean or fuzzy integral [155].

For the purposes of staff notification, an alarm sequence (see Figure 4.4) could be considered for event level evaluation. It could be based on the period-level decisions provided by the detection systems, e.g. an alarm sequence is regarded as detected if any of its periods is detected. As in the proposed block-based metric, application-specific error costs could be introduced penalising the false alarm errors. Also, miss error costs could depend on the alarm sequence duration, where a sequence containing only one period may be accidental and non-relevant and its detection may not be important, while missing an alarm sequence of a long duration may be critical.

We further provide the directions for future work for each of the proposed alternative approaches.

7.2.4.1 Signal processing based approach

Due to similarities between some of the alarm classes that lead to deterioration of the detection performance, an additional pre-processing step could be implemented to reduce the cross-correlation between reference signals of the alarm classes (e.g. by removing the common harmonics). Possible system enhancements could also concern the choice of a proper decision threshold U, in particular:

1) During training only the data segments that do not contain any alarms could be used to avoid situations when the decision threshold is too high due to the presence of similar alarms.
2) The probability density function of the morphological envelope could be employed for the threshold choice. It has been observed that such representation provides distinct peaks for alarm and non-alarm data segments, and the decision threshold would correspond to the least probable value between these two peaks.

3) More complex ways of combining the training thresholds could be explored that may lead to a more optimal testing threshold value.

7.2.4.2 Knowledge-based approach

The detection error obtained by the knowledge-based system is rather high, in part, due to the scarcity of the data available, so acquiring more annotated data is important for advances in this approach.

Feature extraction schemes that would explicitly capture a peculiar spectro-temporal structure of alarms could be tried. For instance, some modulation features that describe a temporal evolution in spectral domain (e.g. [70]) or two-dimentional Gabor-based features [73] that would be oriented along time axis could be employed. Also, a more sophisticated algorithm for sinusoidal detection could be used.

Further work could focus on improving the period probability estimation algorithm, where an adaptive thresholding could be implemented for period detection, and the way of combining it with smoothing might be investigated.

Alternatively, the information about the temporal structure of alarms could be implicitly captured by classification models like recurrent neural networks or hidden Markov models. Note that in our preliminary experiments the use of hidden Markov models was not yielding the performance improvement.

7.2.4.3 Neural network based approach

Obtaining a larger database will be the most crucial factor for the advancing in this approach as it will allow to explore more complex neural network structures. More training data could be obtained automatically through data augmentation, which is a strategy typically used for deep learning in speech recognition in order to avoid overfitting and improve robustness [156]. Following such strategy, clean alarm samples (e.g. reference signals used in signal processing based approach) would be mixed with noise samples. Note that the noise samples and the generated data in general should represent as close as possible the sound diversity of the NICU environment.

Having more training data, network structures with larger number of hidden layers could be trained, where deeper hidden layers correspond to feature representations of higher level of abstraction (Deep Neural Networks (DNNs)). Also, the system performance may improve when a larger temporal context (e.g. an alarm period for the particular approach) is introduced at the input of the network. The evaluation over cross-validation data for the early stopping strategy and dropout [157] could be employed to avoid overfitting in networks with a large number of parameters.

In our preliminary experiments, the long short-term memory networks [86], which are typically used to model a sequential data and could capture the temporal recurrence inherent to alarm sounds, didn't achieve an improvement in detection performance. More work is required in this direction to get a better insight about the obtained results, and possibly improve them. On the other hand, more filters could be included in the proposed partially connected hidden layers for performing convolution in time and in frequency. Future work could concern the implementation of other pooling schemes (e.g. [148]) that would emphasize the alarm-specific properties.

As the trained neural networks deal independently with each of the considered alarm classes, to account for similarities between some of the alarm classes, the outputs of neural networks or of their penultimate layers could be shared via a jointly trained layer. The idea of a multi-label DNN that was proposed in [158] could be used to deal with the problem of temporal overlaps between alarm sounds.

7.2.5 Vocalization detection

The binary detection of vocalizations presented in this work is a first step towards the correlation study and will have to be followed by the detection of each relevant type of vocalization sounds. In particular, more specific tasks of detecting higher intensity vocalizations (i.e. foreground speech and shouts, the detection of which may be easier), parental voices could be considered, as these sounds are supposed to affect a preterm baby the most [3]. Also, the task of baby crying detection alone could be relevant for the medical application.

More sophisticated combinations of Spectral Subtraction (SS) and Non-negative Matrix Factorization (NMF) techniques could be explored in future work. First, NMF could be used in SS with Minima-Controlled Recursive-Averaging (MCRA) algorithm for a more adequate estimation of the speech-presence probability. Second, the noise estimate obtained from MCRA could be integrated in the NMF-based enhancement. E.g. in [159] an unsupervised noise estimate was incorporated in bases training, and improvement over straightforward cascading of SS and NMF techniques was reported.

In this work, the focus in development of the system for vocalization detection was put on the pre-processing enhancement step, but improvements could be introduced at other steps. For instance, other feature extraction schemes and their combinations could be tried, which might allow better discrimination between vocalization and non-vocalization classes. A comprehensive study, like the one reported in [160], where the performance of various types of features (i.e. based on energy, harmonicity, formants, stationarity and modulation) for voice activity detection was analysed, can serve as a starting point.

Appendix A

Description of the recording sessions set

This appendix provides information about the entire set of recording sessions carried out while acquiring the database used in this thesis work. Table A.1 contains a description of each session, where sessions included in the database are marked in green.

To begin with, the recording sessions RS1 and RS2 were performed for exploring the NICU acoustic environment: to perform its general analysis and for initial acquisition of the acoustic scenarios. No external microphones were connected to the recording device during these sessions. Starting from the recording session RS3, the recording setup described in Section 3.2 was adopted, where the defined set of ten acoustic scenarios was recorded using two microphones connected to the recording device. It should be noted that the recording device setup was strictly checked only starting from the recording session RS7.

Table A.1: Information about all the performed recording sessions

Session code	Date	Time	Incubator	Duration (s)	Comments
RS1	22.01.2013	*unknown*	–	5419.21	The recording device was placed at the table closest to the center of the room (see Figure 3.1). There are 6 recordings each 5 minutes long and an hour-long recording.

RS2	20.02.2013	*unknown*	*unknown*	545.58	The recording device was placed inside the incubator. There are recordings of the following scenarios: 4 of *nursery care* (without specifying the concrete acoustic scenarios), 1 of changing medications, 2 of pediatric observation and 1 neutral.
RS3	07.03.2013	15.30	3	587.16	
RS4	20.03.2013	13.00	1	481.28	
RS5	21.03.2013	13.00	*unknown*	258.03	Only changing an oxygen sensor, cleaning respiratory secretions and measuring temperature scenarios were recorded.
RS6	22.03.2013	15.30	1	368.65	Due to the incubator design, the conventional microphones positions were not followed.
RS7	04.04.2013	16.45	–	748.34	These recording sessions were carried out in another NICU room,
RS8	05.04.2013	16.30	–	1035.61	which is to the right from the room depicted in Figure 3.1.
RS9	08.04.2013	05.00	4	382.77	Night time recording session.
RS10	11.04.2013	15.10	–	74.05	Recording of the kangaroo care.
RS11	16.04.2013	09.25	4	1030.9	
RS12	16.04.2013	13.00	3	581.87	
RS13	04.06.2013	17.10	2	683.58	
RS14	11.06.2013	17.25	4	918.19	
RS15	13.06.2013	09.05	4	659.76	
RS16	20.06.2013	09.15	4	804.03	
RS17	27.06.2013	17.00	3	658.63	The preterm infant was attended by two nurses at the same time, so all the four doors of the incubator were open during manipulations (only two doors were open in other sessions).
RS18	09.07.2013	17.20	2	429.03	

Appendix B

Neutral scenario database

An additional database of continuous acoustic environment recordings was collected in April 2014, which was recorded but not annotated. Besides audio, a video signal from the cardiorespiratory monitor screen was recorded to be able to annotate posteriorly the physiological variables. For this database, only the *neutral* scenario was considered in order to exclude the periods during which the preterm infant is exposed to tactile stimuli. Note that if the baby had to be manipulated during the recording session, recording was interrupted and a stabilization period of several minutes was used to take into account the tactile stimuli influence decay.

These recordings are supposed to capture the NICU environment during various common activities and events (see Section 3.2), possibly happening close to the incubator with which the recordings are performed. For example, in case of capturing the nursery care activities, it is preferable that a nurse attends a preterm infant in the neighboring incubator. The particular activities and scenarios that were taking place simultaneously with the recording sessions, as well as the information about these sessions can be found in Table B.1. The activities are provided in temporal order and the incubator with which the activity occurred is noted down in round brackets.

The total duration of the acquired audio data is 235.1 minutes, and part of this data was discarded during synchronization of audio and video modalities. The overall duration of the relevant audio data is 201.66 minutes, which is around 3.3 hours. All the recording sessions were performed with the incubator at position 2. Moreover, recording sessions RS20–RS22 were made with the same preterm infant. The state of concious of the preterm was noted down, but not thoroughly and without previous assessment of the medical staff.

In general, compared to the first database of acoustic scenarios, this audio database has more controlled conditions and is less diverse. Namely, only one scenario (*neutral*) is recorded, the incubator position and the medical equipment set are fixed, the doors of the incubator are always closed, the recordings are performed during limited number of time slots.

Table B.1: Information about the recording sessions of the second database

Session code	Date	Time	Duration (s)	Comments
RS19	*20.01.2014*	*unknown*	*750.37*	*Pilot recording session carried out to check the technical setup. Is not included in the database.*
RS20	01.04.2014	10.20	3941.77	Surgical intervention (3), nursery care (1), visit of parents (3), changing medications (2), nursery care(4).
RS21	02.04.2014	15.20	5875.49	The environment was very quiet and only changing medication (2) occurred during this session.
RS22	03.04.2014	09.15	2282.55	Examination (X-Ray; 3), pediatric observation (1), nursery care(1), examination (ultrasound; 1), preterm's relocation (3).

Appendix C

Labelling protocol changes

This appendix summarizes the changes in the labelling protocol used during the two stages of the annotation campaign. In particular, Table C.1 outlines the differences in the list of acoustic events considered during the first stage with regards to the final list presented in Table 3.2. The three alarms not labelled during the first stage were found after it. Also, more vocalization classes were considered at the second stage.

The first stage also had more exploratory character and some of the audio recordings (namely, 18 files) were annotated fully. Moreover, for these files the number of speakers in vocalization intervals was specified. The *events* tier contains acoustic events specific to the considered scenarios, and acoustic events in the *noises* tier are the ones recognized by the annotator during the work. Note that one of these events is cough (nn[co]), which was added to the protocol at the second stage of the campaign.

Table C.1: Differences in the list of acoustic events considered for annotation during the first stage of the annotation campaign

Tier	Label	Acoustic event
		Not considered for annotation
Alarms	a14	Incubator Atom
	a15	Infusion pump Alaris GH Plus
	a16	Monitor Philips IntelliVue MP70
Vocalizations	sh	Shout
	lg	Laughter
	co	Cough
	bci	Baby crying/voice (heard only inside the incubator)
Events	nn[bm]	Buttons of the monitor
	nn[bp]	Buttons of the infusion pump

	Annotated for a part of the recordings	
Events	dp	Diaper
	ab	Putting on/off sphygmomanometer
	hi	Hissing (rhythmic)
	nn[xi]	Squeak
	ts	Taking sensor off
	pl	Plastic wrapping
	cs	Secretions cleaning
	cb	Click of the infusion pump
	od/cd	Incubator door opening/closing
CPAP	nc	Continuous Positive Airway Pressure (CPAP) noise
Noises	nn[kn]	Knock
	nn[ch]	Chair moving
	nn[dr]	Drawer
	nn[st]	Steps
	nn[te]	Telephone ring
	nn[co]	*Cough*
	nn[ma]	Mobile phone
	nn[sp]	Spray
	nn[clh]	Cleaning hands
	nn[do]	Door slam

Appendix D

Number of annotated alarm and vocalization samples

This appendix provides information about the number of annotated samples that belong to two types of acoustic events: equipment alarms (see Table D.1) and vocalizations (see Table D.2). These are the major acoustic event classes for which the detection systems were developed in this work. In the tables, the classes *al* and *vo* correspond to the generic classes used in the binary detection. Note that the total number of samples along all classes may not be equal to that number in the binary case due to the temporal overlaps between acoustic events.

Table D.1: Number of annotated alarm samples in each recording session

Session	Class																
code	*al*	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16
RS3	67	27	0	15	0	0	0	7	4	0	5	7	2	0	0	0	0
RS4	193	0	0	17	0	0	53	0	146	12	0	0	0	0	0	0	0
RS11	104	60	6	20	0	0	7	0	3	10	19	0	0	0	0	0	0
RS12	113	1	3	0	0	0	35	0	29	7	33	0	0	0	0	0	17
RS13	203	10	4	9	0	0	0	0	147	0	0	18	0	0	0	0	36
RS14	106	5	0	0	0	0	0	42	12	0	0	0	0	0	0	0	53
RS15	78	24	0	36	0	0	0	15	3	1	1	0	0	0	0	0	0
RS16	91	25	0	9	0	0	0	35	0	0	0	0	5	0	0	0	25
RS17	146	67	0	3	0	0	108	0	14	0	15	7	0	0	0	0	4
RS18	136	19	0	21	0	0	0	15	95	0	2	0	0	0	0	0	0
Total	1237	238	13	130	0	0	203	114	453	30	75	32	7	0	0	0	135

Table D.2: Number of annotated vocalization samples in each recording session

Session	Class							
code	*vo*	bv	fv	bc	co	sh	lg	bci
RS3	60	52	6	8	1	2	3	0
RS4	70	48	19	37	0	0	2	0
RS11	27	27	0	0	0	0	0	0
RS12	50	32	8	19	1	0	0	0
RS13	63	53	9	0	1	1	5	0
RS14	64	47	6	48	0	0	0	13
RS15	72	61	7	12	0	4	4	0
RS16	86	36	0	61	0	1	0	5
RS17	70	45	25	8	1	2	3	3
RS18	92	59	32	5	3	3	2	0
Total	654	460	112	198	7	13	19	21

Publications

- **G. Raboshchuk**, C. Nadeu, S. Vidiella Pinto, O. Ros Fornells, B. Muñoz Mahamud, and A. Riverola de Veciana, “Automatic detection of vocalization sounds in a neonatal intensive care unit environment using enhancement techniques,” *(submitted)*.

- **G. Raboshchuk**, P. Jančovič, C. Nadeu, A. Peiró Lilja, M. Köküer, B. Muñoz Mahamud, and A. Riverola de Veciana, “A knowledge-based approach to automatic detection of equipment alarm sounds in a neonatal intensive care unit environment,” *(submitted)*.

- B. Muñoz Mahamud, **G. Raboshchuk**, M. José Troyano Martos, M. Padró Hernández, and A. Riverola de Veciana, “Comparison of ambient sound levels in two level III neonatal intensive care units,” *Metas de Enfermería*, vol. 19, no. 3, pp. 57–63, Apr. 2016.

- **G. Raboshchuk**, C. Nadeu, S. Vidiella Pinto, O. Ros Fornells, B. Muñoz Mahamud, and A. Riverola de Veciana, “Audio enhancement for vocalization detection in a neonatal intensive care unit environment,” *Proceedings of International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, pp. 379–384, Apr. 2016.

- **G. Raboshchuk**, B. Muñoz Mahamud, C. Nadeu, and A. Riverola de Veciana, “Acoustic analysis of equipment alarms in a neonatal intensive care unit,” *Journal of Perinatal Medicine*, vol. 43, issue s1, Oct. 2015, *(abstract; presented at 12th World Congress of Perinatal Medicine)*.

- **G. Raboshchuk**, P. Jančovič, C. Nadeu, A. Peiró Lilja, M. Köküer, B. Muñoz Mahamud, and A. Riverola de Veciana, “Automatic detection of equipment alarms in a neonatal intensive care unit environment: a knowledge-based approach,” in *Proceedings of INTERSPEECH*, Sep. 2015, pp. 2902–2906, *(awarded ISCA travel grant)*.

- N. Torre, A. Riverola de Veciana, **G. Raboshchuk**, B. Muñoz Mahamud, C. Nadeu, and S. Navarro Hervas, “Acoustic environment study in a neonatal intensive care unit,” *Archives of Disease in Childhood*, vol. 99, suppl. 2, p. A465, Oct. 2014, *(abstract; presented at 5th Congress of the European Academy of Paediatric Societies)*.

- **G. Raboshchuk**, C. Nadeu, O. Ghahabi, S. Solvez, B. Muñoz Mahamud, A. Riverola de Veciana, and S. Navarro Hervas, "On the acoustic environment of a neonatal intensive care unit: initial description, and detection of equipment alarms," in *Proceedings of INTERSPEECH*, Sep. 2014, pp. 2543–2547.
- **G. Raboshchuk**, C. Nadeu, B. Muñoz Mahamud, A. Riverola de Veciana, and S. Navarro Hervas, "Acoustic study of a neonatal intensive care unit: preliminary results," in *Proceedings of International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, vol. 2, Apr. 2014, pp. 1418–1426.

Bibliography

[1] D. Wilson-Costello, H. Friedman, N. Minich, A. A. Fanaroff, and M. Hack, “Improved survival rates with increased neurodevelopmental disability for extremely low birth weight infants in the 1990s,” *Pediatrics*, vol. 115, no. 4, pp. 997–1003, 2005.

[2] K. A. Thomas and A. Uran, “How the NICU environment sounds to a preterm infant: update,” *The American journal of maternal child nursing*, vol. 32, no. 4, pp. 250–253, 2007.

[3] L. Gray and M. K. Philbin, “Effects of the neonatal intensive care unit on auditory attention and distraction,” *Clinics in perinatology*, vol. 31, no. 2, pp. 243–260, vi, 2004.

[4] A. Salavitabar, K. K. Haidet, C. S. Adkins, E. J. Susman, C. Palmer, and H. Storm, “Preterm infants’ sympathetic arousal and associated behavioral responses to sound stimuli in the neonatal intensive care unit,” *Advances in neonatal care*, vol. 10, no. 3, pp. 158–166, 2010.

[5] E. M. Wachman and A. Lahav, “The effects of noise on preterm infants in the NICU,” *Archives of Disease in Childhood - Fetal and Neonatal Edition*, vol. 96, no. 4, pp. F305–F309, 2010.

[6] Committee on Environmental Health, “Noise: A hazard for the fetus and newborn,” *Pediatrics*, vol. 100, no. 4, pp. 724–727, 1997.

[7] S. N. Graven and J. V. Browne, “Sleep and brain development,” *Newborn and Infant Nursing Reviews*, vol. 8, no. 4, pp. 173–179, 2008.

[8] S. M. A. Hassanein, N. M. El Raggal, and A. A. Shalaby, “Neonatal nursery noise: practice-based learning and improvement,” *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 26, no. 4, pp. 392–395, 2013.

[9] M. D. Livera, B. Priya, A. Ramesh, P. N. Suman Rao, V. Srilakshmi, M. Nagapoornima, A. G. Ramakrishnan, M. Dominic, and Swarnarekha, “Spectral analysis of noise in the neonatal intensive care unit,” *Indian journal of pediatrics*, vol. 75, no. 3, pp. 217–222, 2008.

[10] R. D. White, “Recommended standards for the newborn ICU design,” *Journal of Perinatology*, vol. 33, pp. S2–S16, 2013.

[11] C. Krueger, S. Wall, L. Parker, and R. Nealis, "Elevated sound levels within a busy NICU," *Neonatal network*, vol. 24, no. 6, pp. 33–37, 2005.

[12] M. K. Philbin, "Planning the acoustic environment of a neonatal intensive care unit," *Clinics in Perinatology*, vol. 31, no. 2, pp. 331–352, 2004.

[13] C. Jousselme, R. Vialet, E. Jouve, P. Lagier, C. Martin, and F. Michel, "Efficacy and mode of action of a noise-sensor light alarm to decrease noise in the pediatric intensive care unit: a prospective, randomized study," *Pediatric critical care medicine*, vol. 12, no. 2, pp. e69–72, 2011.

[14] M. Aita, C. Johnston, C. Goulet, T. F. Oberlander, and L. Snider, "Intervention minimizing preterm infants' exposure to NICU light and noise," *Clinical nursing research*, vol. 22, no. 3, pp. 337–358, 2013.

[15] R. Duran, N. A. Ciftdemir, U. V. Ozbek, U. Berberoğlu, F. Durankuş, N. Süt, and B. Acunaş, "The effects of noise reduction by earmuffs on the physiologic and behavioral responses in very low birth weight preterm infants," *International journal of pediatric otorhinolaryngology*, vol. 76, no. 10, pp. 1490–1493, 2012.

[16] F. Benini, V. Magnavita, P. Lago, E. Arslan, and P. Pisan, "Evaluation of noise in the neonatal intensive care unit," *American journal of perinatology*, vol. 13, no. 1, pp. 37–41, 1996.

[17] P. E. Marik, C. Fuller, A. Levitov, and E. Moll, "Neonatal incubators: a toxic sound environment for the preterm infant?" *Pediatric critical care medicine*, vol. 13, no. 6, pp. 685–689, 2012.

[18] A. W. Trickey, C. C. Arnold, A. Parmar, and R. E. Lasky, "Sound levels, staff perceptions, and patient outcomes during renovation near the neonatal intensive care unit," *Health Environments Research & Design Journal*, vol. 5, no. 4, pp. 76–87, 2012.

[19] P. Kuhn, C. Zores, T. Pebayle, A. Hoeft, C. Langlet, B. Escande, D. Astruc, and A. Dufour, "Infants born very preterm react to variations of the acoustic environment in their incubator from a minimum signal-to-noise ratio threshold of 5 to 10 dBA," *Pediatric research*, vol. 71, no. 4, pp. 386–392, 2012.

[20] A. Vorstermans, J. P. Martens, and B. Van Coile, "Automatic segmentation and labelling of multilingual speech data," *Speech Communication*, vol. 19, no. 4, pp. 271–293, 1996.

[21] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proceedings of ICASSP*, 2009, pp. 165–168.

[22] A. Freudenthal, M. v. Stuijvenberg, and J. B. v. Goudoever, "A quiet NICU for improved infants' health, development and well-being: a systems approach to reducing noise and auditory alarms," *Cognition, Technology & Work*, vol. 15, no. 3, pp. 329–345, 2013.

[23] C. van Pul, H. v. d. Mortel, J. v. d. Bogaart, T. Mohns, and P. Andriessen, “Safe patient monitoring is challenging but still feasible in a neonatal intensive care unit with single family rooms,” *Acta Paediatrica*, 2015.

[24] E. McMahon, P. Wintermark, and A. Lahav, “Auditory brain development in premature infants: the importance of early experience,” *Annals of the New York Academy of Sciences*, vol. 1252, pp. 17–24, 2012.

[25] M. Caskey, B. Stephens, R. Tucker, and B. Vohr, “Importance of parent talk on the development of preterm infant vocalizations,” *Pediatrics*, vol. 128, no. 5, pp. 910–916, 2011.

[26] A. Robertson, J. Kohn, P. Vos, and C. Cooper-Peel, “Establishing a noise measurement protocol for neonatal intensive care units,” *Journal of perinatology*, vol. 18, no. 2, pp. 126–130, 1998.

[27] R. E. Lasky and A. L. Williams, “Noise and light exposures for extremely low birth weight newborns during their stay in the neonatal intensive care unit,” *Pediatrics*, vol. 123, no. 2, pp. 540–546, 2009.

[28] A. E. Darcy, L. E. Hancock, and E. J. Ware, “A descriptive study of noise in the neonatal intensive care unit. ambient levels and perceptions of contributing factors,” *Advances in neonatal care*, vol. 8, no. 3, pp. 165–175, 2008.

[29] M. K. Philbin and P. Klaas, “Evaluating studies of the behavioral effects of sound on newborns,” *Journal of perinatology*, vol. 20, no. 8 pt. 2, pp. S61–67, 2000.

[30] J. Loewy, K. Stewart, A.-M. Dassler, A. Telsey, and P. Homel, “The effects of music therapy on vital signs, feeding, and sleep in premature infants,” *Pediatrics*, pp. 902–918, 2013.

[31] M. L. Mark Tramo, “Effects of music on physiological and behavioral indices of acute pain and stress in premature infants: clinical trial and literature review,” *Music and Medicine*, vol. 3, no. 2, pp. 72–83, 2011.

[32] J. Panagiotidis and A. Lahav, “Simulation of prenatal maternal sounds in NICU incubators: a pilot safety and feasibility study,” *The journal of maternal-fetal & neonatal medicine*, vol. 23, pp. 106–109, 2010.

[33] A. L. Williams, W. van Drongelen, and R. E. Lasky, “Noise in contemporary neonatal intensive care,” *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 2681–2690, 2007.

[34] M. Shahheidari and C. Homer, “Impact of the design of neonatal intensive care units on neonates, staff, and families: a systematic literature review,” *The Journal of perinatal & neonatal nursing*, vol. 26, no. 3, pp. 260–266, 2012.

[35] R. G. Pineda, J. Neil, D. Dierker, C. D. Smyser, M. Wallendorf, H. Kidokoro, L. C. Reynolds, S. Walker, C. Rogers, A. M. Mathur, D. C. Van Essen, and T. Inder, “Alterations in brain structure and neurodevelopmental outcome in preterm infants hospitalized in different neonatal intensive care unit environments,” *The Journal of pediatrics*, vol. 164, no. 1, pp. 52–60, 2014.

[36] C. Krueger, S. Schue, and L. Parker, “Neonatal intensive care unit sound levels before and after structural reconstruction,” *The American journal of maternal child nursing*, vol. 32, no. 6, pp. 358–362, 2007.

[37] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013.

[38] N. Vanderveer, “Ecological acoustics: human perception of environmental sounds,” Ph.D. dissertation, Cornell University, Ithaca, USA, 2011.

[39] D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: principles, algorithms and applications.* Wiley, 2006.

[40] A. Rabaoui, Z. Lachiri, and N. Ellouze, “Using HMM-based classifier adapted to background noises with improved sounds features for audio surveillance application,” *International Journal of Signal Processing*, vol. 5, no. 1, pp. 531–540, 2009.

[41] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, “Healthcare audio event classification using hidden markov models and hierarchical hidden markov models,” in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 1218–1221.

[42] M. Vacher, F. Portet, A. Fleury, and N. Noury, “Challenges in the processing of audio channels for ambient assisted living,” in *Proceedings of IEEE International Conference on e-Health Networking Applications and Services (Healthcom)*, 2010, pp. 330–337.

[43] S. Chu, S. Narayanan, and C.-C. Kuo, “Environmental sound recognition with time-frequency audio features,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[44] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, “A flexible framework for key audio effects detection and auditory context inference,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1026–1039, 2006.

[45] C. Cotton, D. Ellis, and A. Loui, “Soundtrack classification by transient events,” in *Proceedings of ICASSP*, 2011, pp. 473–476.

[46] D. P. Ellis, "Detecting alarm sounds," in *Proceedings of One-day Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis.* Department of Electrical Engineering, Columbia University, 2001, pp. 59–62.

[47] M. Vacher, D. Istrate, and J.-F. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees," in *Proceedings of EUSIPCO*, 2004, pp. 1171–1174.

[48] C. N. Doukas and I. Maglogiannis, "Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 277–289, 2011.

[49] M. Stäger, P. Lukowicz, N. Perera, T. von Büren, and T. Starner, "Soundbutton: Design of a low power wearable audio classification system," in *Proceedings of IEEE International Symposium on Wearable Computers*, 2003, pp. 12–17.

[50] J. Chen, J. Zhang, A. H. Kam, and L. Shue, "An automatic acoustic bathroom monitoring system," in *Proceedings of IEEE International Symposium on Circuits and Systems*, 2005, pp. 1750–1753.

[51] B. Lukic, "Activity detection in public places," Master's thesis, Royal Institute of Technology, Stockholm, Sweden, 2004.

[52] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *Proceedings of EUSIPCO*, 2013, pp. 1–5.

[53] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[54] M.-A. Carbonneau, N. Lezzoum, J. Voix, and G. Gagnon, "Detection of alarms and warning signals on an digital in-ear device," *International Journal of Industrial Ergonomics*, vol. 43, no. 6, pp. 503–511, 2013.

[55] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans.* Springer, 2007, pp. 311–322.

[56] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Springer, 2008, no. 4625, pp. 3–34.

[57] A. Misra, "Speech/nonspeech segmentation in web videos," in *Proceedings of INTERSPEECH*, 2012.

[58] T. Zhang and C.-C. Kuo, “Audio content analysis for online audiovisual data segmentation and classification,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[59] J. T. Geiger, B. Schuller, and G. Rigoll, “Large-scale audio feature extraction and SVM for acoustic scene classification,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[60] F. Beaufays, D. Boies, M. Weintraub, and Q. Zhu, “Using speech/non-speech detection to bias recognition search on noisy data,” in *Proceedings of ICASSP*, vol. 1, 2003, pp. I424–I427.

[61] M. Mimura, S. Sakai, and T. Kawahara, “Exploring deep neural networks and deep autoencoders in reverberant speech recognition,” in *Proceedings of Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014.

[62] J.-J. Aucouturier, B. Defreville, and F. Pachet, “The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music,” *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[63] D. Castan, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, “Segmentation-by-classification system based on factor analysis,” in *Proceedings of ICASSP*, 2013, pp. 783–787.

[64] C. Cotton and D. Ellis, “Spectral vs. spectro-temporal features for acoustic event detection,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 69–72.

[65] L. Vuegen, B. Van Den Broeck, P. Karsmakers, H. Van hamme, and B. Vanrumste, “Automatic monitoring of activities of daily living based on real-life acoustic sensor data: a preliminary study,” in *Proceedings of Workshop on speech and language processing for assistive technologies (SLPAT)*, 2013, pp. 113–118.

[66] C. Nadeu, D. Macho, and J. Hernando, “Time and frequency filtering of filter-bank energies for robust HMM speech recognition,” *Speech Communication*, vol. 34, no. 1–2, pp. 93–114, 2001.

[67] A. Temko, “Acoustic event detection and classification,” Ph.D. dissertation, Technical University of Catalonia, Barcelona, Spain, 2008.

[68] T. Butko, “Feature selection for multimodal acoustic event detection,” Ph.D. dissertation, Technical University of Catalonia, Barcelona, Spain, 2011.

[69] S. Ganapathy, S. Thomas, and H. Hermansky, “Comparison of modulation features for phoneme recognition,” in *Proceedings of ICASSP*, 2010, pp. 5038–5041.

[70] M. Athineos and D. P. Ellis, “Frequency-domain linear prediction for temporal features,” in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003, pp. 261—266.

[71] G. Roma, W. Nogueira, and P. Herrera, “Recurrence quantification analysis features for environmental sound recognition,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[72] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[73] M. Kleinschmidt, “Localized spectro-temporal features for automatic speech recognition,” in *Proceedings of Eurospeech*, 2003, pp. 2573–2576.

[74] R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik, “Sound retrieval and ranking using sparse auditory representations,” *Neural computation*, vol. 22, no. 9, pp. 2390–2416, 2010.

[75] X. V. González, “Perceptually-based signal features for environmental sound classification,” Ph.D. dissertation, Universitat Ramon LLul, Barcelona, Spain, 2011.

[76] K. Fukunaga, *Introduction to statistical pattern recognition.* Academic Press, 1972.

[77] D. P. W. Ellis and C. V. Cotton, “Subband autocorrelation features for video soundtrack classification,” in *Proceedings of ICASSP*, 2013, pp. 8663–8666.

[78] E. Wold, T. Blum, D. Keislar, and J. Wheaten, “Content-based classification, search, and retrieval of audio,” *IEEE MultiMedia*, vol. 3, no. 3, pp. 27–36, 1996.

[79] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, 2001.

[80] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” vol. 521, no. 7553, pp. 436–444, 2015.

[81] V. Nair and G. E. Hinton, “3D object recognition with deep belief nets,” in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1339–1347.

[82] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[83] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[84] A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[85] G. E. Hinton and S. Osindero, “A fast learning algorithm for deep belief nets,” *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[86] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” vol. 9, no. 8, pp. 1735–1780, 1997.

[87] O. Abdel-Hamid, L. Deng, and D. Yu, “Exploring convolutional neural network structures and optimization techniques for speech recognition,” in *Proceedings of INTERSPEECH*, 2013, pp. 3366–3370.

[88] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, “Feature extraction strategies in deep learning based acoustic event detection,” in *Proceedings of INTERSPEECH*, 2015, pp. 2922–2926.

[89] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *Proceedings of INTERSPEECH*, 2015.

[90] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: an overview,” in *Proceedings of ICASSP*, 2013, pp. 8599–8603.

[91] D. Palaz, R. Collobert, and others, “Analysis of CNN-based speech recognition system using raw speech as input,” in *Proceedings of INTERSPEECH*, 2015, pp. 11–15.

[92] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, and B. Ramabhadran, “Learning filter banks within a deep neural network framework,” in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 297–302.

[93] J. Zelinka, J. Vaněk, and L. Müller, “Simultaneously trained NN-based acoustic model and NN-based feature extractor,” in *Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science, P. Král and V. Matoušek, Eds. Springer International Publishing, 2015, no. 9302, pp. 234–242.

[94] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[95] T. Butko and C. Nadeu, “Audio segmentation of broadcast news: A hierarchical system with feature selection for the albayzin-2010 evaluation,” in *Proceedings of ICASSP*, 2011, pp. 357–360.

[96] C. Lopes and F. Perdigão, “Speech event detection by non negative matrix deconvolution,” in *Proceedings of EUSIPCO*, vol. 1, 2007, pp. 1280–1284.

[97] G. E. Hinton, “A Practical Guide to Training Restricted Boltzmann Machines,” in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Springer Berlin Heidelberg, 2012, pp. 599–619.

[98] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[99] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013.

[100] D. Johnson and D. Dudgeon, Eds., *Array Signal Processing.* Prentice Hall, 1993.

[101] R. Chakraborty, “Acoustic event detection and localization using distributed microphone arrays,” Ph.D. dissertation, Technical University of Catalonia, Barcelona, Spain, 2013.

[102] W. Wang, Ed., *Machine Audition: Principles, Algorithms and Systems*, 1st ed. IGI Global, 2010.

[103] M. Wolf, “Channel selection and reverberation-robust automatic speech recognition,” Ph.D. dissertation, Technical University of Catalonia, Barcelona, Spain, 2013.

[104] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley, 2001.

[105] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural Networks*, vol. 13, pp. 411–430, 2000.

[106] M. Berg, E. Bondesson, S. Y. Low, S. Nordholm, and I. Claesson, “A combined on-line PCA-ICA algorithm for blind source separation,” in *Proceedings of Asia-Pacific Conference on Communications*, 2005, pp. 969–972.

[107] N. Delfosse and P. Loubaton, “Adaptive blind separation of independent sources: a deflation approach,” *Signal Processing*, vol. 45, no. 1, pp. 59–83, 1995.

[108] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[109] J. T. Geiger, J. F. Gemmeke, B. Schuller, and G. Rigoll, “Investigating NMF speech enhancement for neural network based acoustic models,” in *Proceedings of INTERSPEECH*, 2014, pp. 2405–2409.

[110] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, “Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations,” in *Proceedings of ICASSP*, 2015, pp. 151–155.

[111] K. Kumar, R. Singh, B. Raj, and R. Stern, “Gammatone sub-band magnitude-domain dereverberation for ASR,” in *Proceedings of ICASSP*, 2011, pp. 4604–4607.

[112] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, “Discriminative NMF and its application to single-channel source separation,” in *Proceedings of INTERSPEECH*, 2014, pp. 865–869.

[113] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, “Sound event detection in multisource environments using source separation,” in *Proceedings of Workshop on Machine Listening in Multisource Environments*, 2011, pp. 36–40.

[114] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Efficient algorithms for multichannel extensions of itakura-saito nonnegative matrix factorization,” in *Proceedings of ICASSP*, 2012, pp. 261–264.

[115] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[116] P. O’Grady and B. Pearlmutter, “Convolutive non-negative matrix factorisation with a sparseness constraint,” in *Proceedings of IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 2006, pp. 427–432.

[117] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, “Supervised model training for overlapping sound events based on unsupervised source separation,” in *Proceedings of ICASSP*, 2013, pp. 8677–8681.

[118] J. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[119] B. Raj and P. Smaragdis, “Latent variable decomposition of spectrograms for single channel speaker separation,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 17–20.

[120] M. Shashanka, B. Raj, and P. Smaragdis, “Probabilistic latent variable models as nonnegative factorizations,” *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[121] P. Smaragdis, B. Raj, and M. Shashanka, “A probabilistic latent variable model for acoustic modeling,” in *Proceedings of NIPS Workshop on Advances in Models for Acoustic Processing*, vol. 148, 2006.

[122] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proceedings of ICASSP*, vol. 4, 1979, pp. 208–211.

[123] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “Elan: a professional framework for multimodality research,” in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006.

[124] D. Gerhard, “Audio signal classification: history and current techniques,” Dept. of Computer Science, University of Regina, Regina, Tech. Rep., 2003.

[125] M. Cowling, “Non-speech environmental sound classification system for autonomous surveillance,” Ph.D. dissertation, Griffith University, Nathan, Australia, 2004.

[126] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, Inc., 1993.

[127] J. ELliott, “Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics,” in *Proceedings of 8th Australian International Conference on Speech Science and Technology*, 2000, pp. 154–159.

[128] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, “Analysis and synthesis of shouted speech.” in *Proceedings of INTERSPEECH*, 2013, pp. 1544–1548.

[129] C. Zhang and J. H. Hansen, “Analysis and classification of speech mode: whispered through shouted,” in *Proceedings of INTERSPEECH*, 2007, pp. 2289–2292.

[130] J. Korpás, J. Sadlonová, and M. Vrabec, “Analysis of the cough sound: an overview,” *Pulmonary Pharmacology*, vol. 9, no. 5-6, pp. 261–268, 1996.

[131] T. Etz, H. Reetz, C. Wegener, and F. Bahlmann, “Infant cry reliability: Acoustic homogeneity of spontaneous cries and pain-induced cries,” vol. 58, pp. 91–100.

[132] J. Amoh and K. Odame, “Technologies for developing ambulatory cough monitoring devices,” *Critical Reviews in Biomedical Engineering*, vol. 41, no. 6, 2013.

[133] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, “The acoustic features of human laughter,” *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1581–1597, 2001.

[134] G. Raboshchuk, C. Nadeu, B. Muñoz Mahamud, A. Riverola de Veciana, and S. Navarro Hervas, “On the acoustic environment of a neonatal intensive care unit: initial description, and detection of equipment alarms,” in *Proceedings of INTERSPEECH*, 2014.

[135] F. Beritelli, S. Casale, A. Russo, and S. Serrano, “An Automatic Emergency Signal Recognition System for the Hearing Impaired,” in *Proceedings of the IEEE Digital Signal Processing Workshop*, 2006, pp. 179–182.

[136] M.-A. Carbonneau, N. Lezzoum, J. Voix, and G. Gagnon, “Detection of alarms and warning signals on an digital in-ear device,” *International Journal of Industrial Ergonomics*, vol. 43, no. 6, pp. 503–511, 2013.

[137] F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci, and P. Desii, “A real–time siren detector to improve safety of guide in traffic environment,” *Proceedings of EUSIPCO*, 2008.

[138] X. Xiao, H. Yao, and C. Guo, "Automatic Detection of Alarm Sounds in Cockpit Voice Recordings," in *IITA International Conference on Control, Automation and Systems Engineering (CASE)*, 2009, pp. 599–602.

[139] J. Schröder, S. Goetze, V. Grutzmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models." in *Proceedings of ICASSP*, 2013, pp. 493–497.

[140] "The NIST year 2006 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf, accessed: 2015-12-08.

[141] R. Joshi, C. v. Pul, L. Atallah, L. Feijs, S. V. Huffel, and P. Andriessen, "Pattern discovery in critical alarms originating from neonates under intensive care," *Physiological Measurement*, vol. 37, no. 4, p. 564.

[142] P. Jančovič and M. Köküer, "Detection of sinusoidal signals in noise by probabilistic modelling of the spectral magnitude shape and phase continuity," in *Proceedings of ICASSP*, 2011, pp. 517–520.

[143] F. Keiler and S. Marchand, "Survey on extraction of sinusoids in stationary sounds," in *Proceedings of International Conference on Digital Audio Effects (DAFX)*, 2002, pp. 1–8.

[144] P. Jančovič, M. Köküer, M. Zakeri, and M. Russell, "Unsupervised discovery of acoustic patterns in bird vocalisations employing DTW and clustering," in *Proceedings of EUSIPCO*, 2013.

[145] P. Jančovič, M. Zakeri, M. Köküer, and M. Russell, "HMM-based modelling of individual syllables for bird species recognition from audio field recordings," in *Proceedings of ICASSP*, 2015, pp. 768–772.

[146] P. Jančovič and M. Köküer, "Incorporating the voicing information into hmm-based automatic speech recognition in noisy environments," *Speech Communication*, vol. 51, no. 14, pp. 438–451, 2009.

[147] B. Schuller and F. Weninger, "Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization," in *Proceedings of ICASSP*, 2010, pp. 5054–5057.

[148] T. N. Sainath, B. Kingsbury, A.-r. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proceedings of Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 315–320.

[149] G. Raboshchuk, C. Nadeu, B. Muñoz Mahamud, A. Riverola de Veciana, and S. Navarro Hervas, "Acoustic study of a neonatal intensive care unit: Preliminary results," in *Proceedings of International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, 2014, pp. 1418–1426.

[150] S. Telkemeyer, S. Rossi, S. P. Koch, T. Nierhaus, J. Steinbrink, D. Poeppel, H. Obrig, and I. Wartenburger, "Sensitivity of newborn auditory cortex to the temporal structure of sounds," *Journal of Neuroscience*, vol. 29, no. 47, pp. 14 726–14 733, 2009.

[151] K. Kotilahti, I. Nissilä, T. Näsi, L. Lipiäinen, T. Noponen, P. Meriläinen, M. Huotilainen, and V. Fellman, "Hemodynamic responses to speech and music in newborn infants," *Human Brain Mapping*, vol. 31, no. 4, pp. 595–603, 2010.

[152] M. Trapanotto, F. Benini, M. Farina, D. Gobber, V. Magnavita, and F. Zacchello, "Behavioural and physiological reactivity to noise in the newborn," *Journal of Paediatrics and Child Health*, vol. 40, no. 5–6, pp. 275–281, 2004.

[153] A. L. Williams, M. Sanderson, D. Lai, B. J. Selwyn, and R. E. Lasky, "Intensive care noise and mean arterial blood pressure in extremely low-birth-weight neonates," *American Journal of Perinatology*, vol. 26, no. 5, pp. 323–329, 2009.

[154] F. Güiza, J. V. Eyck, and G. Meyfroidt, "Predictive data mining on monitoring data from the intensive care unit," *Journal of Clinical Monitoring and Computing*, vol. 27, no. 4, pp. 449–453, 2012.

[155] A. Temko, D. Macho, and C. Nadeu, "Fuzzy integral based information fusion for classification of highly confusable non-speech sounds," *Pattern Recognition*, vol. 41, no. 5, pp. 1831–1840, 2008.

[156] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of INTERSPEECH*, 2015.

[157] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," vol. 15, no. 1, pp. 1929–1958.

[158] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[159] C. Joder, F. Weninger, D. Virette, and B. Schuller, "Integrating noise estimation and factorization-based speech separation: a novel hybrid approach," in *Processing of ICASSP*, 2013, pp. 131–135.

[160] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–15, 2015.