

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies
de l'Information et de la Communication*
Spécialité : *Signal, Image, Vision*

Par

Sandie Cabon

Monitoring of premature newborns by video and audio analyses

Thèse présentée et soutenue à Rennes, le 12 juillet 2019

Unité de recherche : LTSI, UMR Inserm 1099 Laboratoire Traitement du Signal et de l'Image

Thèse N° :

Composition du Jury :

Norbert Noury

Professeur, Institut National des Sciences Appliquées de Lyon, Rapporteur

Catherine Achard

Maître de conférences (HdR), Sorbonne Université, Rapporteur

Ludovic Macaire

Professeur, Université de Lille, Président du jury

Fabienne Porée

Maitre de Conférences (HdR), LTSI-INSERM, Université de Rennes I, Examineur

Olivier Rosec

Directeur R&D, Voxygen, Lannion, Co-directeur de thèse

Guy Carrault

Professeur, LTSI-INSERM, Université de Rennes I, Directeur de thèse

Antoine Simon

Maitre de Conférences (HdR), LTSI-INSERM, Université de Rennes I, Membre invité

Patrick Pladys

Professeur-Praticien Hospitalier, LTSI-INSERM, Université de Rennes I, Membre invité

RÉSUMÉ EN FRANÇAIS

Cette thèse s'est déroulée dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE) et a commencé en février 2016. Ce manuscrit est donc le résultat d'un travail collaboratif entre l'entreprise Voxygen et le Laboratoire Traitement du Signal et de L'Image (LTSI). La plupart des travaux ont été réalisés en parallèle du projet européen Digi-NewB, démarré en mars 2016. Digi-NewB a pour but de proposer de nouvelles solutions de monitoring du nouveau-né prématuré à partir de trois types de données (électrophysiologiques, cliniques et vidéo&audio), afin d'aider les cliniciens dans leur prise de décision. Deux aspects principaux de la santé néonatale sont concernés : l'infection et la maturation neurocomportementale. Dans ce contexte, l'attention de cette thèse s'est concentrée sur le développement de techniques non invasives, en particulier l'analyse vidéo et audio, pour le suivi des nouveau-nés prématurés dans leur évolution neurocomportementale.

Motivations

Dans le monde, 15 millions de bébés naissent prématurément chaque année et ce nombre augmente dans presque tous les pays dont les données sont accessibles [2, 6]. En France, cela représente 6% des naissances, soit environ 60 000 naissances par an. La prématurité est la principale cause de mortalité néonatale et des solutions pour améliorer la prise en charge peuvent encore être développées.

Les prématurés ont plusieurs fonctions immatures notamment digestives, cardio-respiratoires, immunologiques ou neurologiques et reçoivent donc des soins spécialisés. Leur développement est optimisé grâce à une forte surveillance assurée par le personnel médical des unités de soins intensifs néonatales.

Les nouveau-nés prématurés sont équipés de plusieurs dispositifs médicaux en fonction de la gravité de leur immaturité. Néanmoins, bien que vitaux, certains de ces équipements sont invasifs, comme l'intubation pour l'assistance respiratoire, la perfusion intraveineuse et le cathéter pour le soutien alimentaire. De plus, environ deux fois par jour, les nouveau-nés prématurés subissent des tests sanguins. Ils sont notamment réalisés pour vérifier le taux d'oxygénation du sang ou pour détecter des infections. Malheureusement, toutes ces procédures invasives peuvent également dégrader l'état du nouveau-né en provoquant des infections nosocomiales [5].

En même temps, leur état de santé est surveillé en permanence, en particulier les activités cardiaque et respiratoire. En effet, des techniques informatiques ont été déployées pour déclencher des alarmes en cas d'événements de détresse comme les apnées ou les bradycardies. Pour cela, des électrodes sont placées sur le corps du nouveau-né et des signaux électrophysiologiques sont constamment acquis.

De manière plus ponctuelle, leur maturation neurocomportementale est également évaluée. Cela s'effectue principalement par l'évaluation du sommeil. En effet, le sommeil est un indicateur direct du développement neurocomportemental puisqu'il a été démontré que sa structure évolue avec l'âge du nouveau-né [4]. Une première façon de réaliser ce suivi est d'étudier l'électroencéphalogramme du

nouveau-né prématuré. Néanmoins, cela nécessite de déployer un système ambulatoire et de placer des électrodes sur la tête du bébé. C'est pourquoi, dans la pratique, cette technique est peu utilisée et sert principalement à confirmer des soupçons de pathologies neurologiques. Comme alternative, des techniques observationnelles peuvent être utilisées. Aujourd'hui, ces observations sont effectuées en présence du nouveau-né par des infirmières qualifiées dans le cadre du Newborn Individualized Developmental Care and Assessment Program (NIDCAP). L'idée sous-jacente de ce programme est que le comportement du nouveau-né est un indicateur de ses besoins et peut donc être observé afin d'individualiser les soins [1]. Au cours de ces observations, plusieurs composantes sont observées telles que les stades de sommeil, les activités vocales, motrices ou faciales. Ces éléments se sont également avérés pertinents pour la détection de divers troubles neurologiques [3, 7, 8]. Toutefois, plusieurs limites entravent la généralisation de ces observations. En effet, cette opération prend beaucoup de temps et seule une petite partie des nouveau-nés peut alors bénéficier de ce suivi. De plus, bien qu'elle soit effectuée par des infirmières spécialement formées, ces observations restent subjectives.

Objectifs

A la lumière de ces observations, il semble que de nouvelles solutions pour assurer une surveillance neurocomportementale continue pourraient améliorer la prise en charge des nouveau-nés. En effet, avec une telle solution, le comportement des nouveau-nés pourraient être surveillé en continu et les cliniciens pourraient alors disposer d'éléments supplémentaires pour évaluer leur développement neurocomportemental. Cependant, les nouveau-nés étant déjà encombrés par plusieurs équipements, il est important de développer des stratégies non invasives.

Contributions

Dans ce but, l'attention de cette thèse s'est concentrée sur des traitements vidéo et audio.

Premièrement, la pertinence de telles analyses dans le contexte de la santé pédiatrique a été étudiée. Ainsi, plus de 150 documents ont été examinés et les méthodes proposées ont été évaluées au regard de la possibilité de les intégrer à un système de surveillance continue. Ce travail conséquent a été publié dans la revue *Physiological Measurement* :

- Cabon, S., Porée, F., Simon, A., Rosec, O., Pladys, P., Carrault, G. Video and audio processing in paediatrics : a review. *Physiological Measurement*, 40(2), 1-20. (2019).

Simultanément, une étude préliminaire sur l'estimation des états de sommeil chez le nouveau-né prématuré a été menée. Pour la première fois, une approche semi-automatique combinant le traitement vidéo et audio a été développée. Ce travail a été publié dans la revue *Biomedical Signal Processing and Control* :

- Cabon, S., Porée, F., Simon, A., Met-Montot B., Pladys, P., Rosec, O., Nardi, N. and Carrault, G. Audio- and Video-based estimation of the sleep stages of newborns in Neonatal Intensive Care Unit. *Biomedical Signal Processing and Control*, 52, 362-370. (2019).

A partir de là, plusieurs décisions ont été prises concernant l'orientation du travail et les développements prioritaires pour se rapprocher d'une solution de surveillance entièrement automatique.

Dans un premier temps, un nouveau système d'acquisition audio-vidéo a été proposé. Son intégration dans les unités de soins intensifs néonatales a été étudiée afin de s'adapter à une grande variété de configurations de chambres, en gardant à l'esprit l'objectif non invasif, tant pour les nouveau-nés que pour le personnel médical. Il importe de souligner que cet aspect important de la surveillance audio et vidéo a été très peu abordé dans la littérature et que ce travail constitue une contribution importante.

Deuxièmement, l'attention s'est portée sur les analyses de mouvement. Un processus allant de l'acquisition vidéo à l'extraction de caractéristiques pertinentes pour la caractérisation neurocomportementale a été proposé. Il est basé sur l'extraction de séries de mouvements au moyen de techniques de traitement vidéo classiques (e.g., différence inter-images, opérateurs morphologiques). La force de ce travail repose sur le fait que plusieurs difficultés, inhérentes au suivi sur de longues périodes, ont été surmontées. La plus importante est la détection automatique des périodes d'analyses non pertinentes, en raison de la présence d'un adulte dans le champ de la caméra. Cette partie spécifique du processus a été présentée à la conférence Recherche en Imagerie et Technologies pour la Santé et publiée dans la revue *Innovation and Research in BioMedical engineering* :

- Cabon S., Porée F., Simon A., Ugolin M., Rosec O., Carrault G. and Pladys P. Caractérisation du mouvement chez les nouveau-nés prématurés par analyse automatique de vidéos. Journées RITS 2017. Lyon, France, résumé. (2017).
- Cabon, S., Porée, F., Simon, A., Ugolin, M., Rosec, O., Carrault, G. and Pladys, P. Motion Estimation and Characterization in Premature Newborns Using Long Duration Video Recordings. *IRBM*, 38(4), 207-213. (2017) ;

De plus, un ensemble de paramètres caractérisant l'organisation du mouvement a été calculé après une étape de classification. Les résultats préliminaires révèlent que l'ensemble de paramètres proposé est pertinent pour apprécier l'évolution de l'organisation du mouvement des nouveau-nés prématurés.

Enfin, un processus allant de l'enregistrement audio à l'extraction automatique des pleurs a été proposé. De la même façon que pour l'analyse du mouvement, la force de ces travaux réside dans le fait qu'il a été conçu pour s'adapter aux conditions réelles des unités de soins intensifs néonatales. En effet, dans ces unités, divers sons non pertinents, dans le cadre de notre étude, peuvent être enregistrés, comme des voix d'adultes, des alarmes de dispositifs ou des bruits de fond. Notre approche pour l'extraction des pleurs est basée sur un ensemble de caractéristiques calculées à partir d'une analyse "harmonique plus bruit". Cette technique, généralement appliquée en synthèse vocale, n'a jamais été abordée dans ce but et les résultats s'avèrent équivalents, sinon supérieurs, à ceux observés dans la littérature. Ils montrent qu'en dépit d'un contexte clinique difficile une segmentation et une classification automatique des pleurs est possible et que la fréquence fondamentale des pleurs est un très bon marqueur de la maturation du bébé prématuré.

Enfin, ce travail a été réalisé dans le contexte du projet européen Digi-NewB où une base de données conséquente a pu être acquise, il a également servi de support à l'annotation manuelle de vidéos et de signaux sonores pour l'évaluation des méthodes automatiques qui ont été proposées. Ceci représente donc également à nos yeux une contribution importante qu'il convient de mentionner.

Contenu du manuscrit

Le manuscrit qui résulte de ces travaux est divisé en sept chapitres. Le chapitre 1 est consacré à la définition des objectifs cliniques de la thèse. Une attention particulière est accordée aux soins des nouveau-nés prématurés et à la présentation des unités de soins intensifs néonatales. L'importance de développer de nouvelles solutions de suivi, notamment pour l'évaluation du développement comportemental, y est soulignée.

Dans les deux chapitres suivants, le contexte méthodologique de ce travail est présenté. Dans le chapitre 2, nous proposons un examen approfondi de la littérature. Les méthodes audio et vidéo existantes développées pour la santé pédiatrique sont présentées et leur pertinence dans le cadre d'un objectif de surveillance est discutée. L'analyse de ce chapitre a démontré l'importance des méthodes d'exploration et de classification des données pour aller vers des solutions entièrement automatiques. Ainsi, le chapitre 3 se concentre sur la présentation des méthodes utiles à la classification. De plus, une grande partie des résultats présentés dans cette thèse proviennent directement de l'utilisation de ces techniques dans trois contextes différents : l'analyse du sommeil, la détection de mouvement et l'extraction de pleurs.

Une étude, reportée dans le chapitre 4, a été menée afin d'évaluer la pertinence des méthodes audio et vidéo pour l'identification des stades du sommeil chez les nouveau-nés prématurés. Ces travaux ont servi de base pour les chapitres suivants.

Un nouveau système d'acquisition audio-vidéo, développé dans le cadre de Digi-NewB, a ainsi été proposé. Dix-huit systèmes ont été déployés dans six hôpitaux. Leur intégration dans les unités de soins intensifs néonatales a été étudiée. La présentation et l'évaluation de ces travaux sont présentées dans le chapitre 5.

Le chapitre 6 présente les traitements vidéo mis au point pour caractériser l'organisation du mouvement des nouveau-nés prématurés. Ce processus a été conçu pour répondre en grande partie aux difficultés induites par la surveillance à long terme telles que les diverses configurations d'environnement ou la présence d'adultes dans le champ de la caméra. Les séries de mouvements ont été extraites de la vidéo et des paramètres caractérisant l'organisation du mouvement en termes de durée et de nombre d'intervalles de mouvement et de non-mouvement ont été estimés. Les résultats reportés démontrent tout le bien fondé de l'importance de quantifier le mouvement dans le suivi de la maturation du nouveau-né prématuré.

Avec ces mêmes objectifs, un processus audio, basé sur la classification, a été conçu afin de caractériser les pleurs des nouveau-nés prématurés et est décrit au chapitre 7. Dans ce contexte, nous nous sommes intéressés à l'extraction automatique des pleurs émis par le bébé en excluant les autres sons (e.g., voix d'adultes, alarmes) grâce à des méthodes de classification. De la même manière que pour le mouvement, des résultats préliminaires ont été obtenus concernant l'évolution des pleurs, notamment en ce qui concerne la fréquence fondamentale, chez les prématurés.

Bibliographie

[1] ALS, H., LAWHON, G., DUFFY, F. H., McANULTY, G. B., GIBES-GROSSMAN, R., AND BLICKMAN,

- J. G. Individualized developmental care for the very low-birth-weight preterm infant : medical and neurofunctional effects. *Jama* 272, 11 (1994), 853–858.
- [2] BLENCOWE, H., COUSENS, S., OESTERGAARD, M. Z., CHOU, D., MOLLER, A.-B., NARWAL, R., ADLER, A., GARCIA, C. V., ROHDE, S., SAY, L., ET AL. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries : a systematic analysis and implications. *The Lancet* 379, 9832 (2012), 2162–2172.
- [3] BOS, A. F., MARTIJN, A., OKKEN, A., AND PRECHTL, H. F. R. Quality of general movements in preterm infants with transient periventricular echodensities. *Acta Paediatrica* 87, 3 (1998), 328–335.
- [4] CURZI-DASCALOVA, L. Développement du sommeil et des fonctions sous contrôle du système nerveux autonome chez le nouveau-né prématuré et à terme. *Archives de pédiatrie* 2, 3 (1995), 255–262.
- [5] MCGUIRE, W., CLERIHEW, L., AND FOWLIE, P. W. Infection in the preterm infant. *British Medical Journal* 329, 7477 (2004), 1277–1280.
- [6] WORLD HEALTH ORGANIZATION AND OTHERS. Born too soon : the global action report on preterm birth. (2012)
- [7] PRECHTL, H. F. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Human Development* 23, 3 (1990), 151–8.
- [8] PRECHTL, H. F., EINSPIELER, C., CIONI, G., BOS, A. F., FERRARI, F., AND SONTHEIMER, D. An early marker for neurological deficits after perinatal brain lesions. *Lancet* 349, 9062 (1997), 1361–3.

TABLE OF CONTENTS

List of acronyms	11
Remerciements	13
Introduction	15
1 Clinical context and objectives	19
1 Prematurity	19
1.1 Definitions	19
1.2 Risks due to prematurity	20
1.3 Presentation of Neonatal Intensive Care Units	21
1.4 Sleep assessment	23
2 Objectives of the work	25
2.1 The thesis	25
2.2 The Digi-NewB project	25
3 Conclusion	26
Bibliography	27
2 Methodological context: video and audio processing in paediatrics	31
1 Background	31
2 Video analysis	32
2.1 Clinical applications	33
2.2 Methods for video processing	34
3 Audio analysis	40
3.1 Clinical applications	40
3.2 Methods for acoustic signal processing	42
4 Discussion and Conclusion	49
Bibliography	50
3 Methodological context: methods for classification	65
1 Problem formulation	65
2 Dimensionality reduction	67
2.1 Feature selection	69
2.2 Feature extraction	70
3 Machine learning algorithms	72
3.1 Linear algorithms	73

3.2	Non-linear algorithms	75
4	Techniques for performance evaluation	80
4.1	Metrics	80
4.2	Cross-validation	82
5	Conclusion	83
	Bibliography	84
4	Preliminary behavioral sleep states estimation	87
1	Introduction	87
2	Methods	89
2.1	Database	89
2.2	Vocalizations' extraction	90
2.3	Motion estimation	90
2.4	Eye state estimation	91
2.5	Sleep stage estimation	94
3	Results	94
3.1	Software and platforms	95
3.2	Tuning of the parameters	95
3.3	Performances of the eye state estimation method	96
3.4	Results of sleep stage classification from extracted descriptors	97
4	Discussion	99
5	Conclusion	100
	Bibliography	100
5	Voxyvi, a new audio-video acquisition system for Neonatal Intensive Care Unit	105
1	Introduction	105
2	Materials and methods	106
2.1	Description of the acquisition system	106
2.2	Integration in Neonatal Intensive Care Unit	110
2.3	Data acquisition protocol	113
3	Evaluation of the acquisition system integration in NICU	114
3.1	Inclusion progress	114
3.2	Clinicians feedback	116
3.3	Difficulties	119
4	Discussion and conclusion	120
	Bibliography	121
6	Video-based characterization of newborn motion for neuro-developmental monitoring	123
1	Methods to characterize motion organization in newborns	123
1.1	Estimation of the amount of motion	124
1.2	Adult detection	125
1.3	Motion segmentation	127

1.4	Feature extraction	131
2	Evaluation of the process	131
2.1	Evaluation of the adult detection method	131
2.2	Evaluation of the motion segmentation method	134
2.3	Evaluation of the accuracy of motion features	139
3	Evolution of the motion organization in preterm newborns	140
3.1	Maturation dataset	140
3.2	Comparison between $D1$ and $D2$ in $G1$	141
3.3	Comparison between $D1$ and $D2$ in $G2$	143
3.4	Comparison of motion activity near discharge	143
3.5	Control map for assessing motion organization development in preterm infants. . .	144
4	Discussion and conclusion	145
	Bibliography	146
7	Audio-based characterization of newborn cries for neuro-developmental monitoring	149
1	Methods to extract newborn cries	149
1.1	Annotated database	150
1.2	Feature engineering	151
1.3	Classifiers	156
2	Evaluation of the process	156
2.1	Evaluation strategy	156
2.2	Dimensionality reduction	156
2.3	Tuning of the parameters	158
2.4	Classification results	159
2.5	Evaluation of the model accuracy for fundamental frequency analyses	160
3	Evolution of the fundamental frequency in very preterm newborns	166
4	Discussion and conclusion	167
	Bibliography	168
	Conclusion and perspectives	171
	List of publications	174
	Appendices	175
A	User Manual	176

LIST OF ACRONYMS

AA	Active Alert
AAM	Active Appearance Models
ANN	Artificial Neural Networks
APSS	Association of the Psychophysiological Study of Sleep
AR	Auto Regressive
AS	Active Sleep
ASA/ASTA	Australasian Sleep Association/Australasian Sleep Technologists' Association
AU	Action Unit
AUC	Area Under Curve
B&W	Black and White
CCHS	Congenital Central Hypoventilation Syndrome
CNN	Convolution Neural Networks
CNS	Central Nervous System
CP	Cerebral Palsy
CU	Cry Unit
CWT	Continuous Wavelet Transform
D	Drowsiness
DAN	Douleur Aigüe du Nouveau-né
DCT	Discrete Cosinus Transform
DSS	Decision Support System
EEG	ElectroEncephaloGraphic
EMD	Empirical Mode Decomposition
FFNN	Feed Forward Neural Networks
FFT	Fast Fourier Transform
FLN	Five-Line Method
FP	False Positive
FN	False Negative
GA	Gestational Age
GM	General Movements
GMA	General Movements Assessment
GMM	Gaussian Mixture Model
HINE	Hammersmith Infant Neurological Examinations
HMM	Hidden Markov Model
HNH	Harmonic plus Noise Model
HOG	Histogram of Oriented Gradients
IUGR	IntraUterine Growth Retardation

KLT	Kanade-Lucas-Tomasi
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LOOCV	Leave-one-out cross-validation
LPCCs	Linear Prediction Cepstral Coefficients
LR	Logistic Regression
LTAS	Long Time Average Spectrum
MFCCs	Mel Frequency Cepstral Coefficients
MLE	Maximum Likelihood Estimation
MLP	Multi-Layer Perceptron
NBAS	Neonatal Behavioral Assessment Scale
NFCS	Neonatal Facial Coding System
NICU	Neonatal Intensive Care Units
NIDCAP	Newborn Individualized Developmental Care and Assessment Program
NIR	Near-InfraRed
NLEO	Non Linear Energy Operator
NUC	Next Unit of Computing
OSA	Obstructive Sleep Apnea
PCA	Principal Component Analysis
PMA	PostMenstrual Age
PPV	Positive Predictive Value
PRM	Premature Rupture of Membranes
QA	Quiet Alert
QS	Quiet Sleep
RF	Random Forest
ROC	Receiver Operating Characteristics
ROI	Region Of Interest
SIDS	Sudden Infant Death Syndrome
SIFT	Simple Inverse Filter Tracking
SVM	Support Vector Machine
SSM	Smoothed Spectrum Method
STE	Short Time Energy
TP	True Positive
TN	True Negative
V/UV	Voiced/UnVoiced
WP	Work Package
ZCR	Zero Crossing Rate

REMERCIEMENTS

Je tiens tout d'abord à remercier Ludovic Macaire, président du jury et Professeur à l'Université de Lille, Catherine Achard, Maître de Conférences à l'Université Pierre et Marie Curie ainsi que Norbert Noury, Professeur à l'Institut National des Sciences Appliquées de Lyon, d'avoir apporté leurs expertises à ces travaux.

Je remercie mes directeurs de thèse, Guy Carrault, Professeur à l'Université de Rennes 1 et Olivier Rosec, Directeur R&D de Voxygen pour leurs conseils et leur soutien. Je les remercie de m'avoir permis de travailler avec une très grande liberté.

Je souhaite particulièrement remercier Patrick Pladys qui porte, avec Guy Carrault, le beau projet Digi-NewB.

Je tiens aussi à souligner l'entente et les échanges exceptionnels que nous avons eus avec mes deux encadrants, Fabienne Porée et Antoine Simon. Merci pour tout ce que vous m'avez apporté et m'apporterez encore.

Je remercie toute l'équipe de Digi-NewB pour son implication dans le projet. Merci à Florence pour sa disponibilité, nos échanges et nos tournées des centres hospitaliers et aussi à mon « binôme » de Voxygen, Guillaume. Je souhaite aussi remercier Bertille et Raphaël pour leurs travaux sur les analyses audio et vidéo.

Je remercie le Laboratoire de Traitement du Signal et de l'Image et particulièrement son directeur, Lotfi Senhadji pour l'accueil qui m'a été réservé depuis février 2015.

Je remercie les collègues pour l'ambiance et pour les discussions que nous avons pu avoir et plus particulièrement celles avec mes trois amis : Karim, Pablo et Matthieu. Mention spéciale aux locataires du bureau 404.

Pendant ces trois ans, j'ai aussi pu compter sur le soutien sans faille de mes amis d'enfance et colocataires de la Tucherie (CT, Gaby, Chris et Ronan). Je les remercie d'avoir pris soin de moi et de m'avoir nourrie lors des 6 derniers mois !

Merci à ma famille et ma belle-famille. Ma mère, à qui je dois tout. Sylvain et Yoan, mes frères, je suis très fière de vous. Je remercie aussi mon papy et ma mamie qui m'ont toujours soutenue. Je remercie ma belle-mère, Fabienne, qui m'a toujours accompagnée. Mes super belles sœurs : Nina, Marie, Barbara et Elisabeth, merci pour votre soutien.

Bien entendu, je remercie tout-e-s mes ami-e-s de collège et de lycée (oh les gars, le garage), de l'École Nationale d'Ingénieurs de Brest (BDE SWAG) et ceux que j'ai rencontrés en dehors de ma scolarité.

Merci à mon amour, mon partenaire, mon coéquipier, Ronan, pour ta présence essentielle à mon équilibre.

INTRODUCTION

Worldwide, 15 million of babies are born prematurely each year and this number is rising in almost all countries with reliable data [2, 8]. In France, it represents 6% of the total births which means, approximately, 60 000 births every year. Prematurity is the main cause of neonatal mortality and solutions to enhance their care still need to be developed.

Premature babies have several immature functions such as digestive, cardio-respiratory, immunological or neurological and thus receive a specialized care. Their optimal development is ensured thanks to a high medical supervision that is provided by medical staff in Neonatal Intensive Care Units (NICU).

Premature newborns are equipped with several medical devices depending on the severity of their immaturity. The better their condition evolves, the less equipment will be used until their discharge home. Nevertheless, most of this equipment, although vital, is invasive such as intubation for respiratory assistance, intravenous infusion and catheter for food support. Moreover, about twice a day, premature newborns experience blood testing, performed in order to check blood oxygenation or to detect infections. Unfortunately, all of these invasive procedures can also degrade the newborn condition by provoking nosocomial infections [5].

In the meantime, their health status is continuously monitored, especially regarding cardiac and respiratory activities. In fact, computational techniques have been deployed in order to trigger alarms in case of distress events (e.g., apneas, bradycardias). For that purpose, electrodes are placed on the newborn's body and electrophysiological signals are constantly acquired.

In a more punctual manner, their neuro-behavioral maturation is also assessed. It is mainly performed through sleep evaluation. In fact, the sleep mechanism is a direct indicator of neuro-behavioral development since it has been shown to evolve with the age of the newborn [4]. For this purpose, ElectroEncephaloGrams (EEG) can be acquired and studied. Nevertheless, it requires to deploy an ambulatory system and to place electrodes on the baby's head. Hence, in practice, it is sparsely performed and mainly in order to confirm medical suspicion of neurological pathologies. As an alternative, observational techniques can be used. To date, these observations are performed in presence of the newborn by trained nurses as part of the Newborn Individualized Developmental Care and Assessment Program (NIDCAP). The underlying idea of this program is that the behavior of the newborn is an indicator about his/her needs and thus can be observed to individualize the care [1]. During these observations, several components are observed such as respiratory, vocal, eyes, motor or facial activities. These elements have been shown to be related to sleep states but also to be relevant for the detection of various neurological impairments [3, 6, 7]. However, several limitations hinder the generalization of these observations. This operation is very time-consuming and thus only a small part of the newborns benefits this follow-up. Additionally, although it is performed by trained nurses, these observations remain subjective.

On the light of these observations, it appears that new solutions to provide continuous neuro-behavioral monitoring could enhance the care of newborns. Indeed, with such solutions, all newborns may be con-

stantly monitored and clinicians may have additional elements to assess neuro-behavioral maturation. However, newborns being already cluttered by several devices, it is important to develop non-invasive strategies.

Among the non-invasive techniques, the use of cameras associated with microphones seems to be one of the most relevant to provide a behavioral characterization close to the observations made by nurses. Indeed, that way, vocal, motion or facial activities can be captured. In addition, their set-up requires no interaction and no contact with the newborn. For these reasons, this thesis was focused on the development of video and audio analyses for behavioral monitoring.

This thesis was conducted in the context of a Convention Industrielle de Formation par la Recherche (CIFRE) and began in February 2016. Hence, this work is the result of a collaboration between Voxygen company and the Laboratoire Traitement du Signal et de l'Image (LTSI). Most of this work was performed in the context of the European Project Digi-NewB, started in March 2016. Digi-NewB aims to propose new monitoring systems, based on three data sources (electrophysiological, clinical and audio&video data) to help clinicians in their diagnosis. Two main aspects of neonatal health are targeted: sepsis and neuro-behavioral maturation. It is on this last point that the work of this thesis is focused.

The resulting manuscript is divided into seven chapters. Chapter 1 is dedicated to the definition of the clinical objectives of this thesis. A particular attention is paid to the current care of premature newborns and the presentation of Neonatal Intensive Care Units. The importance of developing new monitoring solutions, notably for the assessment of behavioral development, is underlined.

The methodological context of this work is considered in the two following chapters. In Chapter 2, a review of the literature is proposed. Existing audio and video methods developed for pediatric health are presented and their relevance within a monitoring objective is discussed. Then, Chapter 3 focuses on methods for classification since a large part of the results presented in this thesis comes directly from the use of such techniques.

In Chapter 4, a study is presented in order to evaluate the relevance of audio and video methods for sleep stages identification in preterm newborns. The proposed approach combines audio and video analyses.

Then, a new audio-video acquisition system, developed in the context of Digi-NewB, is proposed in Chapter 5. In particular, its integration in Neonatal Intensive Care Units is evaluated as well as its acceptability by the medical staff.

In Chapter 6, a video-based processing developed in order to characterize the motion organization of preterm newborns is presented. This process is designed in order to answer a large part of the difficulties induced by long time monitoring (e.g., various configurations of the environment, presence of adults in the camera field).

Within the same objectives, an audio-based process based on classification, is proposed in order to automatically extract cry events from all the other sounds encountered in NICU (e.g., alarms, adult voices). This work is reported in Chapter 7.

Bibliography

[1] ALS, H., LAWHON, G., DUFFY, F. H., McANULTY, G. B., GIBES-GROSSMAN, R., AND BLICKMAN,

- J. G. Individualized developmental care for the very low-birth-weight preterm infant: medical and neurofunctional effects. *Jama* 272, 11 (1994), 853–858.
- [2] BLENCOWE, H., COUSENS, S., OESTERGAARD, M. Z., CHOU, D., MOLLER, A.-B., NARWAL, R., ADLER, A., GARCIA, C. V., ROHDE, S., SAY, L., ET AL. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet* 379, 9832 (2012), 2162–2172.
- [3] BOS, A. F., MARTIJN, A., OKKEN, A., AND PRECHTL, H. F. R. Quality of general movements in preterm infants with transient periventricular echodensities. *Acta Paediatrica* 87, 3 (1998), 328–335.
- [4] CURZI-DASCALOVA, L. Développement du sommeil et des fonctions sous contrôle du système nerveux autonome chez le nouveau-né prématuré et à terme. *Archives de pédiatrie* 2, 3 (1995), 255–262.
- [5] MCGUIRE, W., CLERHEW, L., AND FOWLIE, P. W. Infection in the preterm infant. *British Medical Journal* 329, 7477 (2004), 1277–1280.
- [6] PRECHTL, H. F. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Human Development* 23, 3 (1990), 151–8.
- [7] PRECHTL, H. F., EINSPIELER, C., CIONI, G., BOS, A. F., FERRARI, F., AND SONTHEIMER, D. An early marker for neurological deficits after perinatal brain lesions. *Lancet* 349, 9062 (1997), 1361–3.
- [8] WORLD HEALTH ORGANIZATION. Born too soon: the global action report on preterm birth.

CLINICAL CONTEXT AND OBJECTIVES

This first chapter describes the clinical context of our work. It gives an insight into prematurity in terms of definitions and particularities as well as of their clinical management, especially in Neonatal Intensive Care Units (NICU). A focus is then made on sleep assessment. In addition, the objectives of the thesis are introduced.

1 Prematurity

1.1 Definitions

Several definitions are necessary to clarify the discussion about premature particularities. First of all, according to the American Academy of Pediatrics [5], neonate ages can be reported with several terminologies: Gestational Age (GA), PostMenstrual Age (PMA), chronological age or corrected age (see Figure 1.1). Here, we will generally employ GA and PMA to define newborn ages. In fact, GA represents the duration, in weeks, between the first day of the last menstrual period and the birth date and thus, proposes a fixed age to refer and identify premature babies, independently from their current age (date of assessment). In contrast, PMA, also in weeks, defines the duration from the last menstrual period to the date of assessment and consequently offers an evolving age to highlight newborn evolution.

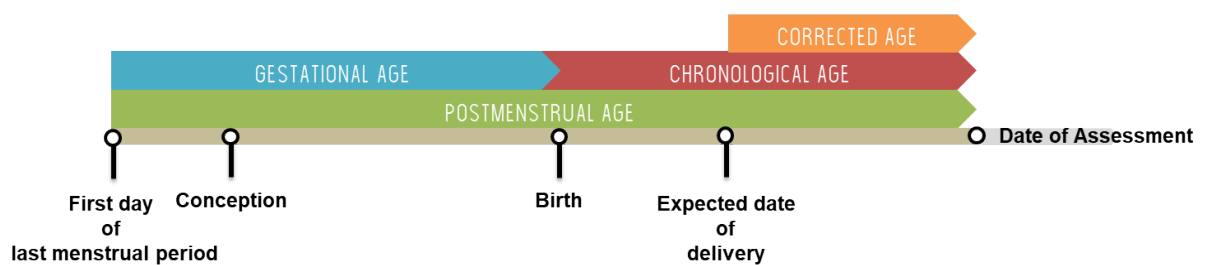


Figure 1.1 – Age terminology during the perinatal period (American Academy of Pediatrics [5]).

A newborn is considered as premature if his/her birth occurred before 37 weeks GA, meaning before eight and a half months of pregnancy. In contrast, full-term neonates are born between 37 and 42 weeks GA. Premature severity is declined in three different categories that represent, respectively, 5, 10 and 85% of the total premature births:

- extremely preterm newborns (under 28 weeks GA);

- very preterm newborns (from 28 to 32 weeks GA);
- moderate or late preterm newborns (from 32 to 37 weeks GA).

Besides these considerations, newborns weighting less than 1500 g, regardless their GA, are also considered as premature and are identified as very low-birth-weight infants.

A preterm birth occurs for a variety of reasons. Among them, multiple pregnancies, infections or chronic conditions, such as diabetes or high blood pressure, have been pointed out to be important factors, although no specific cause is often identified [18]. Two broad classes of preterm births can be distinguished: spontaneous birth and provider-initiated birth [6]. Spontaneous birth regroups spontaneous onset of labor or Premature Rupture of Membranes (PRM) whereas provider-initiated birth (induction of labor or elective caesarean) can be induced in case of maternal or fetal infection or for other non-medical reasons.

1.2 Risks due to prematurity

Prematurity is the main cause of neonatal mortality [16]. However, as depicted by Figure 1.2, survival rates differ according to preterm newborn categories.

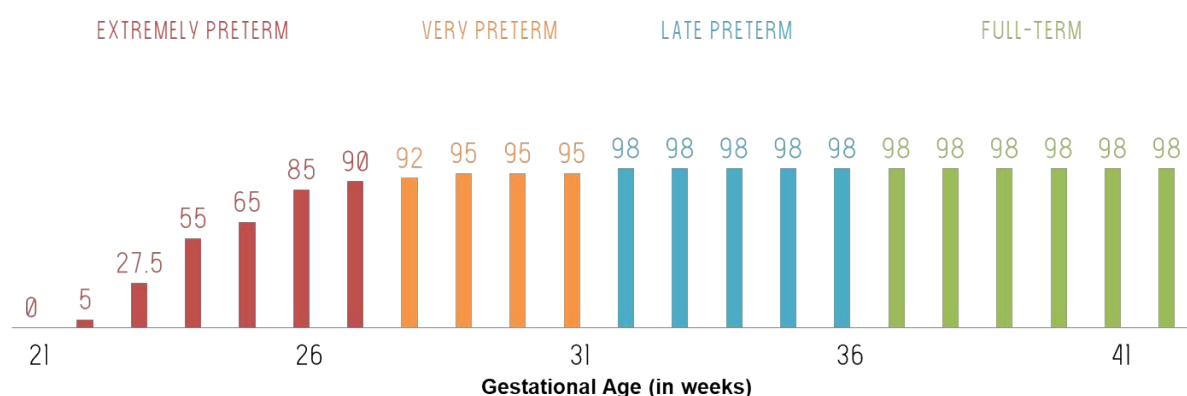


Figure 1.2 – Average survival rates (in %) of premature newborns according to gestational age (in weeks).

Hence, survival rate of extremely preterm newborns is varying between 0 and 90% while infants born after 29 weeks GA are more likely to survive (more than 95%) [17]. Nevertheless, it is important to report that these statistics are different in low-income countries where half of the babies born before 32 weeks GA continue to die because of a lack of feasible and cost-effective care [18]. Due to the immaturity of vital functions such as the central nervous, cardio-respiratory or digestive systems, leaving in extra-uterine environment is challenging for preterm infants. The higher is the degree of prematurity, the more severe are their health problems. The most common are:

- **Neurological:** several diseases such as cerebral palsy or intraventricular hemorrhage are consequences of their immature brain;

- **Cardiorespiratory:** immaturity of the lungs and of the respiratory control system can lead to respiratory distress syndrome, apnea-bradycardia or chronic lung disease;
- **Immunological:** preterm infants are more vulnerable to infection caused by virus or fungi (e.g., sepsis or pneumonia) because of an incomplete immunological system;
- **Thermal:** maintaining corporal temperature is also challenging for premature infants due to their small size and low body fat mass;
- **Digestive:** the gastrointestinal system being not fully developed, breast milk or formula cannot be properly digested. In addition, preterm infants are usually unable to coordinate sucking and swallowing;
- **Metabolic:** the immaturity of organs such as pancreas or liver may induce hypoglycemia or jaundice.

Moreover, over the longer term, premature infants present a higher risk of morbidity than full-term newborns. They are subject to live with long term developmental impairments and behavioral sequelae in the transition to adulthood [13]. Among them, we can cite attention deficit hyperactivity disorder, hearing loss (5 to 10% of extremely preterm) or visual impairments (around 25% of extremely preterm). In children who are born extremely or very preterm, cognitive and neuromotor impairments at 5 years of age increase according to a decreasing gestational age. In fact, near 40% of these preterm infants present motor, sensitive or cognitive impairments while 12% of the full-term children are concerned. Impairments can be divided in three levels: severe (5%), moderate (9%) and mild (25%) [10].

1.3 Presentation of Neonatal Intensive Care Units

Neonatal Intensive Care Units were introduced in the seventies. From there, the neonatal mortality drastically decreased and NICU became essential to neonatal care. These units are designed to provide a specialized medical care for sick term and preterm infants in order to ensure their optimal development. Along his/her hospitalization, the newborn goes through different care configurations that mainly depend on the maturation of each of his/her immature physiological functions. In practice, the three major criteria that lead to the discharge home are the following ones:

- independent thermoregulation to maintain normal body temperature;
- sufficiently mature respiratory control;
- ability to feed by mouth to support appropriate growth.

First, in order to present the whole hospitalization process, the typical care of extremely preterm newborns, regarding the three major criteria, is described below. Nevertheless, each preterm or sick newborn, regarding its disabilities and development, is integrated in this process at the appropriate step. Then, physiological and neurobehavioral monitoring in NICU are presented.

Thermal regulation An extremely preterm newborn system is not capable to regulate the body temperature to deal with extra-uterine environments. Thus, the central element of any NICU room is the

incubator (Figure 1.3(1)), a bed enclosed in a plastic shield that keeps the newborn in a controlled environment regarding temperature and humidity. In addition, it ensures a minimal exposure to germ and external noise. The use and the settings of incubators depend on the special needs of each newborn. Incubator thermal regulation can be based either on a configured targeted ambient temperature or on baby's skin temperature. Then, depending on the development of the baby thermoregulation, clinicians decide to place the infant in successive environments, from radiant warmer (Figure 1.3(2)) to cradle (Figure 1.3(3)) without thermal regulation.

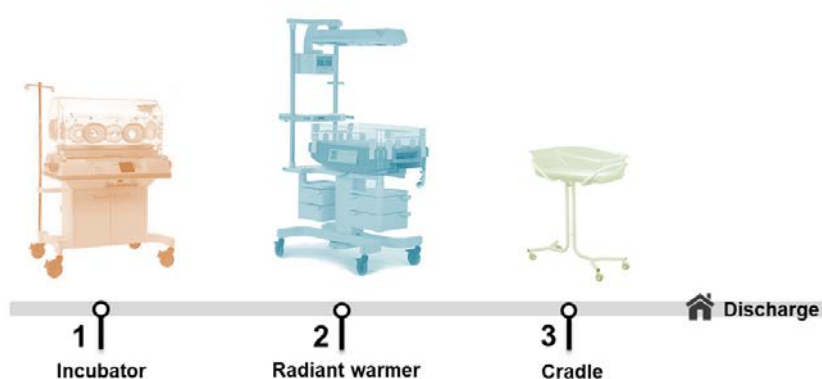


Figure 1.3 – Three examples of beds of NICU.

Respiratory assistance Several breathing assistance techniques are used to fulfill infant oxygen needs. Extremely preterm infants experience several respiratory distress events such as apnea or bradycardia (slow heart rate), that require an immediate intervention. At first, an invasive procedure such as intubation supplemented by a ventilator assistance device is given. Later, in function of the respiratory development of the newborn, intubation is progressively replaced by less invasive elements like nasal cannula. All along the care, clinicians evaluate the needs of the newborn by different means like the analysis of the evolution of the respiratory distress events or blood testing (twice a day).

Feeding A similar strategy is applied to feeding. First, extremely preterm infants are equipped by devices that provide the necessary food: tube feeding from nose to stomach and central catheter that can be supplemented by perfusion. These feeding elements are gradually taken off until the newborn is capable to eat by mouth. Usually, feeding ability is the last step to validate before the discharge.

Physiological monitoring All along the hospitalization, the newborn is continuously monitored. As mentioned above, the temperature of the skin of the baby can be measured to regulate the temperature of incubator or radiant warmer. In addition, cardiorespiratory distresses require immediate nursing actions. The heart and breathing rates or arterial/central venous pressure are then continuously acquired and alarms are triggered in case of distress. These signals are collected by electrodes placed on the infant's body. Non-invasive pulse oximetry can also be performed to measure the blood oxygen saturation and pulse rate through a photodetector. It is used in most NICUs as a detector for desaturation (sudden

decrease of blood oxygenation) and hypoxemia (underoxygenation) events that can cause neurologic sequelae.

Neurobehavioral monitoring Contrary to the physiological monitoring, neurobehavioral monitoring is performed in a more punctual manner. In practice, it is mainly based on sleep analysis.

On request of doctors, ElectroEncephaloGraphic (EEG) examinations can be performed to evaluate possible brain damage. In that case, an ambulatory system is deployed in the newborn room and some electrodes are placed on his/her head for the signal acquisition.

To evaluate the development of the extremely and very preterm newborns, the Newborn Individualized Developmental Care and Assessment Program (NIDCAP) suggests to examine the infant every 15 days until the discharge. This examination lasts one hour and is performed by a trained nurse which visually annotates the behavior of the infant (e.g. sleep states, motor and facial activities) within a 2-minute granularity. However, this observation being time consuming, only a small part of the newborns benefits from this follow-up.

1.4 Sleep assessment

At birth, the brain of preterm infants is less mature than that of full-term infants. This results to immaturity of sleep mechanisms. Sleep cycling can thus be studied as a physiological biomarker of developmental neural plasticity to predict developmental outcomes [1, 4, 8]. Sleep-wake cycling is essential for early neurobehavioural development, learning, memory, and preservation of brain plasticity [7].

The first study on newborn behavior was conducted by Prechtl and Beintena in 1967 [12]. It demonstrated the predictable occurrence of physiological rhythms, also called behavioral states. Four components were observed (eyes state, breathing, movements and crying) to compose five behavioral states (Figure 1.4).

	EYES	BREATHING	MOVEMENT	CRYING
QUIET SLEEP	Closed	Regular	No	No
ACTIVE SLEEP	Closed	Irregular	Small	No
DROWSINESS	Open / Closed	Regular	Small	No
QUIET ALERT	Open	Irregular	Gross	No
ACTIVE ALERT	Open / Closed	Irregular	Gross	Yes

Figure 1.4 – Behavioral states description according Prechtl and Beintena.

To date, several classifications of behavioral states have been proposed. Among them, the most used is the Neonatal Behavioral Assessment Scale (NBAS) integrated in NIDCAP. This assessment is directly performed from visual observations of the newborn behavior and is based on the scale proposed by Prechtl and Beintena. In NBAS, each state has been nominated and associated with a definition:

- **Quiet Sleep (QS).** The newborn is quiet (no movement and no crying), eyes closed, with a regular breathing.
- **Active Sleep (AS).** Contrary to QS, newborns can slightly move with an irregular breathing.
- **Drowsiness (D).** It is an intermediate state that occurs between sleep and wake stages (AS and QA). In this transitional state, eyes can be open, a variable activity level can be observed such as sporadic bursts of movement or moaning. Breathing is mostly regular.
- **Quiet Alert (QA).** The newborn is awake, eyes are open, the infant may twist. Breathing is irregular.
- **Active Alert (AA).** The newborn presents vigorous movements that can involve several limbs. The baby may cry. Eyes are indifferently open or closed.

The evolution of sleep states is graphically represented as a function of time known as hypnogram (cf Figure 1.5).

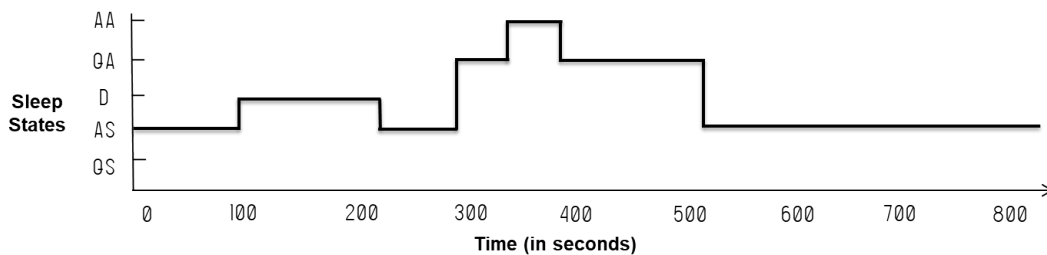


Figure 1.5 – Example of hypnogram for a newborn born at 31 weeks GA and recorded at 32 PMA.

Normal development of preterm newborns is characterized by a rapid evolution of the organization of the sleep states. More precisely, as illustrated by Figure 1.6, the time spent in QS and in AS increases progressively along with conceptual age (i.e., as well as with PMA) [3]. In addition, the amount of waking states (QA and AA) also increases [9].

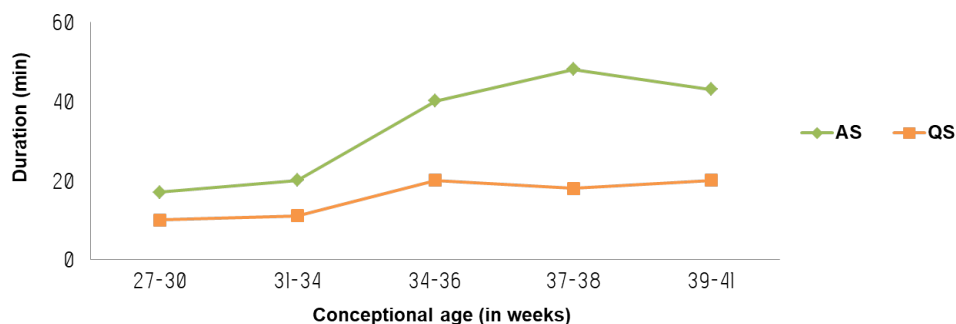


Figure 1.6 – Evolution of the duration of AS and QS from 27 to 41 weeks of conceptual age in neurologically normal infants (taken from [3]).

Nevertheless, this observational technique hardly integrates the clinical routine since it has to be performed in the presence of the newborn. Moreover, behavioral state annotation requires a high training

and experience. In fact, as it can be felt by reading the state definitions, several states can be easily confused and the evaluation remains expert dependent. Hence, the development of automatic and fast procedures is necessary to better monitor neurobehavioral development of newborns.

2 Objectives of the work

2.1 The thesis

The objective of this thesis is to propose non-invasive solutions in order to provide a continuous behavioral monitoring in NICU.

We have seen that observational techniques are usually used in clinic, notably for sleep analysis. However, to date, it is time consuming and sparsely performed by experimented nurses. Within the objective to overcome these limitations, the attention of this work is then focused on the development of video and audio algorithms. Three main objectives were defined:

- Study of the existing literature about the audio- and video-based methods developed in the context of pediatric health;
- Design of an audio-video acquisition system for NICU;
- Development of audio- and video-based methods to continuously monitor the neuro-behavioral development of premature newborns.

2.2 The Digi-NewB project

In parallel with this work, the Digi-NewB project was conducted. This project, summarized by Figure 1.7, is a Horizon 2020 European project that began in March 2016 and will end in March 2020.



Figure 1.7 – Digi-NewB key values.

Using clinical and signal data (cardiac and respiratory traces, video quantification of movement and sound) from a large cohort of hospitalized newborns ($n=780$), Digi-NewB aims to propose a new gen-

eration of non-invasive monitoring in neonatology. The objective is to reduce mortality, morbidity and health costs of hospitalized newborns by assisting clinicians in their decision-making of sepsis risk and of cardio-respiratory and neurobehavioral maturation.

To date, no monitoring system has been proposed to automatically characterize behavior of preterm newborns. Reversely, the development of systems to predict sepsis has been more addressed [2, 11, 15]. Most of these systems were only based on heart rate analyses [2, 11, 14, 15]. Their efficiency to reduce mortality in VLBW patients and to decrease length of stay of the infected surviving VLBW infants was shown during randomized control clinical trials [11, 15]. The same conclusions were reached by a machine learning approach based on the collection of clinical signs [14]. These recent publications show the relevance of developing new monitoring system by collecting data from multiple sources.

To achieve these goals, a consortium of seven partners from four countries (Finland, France, Ireland and Portugal) has been formed, composed by:

- a paediatric clinical network (GCS HUGO, France), responsible of the project coordination and of the clinical protocol conduction in six hospitals;
- two small and medium companies (Voxygen, France and Syncrophi, Ireland);
- four university groups with multidisciplinary expertise in signal processing (Instituto de Engenharia de Sistemas e Computadores, Portugal and LTSI, Université de Rennes 1, France), in cardiovascular modeling (LTSI, Université de Rennes 1, France), in multivariate analyses (Tampere University of Technology, Finland) and in user centered design approach (National University of Ireland Galway, Ireland).

The objective is that this new type of monitoring leads to the development of novel preventive and therapeutic strategies to counteract late diagnosis of sepsis and inappropriate evaluation of maturity.

The overall workflow of the Digi-NewB project is reported in Figure 1.8. Each partner is responsible of a work package that goes from the acquisition of the data to the design of a decision support system.

This thesis is involved in two main work packages: Work Package 2 (WP2) and Work Package 3 (WP3). WP2, driven by Voxygen, concerns the design and integration of a multisensor acquisition system for NICU. WP3, directed by LTSI - Université de Rennes 1, has for objective the processing of the acquired data to extract clinically relevant features regarding the two clinical targets: sepsis risk prediction and maturation evaluation.

3 Conclusion

This chapter underlines the necessity to design new monitoring solutions in order to improve the newborn care. In particular, the examination of neurobehavioral development is essential to evaluate newborns outcomes. However, to date, this examination is time consuming, expert-dependent and sparsely performed.

We have also seen that systems have been already developed but that they did not gather as a whole the physiological data, the behavioral and the clinical signs. One of the objectives of the Digi-NewB project is to enhance the scope of classical collected data (physiological and clinical signs) by the addition of audio and video components in order to propose non-invasive alternatives.

Within this objective, this thesis focuses on the development of audio and video methods. To do this task, it was first important to identify the weak and strong points of audio and video analyses. The next chapter is then dedicated to the audio and video features explored in the literature that were demonstrated clinically relevant in pediatrics.

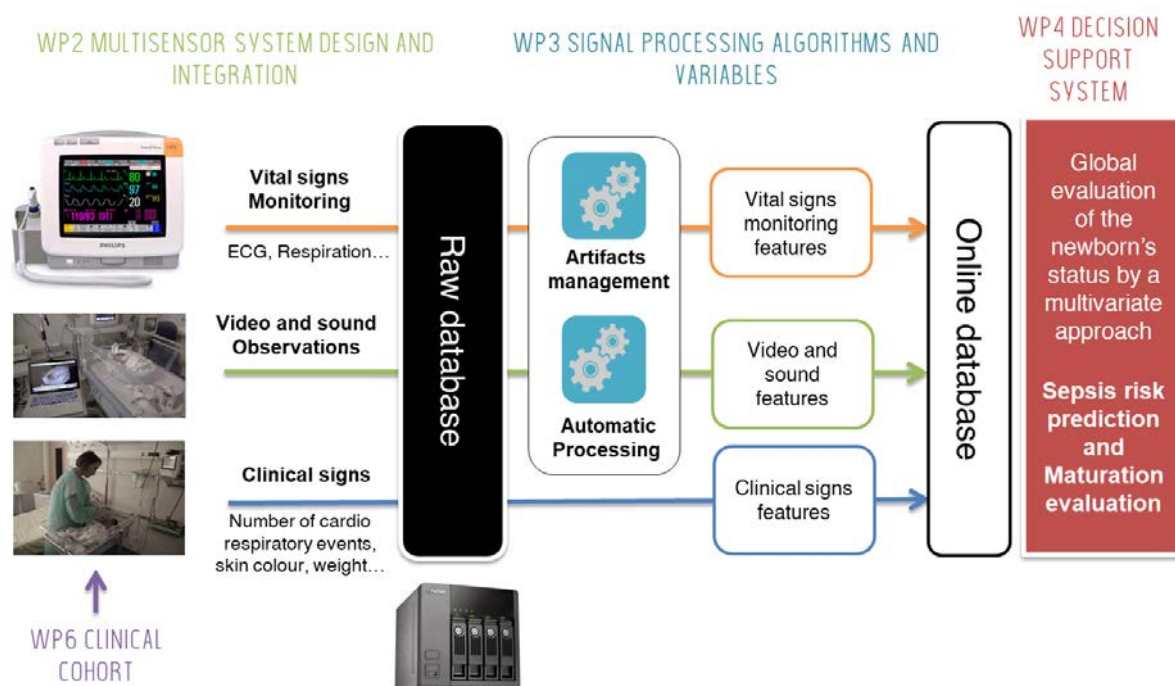


Figure 1.8 – Digi-NewB workflow.

Bibliography

- [1] ANDERS, T. F., KEENER, M. A., AND KRAEMER, H. Sleep-wake state organization, neonatal assessment and development in premature infants during the first year of life. ii. *Sleep* 8, 3 (1985), 193–206.
- [2] COGGINS, S. A., WEITKAMP, J.-H., GRUNWALD, L., STARK, A. R., REESE, J., WALSH, W., AND WYNN, J. L. Heart rate characteristic index monitoring for bloodstream infection in an nicu: a 3-year experience. *Archives of Disease in Childhood-Fetal and Neonatal Edition* 101, 4 (2016), F329–F332.
- [3] CURZI-DASCALOVA, L. Développement du sommeil et des fonctions sous contrôle du système nerveux autonome chez le nouveau-né prématuré et à terme. *Archives de pédiatrie* 2, 3 (1995), 255–262.

- [4] DAVIS, D. H., AND THOMAN, E. B. Behavioral states of premature infants: Implications for neural and behavioral development. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology* 20, 1 (1987), 25–38.
- [5] ENGLE, W. A. Age terminology during the perinatal period. *Pediatrics* 114, 5 (2004), 1362–1364.
- [6] GOLDENBERG, R. L., CULHANE, J. F., IAMS, J. D., AND ROMERO, R. Epidemiology and causes of preterm birth. *The lancet* 371, 9606 (2008), 75–84.
- [7] GRAVEN, S. Sleep and brain development. *Clinics in perinatology* 33, 3 (2006), 693–706.
- [8] HOLDITCH-DAVIS, D., AND EDWARDS, L. J. Temporal organization of sleep–wake states in preterm infants. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology* 33, 3 (1998), 257–269.
- [9] HOLDITCH-DAVIS, D., SCHER, M., SCHWARTZ, T., AND HUDSON-BARR, D. Sleeping and waking state development in preterm infants. *Early human development* 80, 1 (2004), 43–64.
- [10] LARROQUE, B., ANCEL, P.-Y., MARRET, S., MARCHAND, L., ANDRÉ, M., ARNAUD, C., PIERRAT, V., ROZÉ, J.-C., MESSER, J., THIRIEZ, G., ET AL. Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the epipage study): a longitudinal cohort study. *The Lancet* 371, 9615 (2008), 813–820.
- [11] MOORMAN, J. R., CARLO, W. A., KATTWINKEL, J., SCHELONKA, R. L., PORCELLI, P. J., NAVARRETE, C. T., BANCALARI, E., ASCHNER, J. L., WALKER, M. W., PEREZ, J. A., ET AL. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *The Journal of pediatrics* 159, 6 (2011), 900–906.
- [12] PRECHTL, H. F. The behavioural states of the newborn infant (a review). *Brain research* 76, 2 (1974), 185–212.
- [13] SAIGAL, S., AND DOYLE, L. W. An overview of mortality and sequelae of preterm birth from infancy to adulthood. *The Lancet* 371, 9608 (2008), 261–269.
- [14] SHIMABUKURO, D. W., BARTON, C. W., FELDMAN, M. D., MATARASO, S. J., AND DAS, R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ open respiratory research* 4, 1 (2017), e000234.
- [15] SWANSON, J. R., KING, W. E., SINKIN, R. A., LAKE, D. E., CARLO, W. A., SCHELONKA, R. L., PORCELLI, P. J., NAVARRETE, C. T., BANCALARI, E., ASCHNER, J. L., ET AL. Neonatal intensive care unit length of stay reduction by heart rate characteristics monitoring. *The Journal of pediatrics* (2018).
- [16] TUCKER, J., AND MCGUIRE, W. Epidemiology of preterm birth. *British Medical Journal* 329, 7467 (2004), 675–678.

- [17] TYSON, J. E., PARIKH, N. A., LANGER, J., GREEN, C., AND HIGGINS, R. D. Intensive care for extreme prematurity-moving beyond gestational age. *New England Journal of Medicine* 358, 16 (2008), 1672–1681.
- [18] WORLD HEALTH ORGANIZATION. Born too soon: the global action report on preterm birth.

METHODOLOGICAL CONTEXT: VIDEO AND AUDIO PROCESSING IN PAEDIATRICS

This second chapter is important and reports a part of the methodological context of this work. Indeed, it lays the groundwork of the audio and video processing methods that were developed in paediatrics. For this purpose, more than 150 papers were reviewed. For both topics, clinical applications are described according to the considered cohorts, either full-term newborns, infants and toddlers, or preterm newborns. Then, processing methods are presented, in terms of data acquisition, feature extraction and characterization. The content of this chapter has been published in *Physiological Measurement* [19]. In this article, methods for video are first addressed. Audio analyses are then described. To finish, a discussion section is provided in order to give an insight of the strengths and limitations of these methods for monitoring purpose.

1 Background

The analysis of neonatal and early childhood development is at the center of concerns of the medical community. Especially, premature babies, having several vital immature functions, receive a particular attention by the recording, in Neonatal Intensive Care Unit (NICU), of several physiological signals [57], however limited by their fragility. On the other hand, video and audio acquisitions have the advantage to propose contactless and non-invasive ways to collect data for patients being cared for in hospital or at home. Such technologies having known great improvements in the last decades, they are now used in many biomedical applications.

The use of audio and video in paediatrics found certainly its roots in sleep analysis when the Association of the Psychophysiological Study of Sleep (APSS) stated in 1969 the necessity to develop a guide for scoring sleep in infants, since the criteria of Rechtschaffen and Kales [115] were "applicable only to the adult" and had "not taken into account the unique features of the developing infant" [48]. In 1971, Anders *et al.* published a manual recommending to support polysomnographic recordings by behavioral observations [9] and then carried out a study with full-term infants, where behavioral states were scored from video and audio acquisitions, according to eye state, movements and crying vocalizations [10]. Later on, several approaches using audio and video recordings were proposed to analyze sleep either on premature newborns [42], or on children for the evaluation for Obstructive Sleep Apnea (OSA) [87, 127]. Finally in 2011, ASA/ASTA (Australasian Sleep Association/Australasian Sleep Technologists' Association) Paediatric Working Group recommended to record audio and video as additional

information to the electrophysiological signals in the scoring of children sleep [101].

Along with these works, in 1990, Prechtl developed a method to assess the quality of General Movement (GM) based on video observations as a diagnostic tool for early detection of brain dysfunction [109, 110]. From there, General Movement Assessment (GMA) using video has been applied in several clinical contexts. In the same way, the detection of neonatal seizures also led to a lot of works including video acquisitions [104]. Such analyses relied, in the oldest studies, on manual annotations of the videos that began to be automatically processed only in the 2000s, thanks to the improvements in digital video processing.

The development of automated sound processing occurred earlier since studies on newborn cries began from the 1960s with Wasz-Höckert *et al.*, where it was shown, by spectrographic analysis, that four different types of cries could be distinguished as birth, pain, hunger, pleasure (see [147] for an historical review). From there, a huge literature arose, with the analysis of frequency features, in children and newborns. Several studies performed also detailed analyses of crying behavior in preterm newborns either solely, or in comparison with full-term newborns.

The objective of the present paper is to synthesize this abundant literature, in the context of paediatrics, and identify its clinical impacts. Specifically, the motivation of our work is to offer an overview of the existing audio and video processing methods to evaluate their potential application for monitoring in NICU. The paper is organized in two main sections: Section 2 is devoted to video processing, while audio analysis is described in Section 3. For both topics, clinical applications are first described. Since these applications differ depending on the studied population age, especially considering audio analysis, they are presented first for full-term newborns (0 to 2 months old), infants (2 months to 1 year old) and toddlers (1 to 4 years old) and then for preterm newborns. Then, processing methods are presented, in terms of data acquisition, feature extraction and characterization. Finally, last section draws the main limitations of these studies but also gives some propositions, in the objective of developing automatic monitoring systems able to meet clinical needs in NICU.

2 Video analysis

Since the motion of a newborn is one of the most crucial information to describe his physio-pathological state, it has been the most extracted descriptor from video recordings. The estimation of other types of information has also been investigated such as respiration, heart rate and facial expression.

In this section, main clinical applications supported by video recordings are first presented. However, since, for some applications, video recordings were analyzed manually (especially for preterm), some studies with manual video analysis are included in order to present all the potential applications of video analysis. Then, data acquisition and processing methods are described.

2.1 Clinical applications

2.1.1 Full-term newborns, infants and toddlers

General Movement Assessment was used to support many studies regarding different pathologies. In [51], the visual assessment of motion patterns from video enabled to early detect hemiplegia (complete or partial palsy) in infants with cerebral infarction (not enough blood supply in a region of the brain). Mazzone *et al.* found that abnormal GMs were early markers of motor impairment (partial or total loss of function of a body part) in infants with Down Syndrome [83]. Nearly ten years after, automatic video processing was applied by different groups to study Cerebral Palsy (CP), which is a disorder of movements caused by an abnormal motor control center of the brain [95, 130].

A large part of studies using video recordings was dedicated to the analysis of neonatal seizures. Whereas first studies focused on observational classifications of seizures from video recordings [86, 137], later, thanks to the development of video processing, several approaches were proposed to automatically detect or classify seizures from motion descriptors [61–68, 89, 121]. For their part, Cuppens *et al.* focused on the specific case of the epilepsy [27, 28].

Studies regarding physiological monitoring with the support of video processing have also been conducted. Respiratory frequency has been estimated in order to prevent Sudden Infant Death Syndrome (SIDS) [35] or to detect repeated apnea events in the context of Congenital Central Hypoventilation Syndrome (CCHS) [22, 23]. Pulse rate has also been estimated from video acquired during Hammersmith Infant Neurological Examinations (HINE) [126].

Emotion and facial expression detection also received a particular attention. Face analyses were performed either to discriminate, between the behavioral states sleep, awake and cry [53], or to automatically analyze infants emotion during interactions [151, 152].

2.1.2 Preterm newborns

Like for the previous population, video has been mainly used for General Movement Assessment regarding preterm infants. Visual general movement assessment on video recording has been proved successful to determine if infants had brain dysfunctions, either transient or persistent, and to identify infants at risk for impaired neurological outcomes [17]. Later, Bos *et al.* also observed in videotape recordings the effects of the dexamethasone therapy on high-risk very preterm infant through GMA [18]. Moreover, Spittle *et al.* showed that abnormal GMs directly reflect white matter injury [129]. Among these applications, no video processing methods were developed to automatize the assessment. More recently, Adde *et al.* focused on the automatic prediction of CP by computer-based video analysis of general movements [6, 7, 113, 114].

In the meantime, video-based analyses were performed to assess sleep quality, either visually, where video was used as a support for actigraphy measurement validation [128, 134] or to evaluate the influence of light exposure on the sleep [60], or semi-automatically, to detect the eye state in different sleep stages [108].

Other clinical investigations through video processing have been conducted. Among them, Pulled-To-Sit examination of infants was performed by the use of object tracking methods [30]. Later, with the

ambition of a contactless and non-invasive monitoring, Villarroel *et al.* elaborated a monitoring solution of vital signs (respiratory rate, oxygen saturation and heart beat) through video analysis [145]. Some authors also succeeded to estimate the infants heart rate from video imaging [1, 140] and Koolen *et al.* estimated respiration rate by motion extraction [69]. Meanwhile, Zamzami *et al.* worked on pain assessment by classifying infant's expression [153].

2.2 Methods for video processing

Automatic video processing in paediatrics has led to a large number of papers, with different objectives, including motion extraction and characterization, respiration and heart rate estimation and face analysis such as depicted, in Figure 2.1. The next section presents these works. As a first step, a special attention is paid to video databases.

2.2.1 Video acquisition

Recording settings are important quality factors and yet, only one group observed the impact of the camera set-up on motion analysis, including spatial and temporal resolutions, compression, illumination and location of moving body parts in the images [27, 28]. Nonetheless, all authors usually respected reasonable values in term of resolution, frame rate or compression and solutions to overcome the impact of the illumination and location variability were developed (noise reduction techniques, computation of features independent of the amplitude. . .).

Two database types can be identified regarding camera/patient positioning. The more common one contained video recordings performed in a controlled environment where cameras were located above a mattress and infants were placed in a supine position with no blanket and fully visible [6, 7, 61–68, 95, 113, 114, 121, 130]. A marginal positioning has also been used in [151, 152] where infants were placed in a baby seat with only the head and shoulders visible. In the second type, video data collection was integrated in the medical care routine during scheduled examinations such as HINE [30, 126] or directly in the bedroom [1, 22, 23, 27, 35, 53, 69, 89, 108, 140, 145]. This ambition to integrate the care routine has led to a higher variability in camera positioning. Nevertheless, cameras were mostly placed above one corner of the foot-end of the bed, recording the full body of the baby [22, 23, 27, 69, 89, 108].

For studies focused on face analysis or heart rate, camera position may differ and be on the middle-left of the bed [1]. Zoom has been applied to focus on the head [1, 53, 153]. Dealing with closed incubators, Villarroel *et al.* installed their camera above the infant, against the plastic wall [145]. Profile point of view was adopted for the recordings performed during HINE [30, 126].

Most of the databases were composed by selected short video sequences from a dozen of seconds to dozens of minutes. Some exceptions were noted in [22] with more than one hour and half of recording and in the work on continuous vital signs monitoring with initial recordings durations between 50 minutes and more than seven hours [145].

The size of the studied population was very varied, from a very small number of infants (one to three) [22, 95] to a more consequent cohort (more than 30 patients) [7, 64].

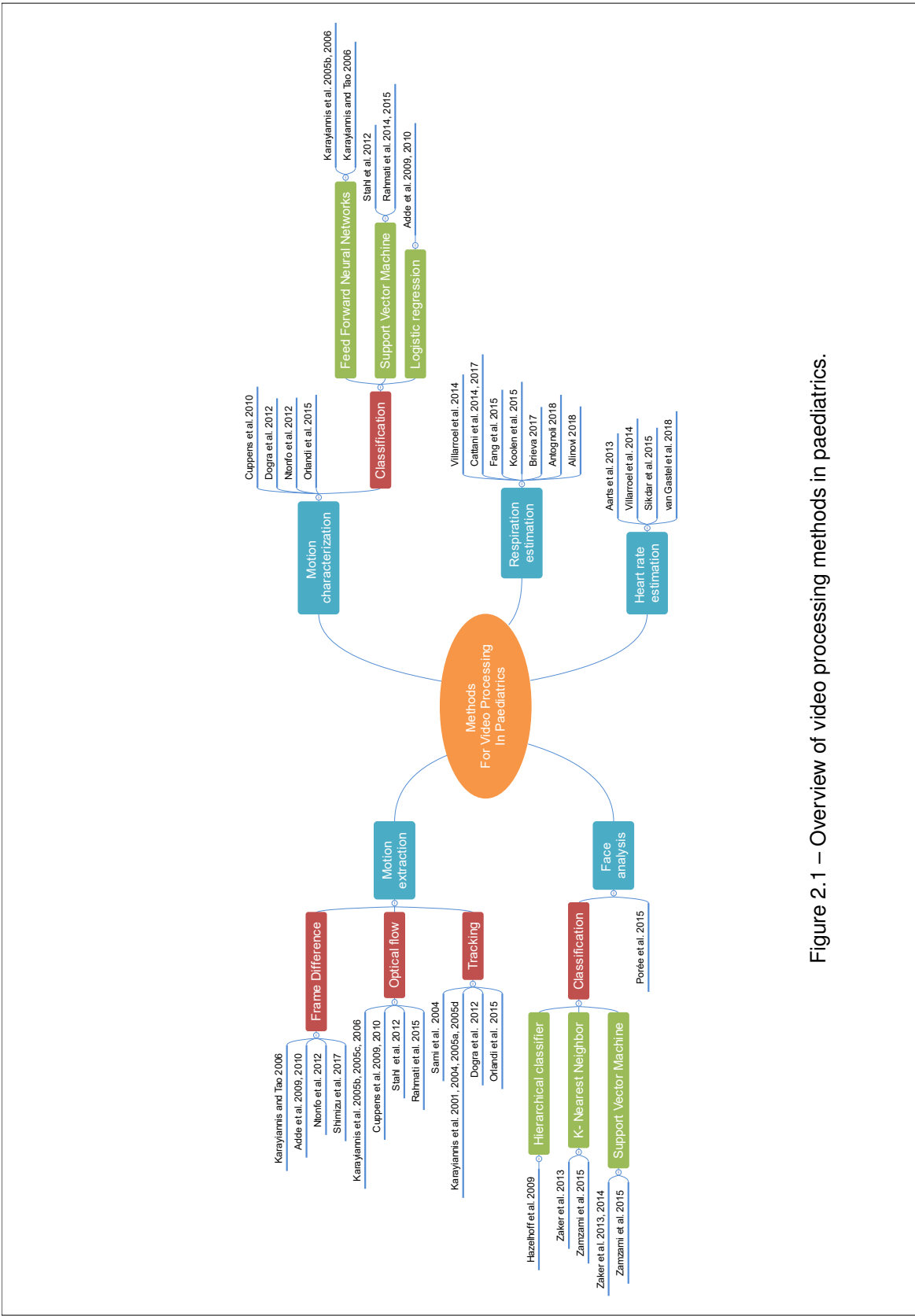


Figure 2.1 – Overview of video processing methods in paediatrics.

2.2.2 Motion extraction

Three popular methods have been used to automatically extract body (global) or limbs (local) movements: frame differencing, optical flow and tracking.

Frame differencing consists in computing the difference between the current and one or multiple previous frames. The resulting difference image contains the global motion information, including the infant's movements. However, this difference image is generally very noisy, since it is sensitive to very low intensity changes and to the original images noise. Moreover, since the detection is based on intensity changes, it enables to identify the contours of the moving parts, while being less sensitive to the motion of areas with homogeneous intensities.

Hence, all the studies have considered post-processing to limit the influence of noise. Intensity thresholding or median filtering have been widely used to remove small intensity differences [64, 89, 124], followed by morphological operators or a threshold applied on objects size to remove the small regions [6, 7]. The use of a vector clustering was also proposed to identify the different regions of interest [64]. Finally, most authors summarized the resulting images by computing a motion signal, sometimes called "motiongram", "quantity of motion" or "motion strength signal", corresponding to the area of the moving parts [6, 7, 64]. This signal was used for further processing in order to characterize the motion (see section 2.2.3).

Optical flow is the velocity field generated by the relative motion between an object and the camera in a sequence of frames. Contrary to frame differencing which detects the moving areas but does not compute the displacements, optical flow estimation methods return a velocity or a displacement vector for each pixel. Many methods for optical flow have been proposed in the literature [133], most of them being based on the hypothesis that the intensity of one pixel remains constant between any two following frames and that the motion between two frames is small. Considering infant's motion extraction, the original method of Horn and Schunck [56] has been used by different teams [27, 28, 65–67]. This approach is based on the minimization of a metric combining two terms, a data term based on the intensity conservation constraint and a smoothness, or regularization, term modelling the spatial distribution of the flow field. The importance of formulating an appropriate smoothness term, e.g. using quadratic functions, and to tune the associated weighting factors has been demonstrated [65]. In the same sense, Stahl *et al.* used a method based on a non-local regularization term, imposing a smoothness assumption within a specified region of the flow field [130]. In order to segment and track infant's individual body parts, Rahmati *et al.* proposed, starting from some manual labels in one or more frames, to use a multi-scale optical flow combined with a tracking of the segmented parts [114]. As for the studies based on image differences, a temporal signal was finally extracted from the computed flow fields. It is generally based, at each frame, on the area of pixels corresponding to velocity vectors of magnitude higher than a threshold.

Tracking methods rely on the selection (either automatically or manually) of points or regions of interest and on their tracking over the image sequence. They rely on the assumption that the considered features keep a constant appearance over time and motion. This tracking is based either on local intensity differences (template matching methods) or on higher level descriptors (feature matching methods). Both approaches have been used to track either specific parts of the infants, generally selected man-

ually (e.g. hands and feet [95] or center of the head [30]) or more general points of interests (e.g. corners [63]). Karayiannis *et al.* especially evaluated different trackers based on block matching. Starting from the Kanade-Lucas-Tomasi (KLT) tracker [63], they considered adaptive block matching [62], predictive block matching [61] and a variety of other block models associated with the automatic localization of moving body parts [121]. If these approaches had the advantage of estimating the motion of specific parts of the infants, and thus enabled to study precisely their motions, they were sensitive to occlusions, complex motions (e.g. rotations) and noise [68].

In short, the considered motion estimation methods have shown their ability to estimate the infant's motion. However, these approaches often relied on parameters which have to be tuned according to the data (e.g. noise, motion amplitude) and are sensitive to occlusions. Moreover, they may not differentiate the infant's motion from other people's (parents, medical staff) motion, thus needing some manual steps to identify regions of interest, making difficult the use of these approaches on long recordings.

2.2.3 Motion characterization

Once the motion signal was extracted, high level features were computed to characterize movements, sometimes in order to automatically classify infants' impairments.

The first step was often to identify motion and non-motion periods (or epochs) from the motion strength signal. For this purpose, Cuppens *et al.* used two methods: a) a fixed threshold determined thanks to a receiver operating characteristic curve and b) a variable threshold adapted to the noise and computed from mean and standard deviation of a selected non-movement period. They concluded that a variable threshold adapted to each video recording was the best approach with an average Predictive Positive Value (PPV) value of 94% [27]. Considering a tracking of different limbs, Orlandi *et al.* defined non-motion periods when all limbs values were lower than a fixed threshold [95].

Many features have been considered to describe the resulting epochs. Features characterizing the distribution of epochs have been used. Mean, median, maximum, standard deviation of the quantity of motion have been computed [6, 7], such as speed and acceleration, skewness of the velocity and kurtosis of the acceleration, the last two being measurements of the complexity of the speed and acceleration values [95]. The periodicity of the motion signal has also been investigated, e.g. using the autocorrelation function, to characterize clonic seizures [89]. The maximum spike duration and the number of spikes have been computed [64, 66]. Positioning features of the centroid of motion (the spatial center of the positive pixels in the motion image) have also been considered [6, 7]: the mean position in the x- and y-directions, the standard deviation, velocity and acceleration. The motion of individual limbs has also been characterized, e.g. using wavelet and frequency analysis [130], the periodicity [113, 114], the correlation between trajectories [95, 113, 114], or the deviation from a smoothed version of the movement to characterize its smoothness [113]. 2-D modeling of the camera scene had also been used to extract the angle between head and torso during HINE [30].

The next step was to characterize or classify pathological situations based on these features. Pulled-to-sit scores have been computed by applying decision rules regarding the relative movement of the head with respect to the infant torso. The comparison with expert scores led to a sensitivity of 92% along a specificity of 96% [30]. Adde *et al.* aimed at determining infant with CP [6, 7]. By the use of logistic

regression, they evaluated the capacity of each feature to classify the absence or presence of fidgety movements (classical circular movements, linked with the absence of CP). Best results were obtained with the standard deviation of the centroid of motion, with a sensitivity of 81.5% and specificity of 70% while other features showed weak specificities of less than 56% [7]. Later, a cerebral palsy predictor was proposed, calculated as a combination of the previous descriptor with the mean and standard deviation of the quantity of motion. It reached a sensitivity of 85% and a specificity of 88% [6]. These works were extended in [113, 114, 130]. Stahl *et al.* classified infants using Support Vector Machine (SVM) with a linear kernel. They reported an accuracy of 93.7% along with a sensitivity of 85.3% [130]. Combined with an automatic segmentation of infant limb's [113, 114], the same approach presented a total accuracy of 87% which was better than the results obtained with electromagnetic sensors [114]. In [64–66], the goal was to distinguish neonatal seizures from random infant behaviors and to differentiate between myoclonic and focal clonic seizures. Multiple Feed Forward artificial Neuronal Networks (FFNN) were trained, using different sets of features. They reached 85% of specificity and sensibility, with an increase of 5% when a frequency feature was added [65].

2.2.4 Respiration estimation

Several groups aimed to characterize respiration from video recordings. Authors mostly proposed methods to estimate the respiration rate [8, 12, 35, 59, 69, 145] and one group focused on the development of an apnea detector [22, 23].

Eulerian video magnification has been used to magnify the motion of low amplitude related to respiration [12, 23, 59, 69]. It is based on the amplification of pixel color variation in a specified frequency band. This was followed by motion extraction, e.g. using frame differences [22, 23] or optical flow estimation [69], to generate a motion signal which was further processed to characterize the periodicity of the motion. A periodic model whose parameters were estimated using a maximum-likelihood method [8, 22, 23] or a short-time Fourier transform [69] was used. If this process generally considered the whole image, one group performed the magnification in several ROIs before selecting the best ones, i.e. with the higher amplitudes of the estimated periodic variations [8]. Respiratory rate detection was successful for most patients during quiet sleep stages [69]. Authors succeeded to identify 90 to 100% of the apneas detected by polysomnography [22].

For their part, Fang *et al.* revealed slight movements by the use of accumulative sum of difference images. Respiratory signals were then obtained from the frame by frame evolution of the intensity of automatically selected pixels in the accumulative sum. The method showed good results in 46 video sequences [35].

Villarroel *et al.* based their method on four steps: a) automatic identification of stable periods (with no interaction between the infant and the medical staff) by the use of a non-parametric statistical background image estimation; b) computation of the mean intensity of Regions of Interest on each channel RGB; c) calculation of the reflectance photoplethysmogram by independent component analysis; d) identification of regular frequencies using auto regressive models e) estimation of vital signals by filtering the frequencies according to physiologically realistic ranges. Respiratory rate was obtained after the application of a band pass filter cutting off at 0.33 Hz and 1.67 Hz (i.e. between 20 and 100 breaths

per minute) and oxygen saturation was calculated from respiratory signal. Respiratory rate and oxygen saturation estimation were comparable with the Philips monitor values [145].

2.2.5 Heart rate estimation

Four studies proposed non-contact heart rate monitoring utilizing camera for preterm infants. These techniques were entirely contactless and used the principle known as pulse-oximetry. In fact, light is more absorbed by blood than by surrounding tissues so that variations in blood volume affect light transmission and reflectance. This phenomenon can be measured by observing slight color variations in a region of interest.

First, a study showing interesting results to monitor heart rate in NICU was presented by Aarts *et al.* in 2013 [1]. They proposed a method based on four steps: a) tracking of a manually selected area of contrast (e.g. eye, eye brow) that also contained a skin area; b) computation of the average green channel pixel values into the skin area ROI; c) computation of the joint time frequency diagrams (called plethysmograph); d) extraction of the dominant frequency, related to pulse rate. In all 19 infants, heart rates were estimated but not continuously. In fact, low ambient light level and infant motion still remained challenging conditions.

Later, as mentioned above (see section 2.2.4), Villarroel *et al.* developed a non-contact vital signs monitoring, among them heart rate estimation based on the same steps as respiratory rate estimation, except that they used a different band pass filter between 1.3 Hz and 5 Hz (corresponding to 78 beats per minute and 300 beats per minute) [145]. Resulting signals showed a great correlation with ECG-derived measurements from the Philips monitor.

Pulse rate was also extracted within a manually chosen region of the trunk of the infant by the mean of a set of different color (RGB) decomposition [126]. Authors found that the RG and GB channel combinations were more accurate in comparison to the RGB or RB channel combination. This observation confirmed that photoplethysmography signal is strongest on the green channel.

It is important to notice that these methods, based on the analysis of slight color changes, were not adapted to acquisitions with low light levels. Recently, a method adapted to infrared illumination was thus proposed, showing a good estimation of heart rate 87% of the time [140].

2.2.6 Face analysis

Infant's facial expression was analyzed to detect the presence of discomfort [53]. Authors automatically segmented the face from the background by skin color modelling and localized the eye, eye-brow and mouth region thanks to shape assumptions. Then, they classified behavioral states by the use of a hierarchical classifier [53]. However, the study showed weakness in eye-brow segmentation under lighting and viewpoint deviations.

Recent studies on facial analysis presented different features extraction methods [108, 151–153]. Zamzami *et al.* detected the nose first and then expanded the mask to include the eyes and surrounding areas. A facial strain algorithm was applied on the remaining area in order to extract strain magnitude. K Nearest-Neighbors (KNN) and SVM classifier have been trained to discriminate expression between two states: pain and no-pain. KNN approach showed an higher accuracy of 96% [153]. Facial features were

also extracted by the mean of Active Appearance Models (AAM) combined with Histogram of Oriented Gradients (HOG) to detect spontaneous facial Actions Units (AUs). Firstly, authors trained classifier for each AU and the best results were given by a SVM classifier with intra-class correlation values up to 0.73 [151]. Then, they improved the results by training classifiers to detect multi-AUs and reported F-scores between 0.58 and 0.91 [152]. Finally, Porée et al. proposed an algorithm estimating the eye state in videos of premature babies combining tracking with segmentation and characterization steps [108]. The proposed approach gave more than 95% of concordance with a sensitivity and specificity, respectively ranging from 78.5% to 100.0% and from 97.69% to 100.0%.

3 Audio analysis

Acoustic analyses in paediatrics mostly relied on cry analysis. Several studies have been conducted for the analysis of the cries of newborn and small infants, healthy or with various diseases, but also of premature newborns. Four research groups largely contributed on infant cry analyses: the group of Wasz-Höckert in Helsinki (Finland), the group of Lester in Providence (USA), the group of Manfredi in Firenze (Italy) and the group of Reyes-Garcia in Tonantzintla (Mexico). If first studies concerned induced pain cries, spontaneous cries have also been considered.

This section is organized as follows. First, we present a non-exhaustive list of studies realized in the analysis of infant cries according to their clinical context. Then, we focus on the methods of data acquisition and acoustic signal processing encountered in these studies. In particular, as the vocabulary of acoustic analysis is very rich, a special attention is paid to feature definition. To a lesser extent, other sounds processing methods (pre-linguistic infant vocalizations, NICU alarms, EEG sonification and lung sound classification) have been proposed and are compiled at the end of this section.

3.1 Clinical applications

3.1.1 Full-term newborns, infants and toddlers

Infant cries were largely studied for the differentiation between normal and pathological cries. In [74], authors compared the cries of typically developing infants and infants possibly suffering from central nervous system insult due to malnutrition. They showed that the cry of the malnourished infant had an initial longer sound, higher pitch, lower amplitude, more arrhythmia, and a longer latency to the next cry sound than the cry of the well-nourished infant. The similarity between the cry of the malnourished infant and the cry of the brain-damaged infant suggested that malnutrition might affect the regulatory function of the central nervous system. Similar conclusions were obtained by a computerized analysis in [31]. It was shown that cries of healthy newborns with prenatal and perinatal complications have different acoustical properties than cries of low-complications newborns [155]. Abnormalities were searched in the cries of newborns with multiple or severe problems during the neonatal period, such as low birth weight, respiratory symptoms, jaundice, apnea, but also infants subsequently victims of presumed sudden infant death syndrome [46]. In the 2000s, normal and pathological cries began to be automatically labeled thanks to a wide variety of machine learning approaches in the context of deafness [97, 119, 132],

hypoxia-based Central Nervous System (CNS) diseases [98], cleft palate [72] and asphyxia [52, 119, 132].

Several studies relied on pain-induced cries. Fuller *et al.* differentiated them from fussy and hunger cries, as well as non-cry cooing vocalizations (pre-linguistic vocalizations occurring around 3 months of age) in 30 infants ranging in age from 2 to 6 months [40, 41]. This study was based on the amplitude of high-frequency components, the fundamental frequency and the spectral energy levels. Formants (vocal tract resonance frequencies) and tenseness were first added in [38] and acoustic cry measures were correlated with four pain levels and four ages (between 0 and 12 months) [39]. Analysis of pain-induced cries was combined [50] with a (manual) coding of the facial expressions (Neonatal Facial Coding System, NFCS) [49], in order to discriminate behavioral reactions between invasive and non-invasive procedures. The objective was to test if a newborn infant's cry could be used to measure pain, after heel-prick stimulus [120]. The analysis showed that the crying sound after the painful stimulus of the heel-prick had a significantly higher fundamental frequency and lasted longer at the first than at the fifth cry. However, while the first cry was more like a cry of pain, the fifth cry more resembled crying for other reasons. The conclusion was that crying could be used to measure pain in newborn infants only when the cause of crying was known. Different pain levels were also applied [14] and relations between cry characteristics and a pain score on the DAN (Douleur Aigüe du Nouveau-né, newborn acute pain) scale were evaluated. Results showed that a correlation existed when the DAN score was greater than 8 (on a 0 to 10 scale), and could be used as an alarm threshold.

Analysis of cries was also developed in other contexts. Characteristics of infant cries were correlated with perception and responses of adults, parents or non-parents in [47, 154, 156]. Acoustic characteristics of the cries of newborn of marijuana users and non-users were compared [76]. Differences observed between both suggested that heavy marijuana use affected the neurophysiological integrity of the infant. The separation distress call in the absence of maternal body contact was evaluated by quantifying the amount of crying during the first 90 minutes after birth when the baby was placed either skin-to-skin with the mother, in a cot, or first in a cot and then skin-to-skin [25].

Spontaneous cries were also processed: i) in the context of profound hearing loss and/or perinatal asphyxia [13, 43, 102, 103, 117, 118, 122, 132, 144, 146], ii) to find possible early signs of autism [96, 123] iii) in the context of monitoring [95], iv) to better understand vocal development and early communication [16, 148, 157]. Cries of hard-of-hearing and healthy infants were compared through duration, amplitude and melody (fundamental frequency fluctuations along a cry) description [141–143]. Recently, baby emotional cries have also been studied in order to be integrated in a robotic baby caregiver [150].

3.1.2 Preterm newborns

Characterization of the premature crying episodes and their differences with cries of full-term infants were also largely explored, in order to explain differences observed in their neurophysiological maturity, and the subsequent impact on their speech development.

First studies focused on the analysis of induced-pain cries. The influence of neurophysiological maturity on induced-pain cries was firstly deducted when full-term newborns cries spectral variability was found to be more complex than for preterm newborns [135]. Later, a comparison between cries

of premature and full-term newborns showed that the cries of smallest premature newborns compared with the controls were shorter, with higher fundamental frequency, and included bi-phonation and glide more often [44, 84]. However, the cry characteristics changed with increasing conceptual age and the older the child the more the cry pattern resembled that of the full-term [138]. Evaluation of pain from facial expression and crying was performed in premature infants, but also in full-term and 2- and 4-month-old infants, each group receiving a dedicated stimulus [58]. Results from this multivariate analysis showed that premature infants were different from older infants, that full-term newborns were different from others, but that 2- and 4-month-olds were similar. Later, Stevens *et al.* included two variables, severity of illness and behavioral state (sleep or awake) in the analysis [131]. Behavioral state was found to influence the facial action variables and severity of illness modified the acoustic cry variables. As for non-premature newborns, some of these works coupled the cry analysis with the facial activity coding NFCS [26, 58, 131].

Analysis of spontaneous cries of preterm infants has been recently investigated [80, 92, 94, 125, 149]. A comparison between spontaneous cries of six premature newborns (three pairs of twins) recorded at different ages (8th-9th week, 15th-17th week and 23rd-24th week) was performed [149]. Essential changes in the cries were observed from the 8th-9th week of life up to the 23rd-24th week of life, and were interpreted as an intentional articulatory activity. The distress during cry was also shown as correlated with central blood oxygenation [92]. Results indicated that a similar decrease in oxygenation level occurs in both groups of patients, but the recovery time after the crying episode was more stable and rapid in full-term newborns than in preterm ones. An acoustic analysis of cries before feeding in both healthy preterm infants at term-equivalent ages and full-term newborns was performed [125]. Effects of gestational age, body size at recording and IntraUterine Growth Retardation (IUGR) were then investigated, and showed that shorter gestational age was significantly associated with a higher fundamental frequency, although no relation was found with smaller body size at recording nor IUGR. Fundamental frequency and formants of preterm newborns were shown to be related to an increasing gestational age [80] and generally higher for full-term newborns [94].

3.2 Methods for acoustic signal processing

Acoustic signal processing in paediatrics dealt with different objectives: extraction of features, cry segmentation, cry classification and assessment of other sounds (Figure 2.2). This section covers all these topics and is supplemented by a paragraph dedicated to feature definition. But first, as for the video analysis section, a particular attention has been paid to audio acquisition methods.

3.2.1 Audio acquisition

Most of the studies have been conducted on real audio signals where microphones were placed from 10 to 30 centimeters of the infant's mouth. In case of pain-induced cry analysis, recordings have been performed from the stimulus to the end of the cry event [46, 50, 85, 143] whereas in other clinical contexts, no further details about recording period were usually reported. In fact, recordings were annotated by experts resulting into small cry signals (few seconds) database. Thus, all authors constructed their own cry signals database supported by specific clinical annotations.

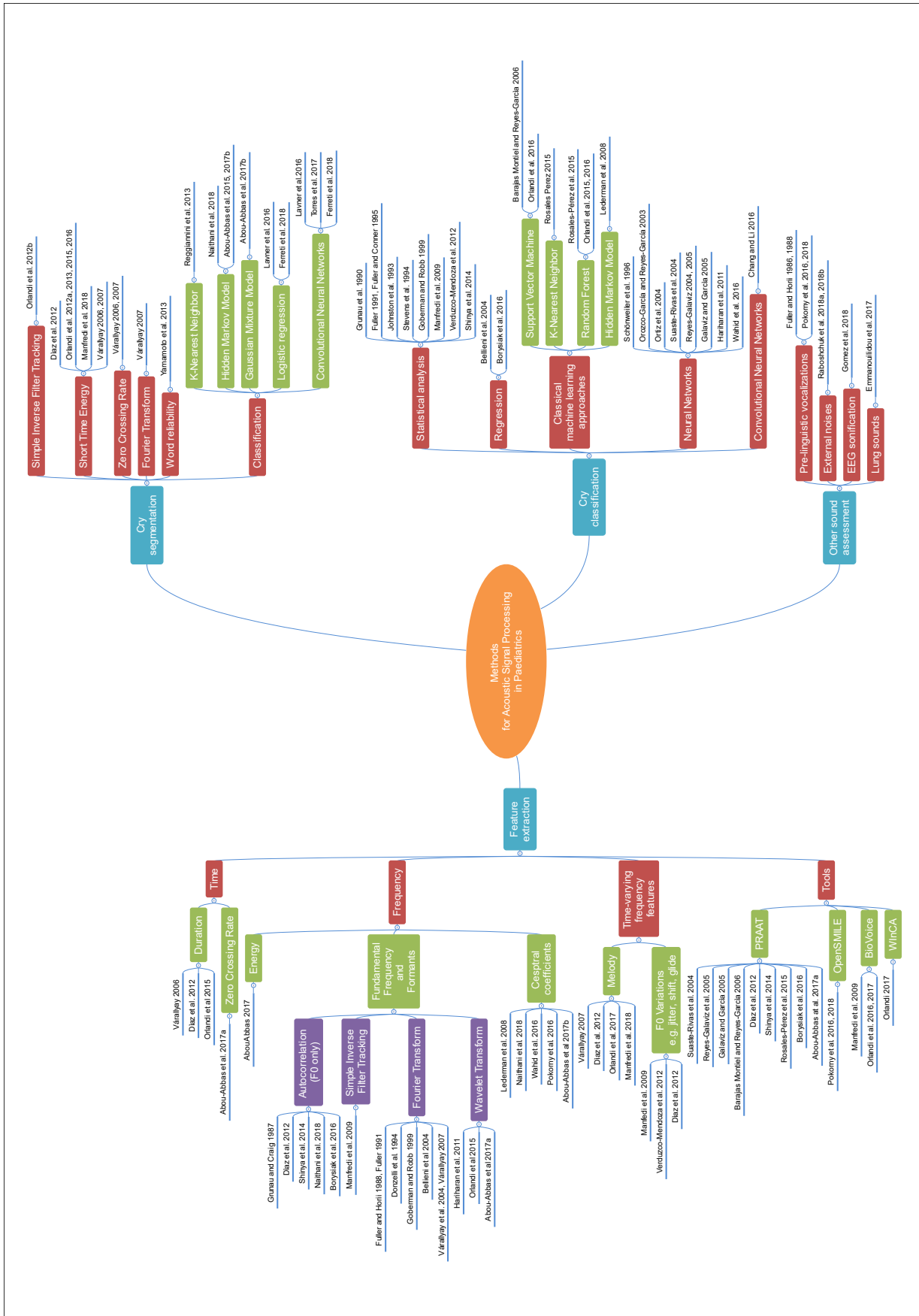


Figure 2.2 – Overview of acoustic signal processing methods in paediatrics.

Among them, we can cite the Baby Chillanto cry signals database, property of the Instituto Nacional de Astrofísica Óptica y Electrónica (Mexico). It was composed by five types of annotated cries (pain, hunger, normal, asphyxia and deaf) and on which relied the work of Reyes-García group [13, 29, 43, 52, 97, 98, 117–119, 132, 144, 146]. Cohort size varied from few infants [72, 150] to more than three hundred babies [142] or more than 120 premature babies in [131].

Several types of noise can affect the audio recordings (e.g. alarms, voice). Their occurrence directly depends on the acquisition environment. Authors usually did not mention where the recording took place except when it was performed at home [106, 142, 150]. However, we reported two types of acoustic environments: i) with low energy noise (e.g. background noise) [29, 79, 92–96, 116, 141, 142] and ii) with high energy noise (e.g. medical device sounds, adults' voices) [3, 5, 71, 88, 150].

Only Manfredi's group dealt with synthetic crying signals [37, 93], and recently, synthetic signals of the melody of cries have been also constructed [79, 91]. It was also proposed to create a simulated database to reproduce realistic acoustic scenarios [36]. Noisy conditions have been created by synthetically adding four noises (human speech, interfering cries, beep sounds and background noise) to a clean set created by convolving 64 infant cry recordings with synthetic impulse responses.

3.2.2 Feature definition

The analysis of cries mainly relied on the computation of features in two domains: time and frequency (see [70] for a review). In addition, several studies investigated the variations of the fundamental frequency along a cry event.

Time features The most common time features were naturally computed to characterize the duration of cries [25, 29, 31, 39, 44, 46, 50, 76, 85, 120, 131, 141]. Among them, we found the total time in crying [120, 141], the ratio between duration of cry and total audio segment duration [44, 141], the mean [31, 141] or the variation coefficient [31] of cry durations. Non-cry episodes were also examined through similar duration metrics [141]. In the case of induced-pain, latency time (interval between the painful stimulus and the first cry) has been considered [50, 85, 120, 131].

In addition, intensity, or loudness, features such as average amplitude [46] or amount of energy [4, 5] of cries have been proposed. Zero Crossing Rate (ZCR) has also been quantified to characterize cries [4].

Frequency features The first way to describe the spectral composition of a signal is to compute spectral energy features. Authors proposed different approaches such as the computation of the overall spectral energy of the signal [39, 58, 116, 131] or the energy only induced by low or high frequencies [38, 41, 44].

Fundamental frequency (F_0)¹ was investigated by virtually all the cited works. In fact, it is a direct measurement of vocal development since it corresponds to the rate of glottal opening and closing in the vocal tract [77]. F_0 has been characterized through different statistical features computed from several

1. The term "pitch" is also employed by some authors [29, 34, 37]. However, in speech production, fundamental frequency and pitch are not identical since pitch strictly refers to the perceptual quality of the frequency i.e. how our auditory system perceives different frequencies [75].

cries such as mean (e.g. [16, 85, 95]), but also maximum and minimum [125], standard deviation [94] or variation coefficient [31].

Formants, or resonance frequencies, are produced by the instantaneous shape of the vocal tract and have also been statically analyzed. Most of the studies that examined resonant frequencies were focused on the first two formants F1 and F2 [31, 38, 95, 149], but some authors also proposed to assess F3 [31, 80, 94].

However, the difficulties met to extract resonance frequencies led to the computation of other coefficients to describe the spectral envelop such as Mel Frequency Cepstral Coefficients (MFCCs) [5, 72, 106, 119] or Linear Prediction Cepstral Coefficients (LPCCs) [132, 146]. MFCCs are obtained through the projection of the signal on a mel-scale inspired by a psychoacoustic model of human auditory perception whereas LPCCs are computed from the modelling of the vocal tract.

Time-varying frequency features The most common time-varying frequency descriptor is the pattern of F0 over a cry, or melody shape. Four main melodic shapes have been defined by Schönweiler *et al.* [122]: falling, rising, falling-rising (or rising-falling) and flat. They were then reduced to three fundamental units (falling, rising and flat) and have been shown as the basis of 77 melodic shapes [142]. A fifth melody shape, called "complex shape", was also considered to cover all melodic patterns composed by more than two fundamental units [79, 91, 148].

Several other features have been defined to assess the F0 variations along a cry unit or during a cry event (succession of cry units). Among them, we can cite jitter (cycle-to-cycle variations of F0) [40], shift (sudden change in F0) [29, 85, 120], glide (rapid variations in F0) [29, 85, 120, 144], vibrato (series of waves with remarkable frequency variations) [84, 144] or glottal roll (phonation of weak intensity and low F0, below the normal average measurement of the tone) [84, 144].

3.2.3 Feature extraction

This section provides an overview of the extraction methods used to estimate acoustics features. They have been organized regarding the three types of feature targeted: time, frequency and time-varying frequency. In addition, a paragraph has been dedicated to tools that were developed and/or used by the different authors.

Time features Computation of duration features relied on a segmentation step in order to extract cry and non-cry epochs. It has been performed either manually [25, 31, 39, 44, 46, 50, 76, 85, 120, 131] or automatically [29, 95, 141] (cf. section 3.2.4). For their part, ZCR or energy were computed directly from the signal by windowing [4].

Frequency features Three types of methods have been employed to estimate frequency features: temporal, spectral and cepstral.

Regarding time domain methods, autocorrelation function has been largely used to estimate the fundamental frequency [16, 29, 49, 88, 125]. In practice, a window, encompassing several pitch periods (F0 was usually searched between 150 Hz and 1000 Hz), was used to calculate short-term autocorrelation

sequence and the fundamental frequency was obtained at the maximum of this sequence. Sometimes, it has been supported by correction of the F0 estimation tracking errors [16]. Manfredi *et al.* also proposed a tuned method of the Simple Inverse Filter Tracking (SIFT) algorithm [81] which gave better F0 estimation [80].

In early studies, frequency feature extraction was generally based on a spectrographic analysis [50, 85, 120], but more recent ones developed automatized estimations methods using Fourier Transform [14, 31, 38, 41, 44, 142, 143] or Wavelet Transform [4, 52, 95]. Most of the time, energy features were directly computed from spectrum and peak-picking procedures were implemented to extract F0 or resonance frequencies [14, 38]. The main limit of this approach was the presence of noise parts due to silence during crying episodes. To overcome this issue, several authors worked with Long Time Average Spectrum (LTAS) that was the average spectrum computed from all selected cry periods of interest (e.g. without pauses or silences) [31, 41, 44]. For their part, Varallyay *et al.* used Smoothed Spectrum Method (SSM). In fact, resulting spectrum only contained cry components thanks to an initial statistical processing removing noise parts induced by silence [142, 143]. Recently, Continuous Wavelet Transform (CWT) approaches, known for their robustness to noise, have also been considered to estimate F0 and formants [95] or to extract energy feature from different component levels [4, 52].

As mentioned in Section 3.2.2, cepstral coefficients (MFCCs and LPCCs) have also been extracted to describe audio signals [5, 72, 88, 106, 146]. MFCCs were traditionally computed in six steps: a) partitioning the signal into short frames, b) computing the power spectrum density, c) projecting the power spectra on mel-filter bank (simulation of human auditory perception) and sum the energy in each filter, d) taking the logarithm of all filter bank energies e) calculating the Discrete Cosinus Transform (DCT) of the log filter bank energies f) keeping DCT coefficients as MFCCs. However, Abou-Abbas *et al.* recently proposed to compute similar MFCCs features after applying Empirical Mode Decomposition (EMD) [4, 5] or by using Discrete Wavelet Transform (DWT) instead of DCT [5]. For their part, LPCCs were computed from the Smoothed Auto Regressive (AR) power spectrum where AR coefficients were estimated by Levinson-Durbin algorithm.

Time-varying frequency features Melody shapes have been identified by two methods based on the projection of F0 pattern into a grid: Five-Line Method (FLN) (fixed 5-lines grid from 330 Hz to 700 Hz) [142] and dodecagram (variable grid depending on the first F0 estimation of the cry) [29]. The dodecagram method showed better results due to its adaptability. Melody shapes have also been synthetically constructed by the mean of rules on F0 variations (e.g. the falling shape was obtained setting a maximum frequency of 650 Hz at 0.4 second, followed by a slow decrease towards 450 Hz) in order to compare tool abilities to estimate F0 patterns [91]. Recently, an approximation of melody shapes by quadratic and fourth order polynomial functions was added [79].

The other F0 variations features (e.g. jitter, shift, glide) were mainly defined by decision rules on the F0 contour, either visually noticed [84, 85] or automatically computed [29, 80, 144]. As an example, in [29], a shift was detected when a sudden change of ascends or descends in the F0 between 100 Hz and 600 Hz within 0.1 second occurred.

Tools Some software tools for the acoustic analysis of infant cries were developed and/or used. The most popular one is PRAAT, initially proposed for adult's voice by Boesrma in 2002 [15]. It was used in [4, 13, 16, 29, 43, 118, 119, 125, 132]. Acoustic parameters have also been extracted by the mean of the openSMILE toolkit [34, 105, 106]. Both softwares provided automatic computation of a wide variety of features (e.g. F0, formants, MFCCs, LPCCs, jitter, shimmer) but had to be manually tuned to give relevant analysis of infant cries.

For their part, Manfredi *et al.* developed BioVoice [80, 94] and WInCA [91], two softwares adapted to infant cry analysis, where different estimation methods of F0 (respectively, SIFT and wavelet) and resonance frequencies (respectively, peak picking in the power spectral density and wavelet) were implemented. When comparing the two approaches with PRAAT, where F0 was computed by autocorrelation method and formants were approached by LPCCs, authors found out that PRAAT gave better results to approximate fundamental frequency whereas BioVoice was more accurate for formant estimation [91].

3.2.4 Automatic cry segmentation

In the analysis of sound, one of the problems relies on the segmentation of the recordings into Voiced/UnVoiced (V/UV) parts, also called detection of Cry Units (CU), in order to extract relevant acoustic parts. If in a large majority of papers, the cry segments were manually selected, some recent studies, described below, proposed solutions to perform this segmentation automatically.

A V/UV detection procedure, based on SIFT, where an interval was selected as voiced if the maximum of the autocorrelation function is greater than a fixed threshold was proposed [96]. Methods based on thresholding the Short-Term Energy (STE) function were also investigated [29, 79, 92–95]. In [93], authors proposed to automatically compute the threshold using Otsu method [100]. Cry segmentation was also performed by combining two short-time methods: STE and ZCR [141]. In fact, STE provided a distinction between audible sounds and silence while ZCR permitted to detect V/UV parts. Then, a threshold was applied to extract CU. Later, authors added a third step to distinguish harmonic and non-harmonic audio segments [142]. It was based on the hypothesis that spectral structure of a crying segment is harmonic since it only contains the fundamental frequency and its harmonics. Nevertheless, no quantitative evaluation was performed. A marginal method has been proposed by Yamamoto *et al.* [150] where a baby cry was detected if at least one of the two following conditions was respected: the word reliability computed by the adult word detection tool "Julius" [73] was under a threshold or the change of the fundamental frequency for a time period was superior to another threshold. A detection accuracy of only 69.4% was obtained.

A few recent approaches considered the cry segmentation as a classification problem. Reggiannini *et al.* proposed an automatized classification in three classes: voiced part, unvoiced part and silence. By the use of KNN, they achieved to discriminate the three states with an Area Under Curve (AUC) of 0.88 [116]. Expiratory and inspiratory phases of the cries were separated in some studies. Hidden Markov Model (HMM) based approaches led to 83.79% of accuracy [3] when six classes (expiratory phases, inspiratory phases, noise, adult speech, silence and beeps) were considered. Later, these results were improved by formulating the problem with three classes: expiratory phases, inspiratory phases and others [5] or residuals [88] (a class regrouping all previous mentioned noisy sounds). HMM

and Gaussian Mixture Model (GMM) methods were compared and GMM gave the best results with a classification error rate of 8.9% [5]. A total accuracy of 89.2% was also reached with HMM in [88].

Newly, deep learning approaches have been considered to detect cries in a domestic environment [71, 139] or in NICU [36]. In all studies, log mel-frequency features were computed on windows and combined to construct the Convolution Neural Network (CNN) input layer. Such methods requiring large dataset, authors proposed to introduce normalization and regularization to adapt CNN to modest dataset [139] or to enhance dataset by adding simulated data [36]. CNN slightly outperformed classification methods in order to detect cries. In fact, CNN gave lower false-positive rate than logistic regression [71] and an AUC upper than 90% was reached [139]. In NICU, an average accuracy of 86.58% was obtained [36].

3.2.5 Automatic cry classification

Automated cry classification have been performed after a manual [13, 14, 16, 38, 39, 43, 44, 50, 52, 58, 72, 97, 98, 117–119, 125, 131, 132, 144, 146] or an automated [80, 94, 95] segmentation step. First studies used classical statistical approaches such as Student T-test [80, 144], ANOVA [38, 44, 50, 125] and MANOVA [39, 58, 131] or regression [14, 16]. It was applied either to compare infant ages [44, 58, 80, 125] and gender [16, 38], to evaluate pain [14, 39, 131] or to recognize pathologies [50, 144].

More recent papers investigated classification approaches using a high number of features. Different numbers of cry classes were considered according to the clinical target: two classes (normal vs abnormal [52, 72, 97, 98, 119] or preterm vs full-term [94, 95]), three classes (normal, hypo acoustic and asphyxia [13, 43, 117, 118, 132], hunger, pain and sleep [24] or hunger, pain and no-pain-no-hunger [13]) and five classes (pain, asphyxia, hunger, deaf and normal [146]). A wide variety of machine learning approaches has been evaluated, regrouping classical methods such as SVM [13, 94], KNN [119], Random Forest (RF) [94, 95, 119], HMM [72] or Neural Networks [43, 52, 97, 98, 117, 118, 122, 132, 146]. Classification results were efficient since some studies reached results above 95% of accuracy [13, 43, 52, 97, 117, 118].

Deep learning was also investigated to classify cries into three categories: hungry, pain and sleep [24]. Spectrogram of cries were computed by Fast Fourier Transform (FFT) and used as input layer of a CNN. The method showed promising results with 78.5% of accuracy.

3.2.6 Other sound assessment

Several recent audio processing methods have been proposed regarding non-cry signals and concerning either pre-linguistic vocalizations (including cooing) [40, 41, 105, 106]. Non-voice analyses were also proposed in different contexts such as external noise detection [111, 112], EEG sonification [99] or lung sound assessment [33].

Cooing, manually selected, were analyzed by Fuller *et al.* . in 30 infants ranging in age from 2 to 6 months, where significant differences were found in F0 and Mean Spectral Energy (MSE) with classical cries (fussy, hungry and pain) [40, 41]. Pre-linguistic vocalizations have also been studied in 7- to 12-month-old infants having received the diagnosis of a neuro-developmental disorder (autism, Rett syndrome, fragile X syndrome) by Pokorny *et al.* . They processed retrospective home video recordings

provided by the family and made during family events, before the disorder was diagnosed. In [106], a comparison between manual and automated segmentation of vocalizations, using machine learning (HMM, SVM and RF), led to an accuracy of only 38.0%, where errors came from confusions with parental voices or voices from television. Later, they evaluated more than 6000 features to differentiate typical and atypical early speech language of one infant with Rett syndrome. Main differences were observed in auditory attributes such as timbre (spectral envelope) and pitch [105].

For their part, Raboshchuk *et al.* focused on the automatic detection of alarms and external vocalizations (e.g. nurses, parents) in NICU. In [111], alarms were detected thanks to the knowledge of alarm characteristics (e.g. frequencies) integrated to a GMM classifier. In [112], several pre-processing approaches were tested: spectral subtraction, non-negative matrix factorization and combination of both. Best results were obtained with non-negative matrix factorization followed by spectral subtraction.

Recently, marginal purposes were investigated through audio processing for example to detect neonatal seizures from EEG or to detect lung sounds abnormalities. EEG signal was converted into an audible audio signal (process is called sonification) in order to hear relative frequency change when a seizure occurred [99]. It was shown that sonification methods perform similarly well, with a smaller inter-observer variability in comparison with visual interpretation. Lung audio recordings of 1000 children were also studied [33]. First, noise suppression techniques were applied to discard ambient sounds, sensors artifacts or crying. Notably, crying episodes were discarded by the use of SVM classifier trained with spectrotemporal features. Finally, normal and pathological lung sounds were classified through SVM classifier with an accuracy of 86.7%.

4 Discussion and Conclusion

This review showed that a lot of works have been published since several decades in the domains of video and audio processing in paediatrics.

The review of video processing showed that video recordings have been mainly exploited for motion analyses, in two major clinical contexts, general movement assessment and neonatal seizures detection and characterization. These studies have shown that the quality and quantity of movements are markers of the infant's neurological health. If recent improvements in digital video processing allowed an increasing automation, most of the above-mentioned studies need to be manually initialized in order to select the considered region (whole baby or limbs). Promising results were recently obtained with CNN but the method was only applied within a controlled environment setup [32]. Furthermore, most of the proposed methods extracted global motion information (i.e. without identifying each limb contribution). This may be improved by exploiting methods developed in adults using more precise body descriptors, such as kinematic or shape models [107].

Video-based respiration and heart rate estimation techniques showed interesting results. However, these techniques only work when the baby is not moving. Nowadays, the most valuable application in term of infant monitoring may be the automatic detection of apneas that generally do not occur when the baby is moving. Similarly, face analysis can't yet be integrated into a monitoring system since none of the proposed techniques are robust to occlusions that can happen in NICU, either from the baby itself or from external adult manipulations. Another limitation is related to the video recording duration and

the constrained acquisition setups. Indeed, for most of the studies, recordings only contained periods of interest and the infants were placed in some specific conditions in order to ensure an appropriate acquisition. Furthermore, infants were generally not covered, with appropriate lightening conditions and no external interventions. On the other hand, long video recordings may include medical staff or parents' presence in the image that had to be detected and eliminated to discard non-suitable periods. This problem has been recently addressed in [20]. Similarly, absences of the newborn in the bed will have to be detected to avoid the analysis of irrelevant periods. A recent study, based on motion analysis, showed encouraging results in that way [78].

The review of acoustic analyses shown that most of the studies was devoted to cries. Initially focused on pain-induced cries, more recent studies considered also spontaneous cries. Processing, most of the time based on frequency features, allowed to distinguish normal and pathological cries but also to classify different types of cries. They were also explored for premature newborns to identify differences in their neurophysiological maturity. A few papers dealing with the processing of pre-linguistic vocalizations, NICU alarms, EEG sonification and lung sound classification were also identified.

As for video, long audio recordings are parasitized by different sources such as nurse or parents' voices, alarms of monitoring devices, ventilation noise, etc. Although some authors worked on audio recordings performed in such environment, the automatic recognition of pathologic cries in NICU still remains difficult. In fact, only a few recent studies showed around 90% of accuracy in cry segmentation but no classification was proposed from there [5, 36, 88]. Additionally, baby sounds other than cries, like coughing but also vowel sounds, were slightly or not investigated. And yet, they are concomitant with many diseases [55] and may be an indicator of maturation [136] and of vocal development [21]. In fact, authors usually discarded them or included them with other sounds without making distinction (e.g. as expiratory sounds).

It is also important to notice that joint audio and video processing was not yet envisaged at this time. Only one study integrating audio and video processing, was, to our knowledge, published in [95], where a contactless system for Audio-Video Infant Monitoring (AVIM) was proposed. The analysis of movements was semi-automatic since the user had to select points to track on the video frame whereas cry analysis was performed automatically after a manual suppression of interfering sounds. Nevertheless, audio and video were processed separately. Interestingly, a combination of these two components could broaden the scope of applications in early clinical diagnosis of several pathologies. Moreover, it could be helpful in automatic sleep analysis, where both motion and baby sounds are important behavioral descriptors.

On the other hand, this review being dedicated to audio- and video-based systems, other acquisition systems have not been included. Briefly, the use of infrared thermography has been investigated to measure the skin temperature of newborns [2, 11, 54] or the respiratory rate and timing [45]. Another example is the use of depth cameras (e.g. Microsoft Kinect) to analyze infants' movements [82, 90].

Finally, if a lot of works have dealt and continue to deal with the processing of video and audio in paediatrics, a fully-automated efficient system does not exist. It will have to tackle above-mentioned difficulties by integrating robust processing methods to cope with unconstrained and long-term acquisition time such as encountered with monitoring systems in NICU. A part of these difficulties is addressed in the following chapters.

Bibliography

- [1] AARTS, L. A. M., JEANNE, V., CLEARY, J. P., LIEBER, C., NELSON, J. S., BAMBANG OETOMO, S., AND VERKRUYSE, W. Non-contact heart rate monitoring utilizing camera photoplethysmography in the neonatal intensive care unit - a pilot study. *Early Human Development* 89 (2013), 943–948.
- [2] ABBAS, A. K., AND LEONHARDT, S. Intelligent neonatal monitoring based on a virtual thermal sensor. *BMC Medical Imaging* 14 (2014), 9.
- [3] ABOU-ABBAS, L., ALAIE, H. F., AND TADJ, C. Automatic detection of the expiratory and inspiratory phases in newborn cry signals. *Biomedical Signal Processing and Control* 19 (2015), 35–43.
- [4] ABOU-ABBAS, L., TADJ, C., AND FERSAIE, H. A. A fully automated approach for baby cry signal segmentation and boundary detection of expiratory and inspiratory episodes. *The Journal of the Acoustical Society of America* 142, 3 (2017), 1318–1331.
- [5] ABOU-ABBAS, L., TADJ, C., GARGOUR, C., AND MONTAZERI, L. Expiratory and inspiratory cries detection using different signals' decomposition techniques. *Journal of Voice* 31, 2 (2017), 259.e13 – 259.e28.
- [6] ADDE, L., HELBOSTAD, J. L., JENSENIUS, A. R., TARALDSEN, G., GRUNEWALDT, K. H., AND STOEN, R. Early prediction of cerebral palsy by computer-based video analysis of general movements: A feasibility study. *Developmental Medicine & Child Neurology* 52, 8 (2010), 773–8.
- [7] ADDE, L., HELBOSTAD, J. L., JENSENIUS, A. R., TARALDSEN, G., AND STOEN, R. Using computer-based video analysis in the study of fidgety movements. *Early human development* 85, 9 (2009), 541–7.
- [8] ALINOVI, D., FERRARI, G., PISANI, F., AND RAHELI, R. Respiratory rate monitoring by video processing using local motion magnification. In *2018 26th European Signal Processing Conference (EUSIPCO)* (2018), IEEE, pp. 1780–1784.
- [9] ANDERS, T. F., EMDE, R. N., AND PARMELEE, A. H. *A manual of standardized terminology, techniques and criteria for scoring of states of sleep and wakefulness in newborn infants*. Los Angeles: UCLA Brain Information Service, NINDS Neurological Information Network, 1971.
- [10] ANDERS, T. F., AND SOSTEK, A. M. The use of time lapse video recording of sleep-wake behavior in human infants. *Psychophysiology* 13, 2 (1976), 155–8.
- [11] ANDERSON, E., WAILOO, M., AND PETERSEN, S. Use of thermographic imaging to study babies sleeping at home. *Archives of Disease in Childhood* 65, 11 (1990), 1266–1267.
- [12] ANTOGNOLI, L., MARCHIONNI, P., NOBILE, S., CARNIELLI, V., AND SCALISE, L. Assessment of cardio-respiratory rates by non-invasive measurement methods in hospitalized preterm neonates. In *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (2018), IEEE, pp. 1–5.

- [13] BARAJAS-MONTIEL, S. E., AND REYES-GARCÍA, C. A. Fuzzy support vector machines for automatic infant cry recognition. In *Intelligent Computing in Signal Processing and Pattern Recognition*, vol. 345. Springer, 2006, pp. 876–881.
- [14] BELLINI, C. V., SISTO, R., CORDELLI, D. M., AND BUONOCORE, G. Cry features reflect pain intensity in term newborns: An alarm threshold. *Pediatric research* 55, 1 (2004), 142–146.
- [15] BOERSMA, P. PRAAT, a system for doing phonetics by computer. *Glott international* 5, 9/10 (2002), 341–345.
- [16] BORYSIK, A., HESSE, V., WERMKE, P., HAIN, J., ROBB, M., AND WERMKE, K. Fundamental frequency of crying in two-month-old boys and girls: Do sex hormones during mini-puberty mediate differences? *Journal of Voice* 31, 1 (2016), 128.e21 – 128.e28.
- [17] BOS, A. F., MARTIJN, A., OKKEN, A., AND PRECHTL, H. F. R. Quality of general movements in preterm infants with transient periventricular echodensities. *Acta Paediatrica* 87, 3 (1998), 328–335.
- [18] BOS, A. F., MARTIJN, A., VAN ASPEREN, R. M., HADDERS-ALGRA, M., OKKEN, A., AND PRECHTL, H. F. Qualitative assessment of general movements in high-risk preterm infants with chronic lung disease requiring dexamethasone therapy. *The Journal of Pediatrics* 132, 2 (1998), 300–6.
- [19] CABON, S., POREE, F., SIMON, A., ROSEC, O., PLADYS, P., AND CARRAULT, G. Video and audio processing in paediatrics: a review. *Physiological Measurement* 40(2) (2019), 1–20.
- [20] CABON, S., POREE, F., SIMON, A., UGOLIN, M., ROSEC, O., CARRAULT, G., AND PLADYS, P. Motion estimation and characterization in premature newborns using long duration video recordings. *IRBM* 38, 4 (2017), 207–213.
- [21] CASKEY, M., STEPHENS, B., TUCKER, R., AND VOHR, B. Importance of parent talk on the development of preterm infant vocalizations. *Pediatrics* 128, 5 (2011), 910–916.
- [22] CATTANI, L., ALINOVI, D., FERRARI, G., RAHELI, R., PAVLIDIS, E., SPAGNOLI, C., AND PISANI, F. A wire-free, non-invasive, low-cost video processing-based approach to neonatal apnoea detection. In *2014 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings* (2014), pp. 67–73.
- [23] CATTANI, L., ALINOVI, D., FERRARI, G., RAHELI, R., PAVLIDIS, E., SPAGNOLI, C., AND PISANI, F. Monitoring infants by automatic video processing: A unified approach to motion analysis. *Computers in Biology and Medicine* 80 (2017), 158–165.
- [24] CHANG, C., AND LI, J. Application of deep learning for recognizing infant cries. In *2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)* (May 2016), pp. 1–2.
- [25] CHRISTENSSON, K., CABRERA, T., CHRISTENSSON, E., UVNAS-MOBERG, K., AND WINBERG, J. Separation distress call in the human neonate in the absence of maternal body contact. *Acta Paediatrica* 84, 5 (1995), 468–473.

-
- [26] CRAIG, K. D., WHITFIELD, M. F., GRUNAU, R. V., LINTON, J., AND HADJISTAVROPOULOS, H. D. Pain in the preterm neonate: Behavioural and physiological indices. *Pain* 52, 3 (1993), 287–299.
- [27] CUPPENS, K., LAGAE, L., CEULEMANS, B., VAN HUFFEL, S., AND VANRUMSTE, B. Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy. *Medical & Biological Engineering & Computing* 48, 9 (2010), 923–31.
- [28] CUPPENS, K., LAGAE, L., AND VANRUMSTE, B. Towards automatic detection of movement during sleep in pediatric patients with epilepsy by means of video recordings and the optical flow algorithm. *IFMBE Proceedings* 22 (2009), 784–789.
- [29] DÍAZ, M. A. R., GARCÍA, C. A. R., ROBLES, L. C. A., ALTAMIRANO, J. E. X., AND MENDOZA, A. V. Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis. *Biomedical Signal Processing and Control* 7, 1 (2012), 43–49.
- [30] DOGRA, D. P., MAJUMDAR, A. K., SURAL, S., MUKHERJEE, J., MUKHERJEE, S., AND SINGH, A. Toward automating hammersmith pulled-to-sit examination of infants using feature point based video object tracking. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2012), 38–47.
- [31] DONZELLI, G. P., RAPISARDI, G., MORONI, M., ZANI, S., TOMASINI, B., ISMAELLI, A., AND BRUSCAGLIONI, P. Computerized cry analysis in infants affected by severe protein energy malnutrition. *Acta Paediatrica* 83, 2 (1994), 204–11.
- [32] DOSSO, Y. S., BEKELE, A., NIZAMI, S., AUBERTIN, C., GREENWOOD, K., HARROLD, J., AND GREEN, J. R. Segmentation of patient images in the neonatal intensive care unit. In *2018 IEEE Life Sciences Conference (LSC)* (2018), IEEE, pp. 45–48.
- [33] EMMANOUILIDOU, D., MCCOLLUM, E. D., PARK, D. E., AND ELHILALI, M. Computerized lung sound screening for pediatric auscultation in noisy field environments. *IEEE Transactions on Biomedical Engineering* 65, 7 (2017), 1564–1574.
- [34] EYBEN, F., WENINGER, F., GROSS, F., AND SCHULLER, B. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia* (New York, NY, USA, 2013), MM '13, ACM, pp. 835–838.
- [35] FANG, C.-Y., HSIEH, H.-H., AND CHEN, S.-W. A vision-based infant respiratory frequency detection system. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on* (2015), IEEE, pp. 1–8.
- [36] FERRETTI, D., SEVERINI, M., PRINCIPI, E., CENCI, A., AND SQUARTINI, S. Infant cry detection in adverse acoustic environments by using deep neural networks. In *2018 26th European Signal Processing Conference (EUSIPCO)* (2018), European Signal Processing Conference, EUSIPCO, pp. 992–996.
- [37] FORT, A., AND MANFREDI, C. Acoustic analysis of newborn infant cry signals. *Med Eng Phys* 20, 6 (1998), 432–42.

- [38] FULLER, B. F. Acoustic discrimination of three types of infant cries. *Nursing Research* 40, 3 (1991), 156–160.
- [39] FULLER, B. F., AND CONNER, D. A. The effect of pain on infant behaviors. *Clinical Nursing Research* 4, 3 (1995), 253–273.
- [40] FULLER, B. F., AND HORII, Y. Differences in fundamental frequency, jitter, and shimmer among four types of infant vocalizations. *Journal of Communication Disorders* 19, 6 (1986), 441–447.
- [41] FULLER, B. F., AND HORII, Y. Spectral energy distribution in four types of infant vocalizations. *Journal of Communication Disorders* 21, 3 (1988), 251–61.
- [42] FULLER, P. W., WENNER, W. H., AND BLACKBURN, S. Comparison between time-lapse video recordings of behavior and polygraphic state determinations in premature infants. *Psychophysiology* 15, 6 (1978), 594–8.
- [43] GALAVIZ, O. F. R., AND GARCÍA, C. A. R. Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system. In *MICAI 2005: Advances in Artificial Intelligence* (2005), Springer Berlin Heidelberg, pp. 949–958.
- [44] GOBERMAN, A. M., AND ROBB, M. P. Acoustic examination of preterm and full-term infant cries: The long-time average spectrum. *Journal of Speech, Language, and Hearing Research* 42, 4 (1999), 850–61.
- [45] GOLDMAN, L. J. Nasal airflow and thoracoabdominal motion in children using infrared thermographic video processing. *Pediatric Pulmonology* 47, 5 (2012), 476–486.
- [46] GOLUB, H. L., AND CORWIN, M. J. Infant cry: A clue to diagnosis. *Pediatrics* 69, 2 (1982), 197–201.
- [47] GREEN, J. A., JONES, L. E., AND GUSTAFSON, G. E. Perception of cries by parents and nonparents: Relation to cry acoustics. *Developmental Psychology* 23, 3 (1987), 370.
- [48] GRIGG-DAMBERGER, M., GOZAL, D., MARCUS, C. L., QUAN, S. F., ROSEN, C. L., CHERVIN, R. D., WISE, M., PICCHIETTI, D. L., SHELDON, S. H., AND IBER, C. The visual scoring of sleep and arousal in infants and children. *Journal of Clinical Sleep Medicine* 3, 2 (2007), 201–240.
- [49] GRUNAU, R. V., AND CRAIG, K. D. Pain expression in neonates: Facial action and cry. *Pain* 28, 3 (1987), 395–410.
- [50] GRUNAU, R. V., JOHNSTON, C. C., AND CRAIG, K. D. Neonatal facial and cry responses to invasive and non-invasive procedures. *Pain* 42, 3 (1990), 295–305.
- [51] GUZZETTA, A., MERCURI, E., RAPISARDI, G., FERRARI, F., ROVERSI, M. F., COWAN, F., RUTHERFORD, M., PAOLICELLI, P. B., EINSPIELER, C., BOLDRINI, A., DUBOWITZ, L., PRECHTL, H. F., AND CIONI, G. General movements detect early signs of hemiplegia in term infants with neonatal cerebral infarction. *Neuropediatrics* 34, 2 (2003), 61–6.

- [52] HARIHARAN, M., YAACOB, S., AND AWANG, S. A. Pathological infant cry analysis using wavelet packet transform and probabilistic neural network. *Expert Systems with Applications* 38, 12 (2011), 15377–15382.
- [53] HAZELHOFF, L., HAN, J., BAMBANG-OETOMO, S., ET AL. Behavioral state detection of newborns based on facial expression analysis. In *International Conference on Advanced Concepts for Intelligent Vision Systems* (2009), Springer Berlin Heidelberg, pp. 698–709.
- [54] HEIMANN, K., JERGUS, K., ABBAS, A. K., HEUSSEN, N., LEONHARDT, S., AND ORLIKOWSKY, T. Infrared thermography for detailed registration of thermoregulation in premature infants. *Journal of Perinatal Medicine* 41, 5 (2013), 613–620.
- [55] HIRSCHBERG, J. Acoustic analysis of pathological cries, stridors and coughing sounds in infancy. *International Journal of Pediatric Otorhinolaryngology* 2, 4 (1980), 287–300.
- [56] HORN, B. K., AND SCHUNCK, B. G. Determining optical flow. *Artificial Intelligence* 17, 1-3 (1981), 185–203.
- [57] HUVANANDANA, J., THAMRIN, C., TRACY, M., HINDER, M., NGUYEN, C., AND MCEWAN, A. Advanced analyses of physiological signals in the neonatal intensive care unit. *Physiological Measurement* 38, 10 (2017), R253.
- [58] JOHNSTON, C. C., STEVENS, B., CRAIG, K. D., AND GRUNAU, R. V. Developmental changes in pain expression in premature, full-term, two-and four-month-old infants. *Pain* 52, 2 (1993), 201–208.
- [59] JORGE BRIEVA, E. M.-A. Phase-based motion magnification video for monitoring of vital signals using the hermite transform. vol. 10572, pp. 10572 – 10572 – 12.
- [60] KANESHI, Y., OHTA, H., MORIOKA, K., HAYASAKA, I., UZUKI, Y., AKIMOTO, T., MORIICHI, A., NAKAGAWA, M., OISHI, Y., WAKAMATSU, H., HONMA, N., SUMA, H., SAKASHITA, R., TSUJIMURA, S.-I., HIGUCHI, S., SHIMOKAWARA, M., CHO, K., AND MINAKAMI, H. Influence of light exposure at nighttime on sleep development and body growth of preterm infants. *Scientific Reports* 6 (2016), 21680.
- [61] KARAYIANNIS, N. B., SAMI, A., FROST, J., WISE, M. S., AND MIZRAHI, E. M. Quantifying motion in video recordings of neonatal seizures by feature trackers based on predictive block matching. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2004), vol. 1, pp. 1447–1450.
- [62] KARAYIANNIS, N. B., SAMI, A., FROST, J. D., WISE, M. S., AND MIZRAHI, E. M. Automated extraction of temporal motor activity signals from video recordings of neonatal seizures based on adaptive block matching. *IEEE Transactions on Biomedical Engineering* 52 (2005), 676–686.
- [63] KARAYIANNIS, N. B., SRINIVASAN, S., BHATTACHARYA, R., WISE, M. S., FROST, J. D., J., AND MIZRAHI, E. M. Extraction of motion strength and motor activity signals from video recordings of neonatal seizures. *IEEE Transactions on Medical Imaging* 20, 9 (2001), 965–80.

- [64] KARAYIANNIS, N. B., AND TAO, G. An improved procedure for the extraction of temporal motion strength signals from video recordings of neonatal seizures. *Image and Vision Computing* 24, 1 (2006), 27–40.
- [65] KARAYIANNIS, N. B., TAO, G., FROST, J. D., J., WISE, M. S., HRACHOVY, R. A., AND MIZRAHI, E. M. Automated detection of videotaped neonatal seizures based on motion segmentation methods. *Clinical Neurophysiology* 117, 7 (2006), 1585–94.
- [66] KARAYIANNIS, N. B., TAO, G., XIONG, Y., SAMI, A., VARUGHESE, B., FROST, J. D., J., WISE, M. S., AND MIZRAHI, E. M. Computerized motion analysis of videotaped neonatal seizures of epileptic origin. *Epilepsia* 46, 6 (2005), 901–17.
- [67] KARAYIANNIS, N. B., VARUGHESE, B., TAO, G., FROST, J. D., J., WISE, M. S., AND MIZRAHI, E. M. Quantifying motion in video recordings of neonatal seizures by regularized optical flow methods. *IEEE Transactions on Image Processing* 14, 7 (2005), 890–903.
- [68] KARAYIANNIS, N. B., XIONG, Y., FROST, J. D., J., WISE, M. S., AND MIZRAHI, E. M. Quantifying motion in video recordings of neonatal seizures by robust motion trackers based on block motion models. *IEEE Transactions on Biomedical Engineering* 52, 6 (2005), 1065–77.
- [69] KOOLEN, N., DECROUPET, O., DEREYMAEKER, A., JANSEN, K., VERVISCH, J., MATIC, V., VANRUMSTE, B., NAULAERS, G., HUFFEL, S. V., AND VOS, M. D. Automated respiration detection from neonatal video data. pp. 164–169.
- [70] LAGASSE, L. L., NEAL, A. R., AND LESTER, B. M. Assessment of infant cry: Acoustic cry analysis and parental perception. *Mental Retardation and Developmental Disabilities Research Reviews* 11, 1 (2005), 83–93.
- [71] LAVNER, Y., COHEN, R., RUINSKIY, D., AND IJZERMAN, H. Baby cry detection in domestic environment using deep learning. In *2016 ICSEE International Conference on the Science of Electrical Engineering* (2016), pp. 1–5.
- [72] LEDERMAN, D., ZMORA, E., HAUSCHILDT, S., STELLZIG-EISENHAEUER, A., AND WERMKE, K. Classification of cries of infants with cleft-palate using parallel hidden markov models. *Medical & Biological Engineering & Computing* 46, 10 (2008), 965–975.
- [73] LEE, A., KAWAHARA, T., AND SHIKANO, K. Julius—an open source real-time large vocabulary recognition engine. 1691–1694.
- [74] LESTER, B. M. Spectrum analysis of the cry sounds of well-nourished and malnourished infants. *Child Development* 47 (1976), 237–241.
- [75] LESTER, B. M., CORWIN, M. J., SEPKOSKI, C., SEIFER, R., PEUCKER, M., MCLAUGHLIN, S., AND GOLUB, H. L. Neurobehavioral syndromes in cocaine-exposed newborn infants. *Child development* 62, 4 (1991), 694–705.

-
- [76] LESTER, B. M., AND DREHER, M. Effects of marijuana use during pregnancy on newborn cry. *Child Development* 60 (1989), 765–771.
- [77] LESTER, B. M., AND ZESKIND, P. S. *A Biobehavioral Perspective on Crying in Early Infancy*. Springer US, Boston, MA, 1982.
- [78] LONG, X., VAN DER SANDEN, E., PREVОО, Y., TEN HOOR, L., DEN BOER, S., GELISSEN, J., OTTE, R., AND ZWARTKRUIS-PELGRIM, E. An efficient heuristic method for infant in/out of bed detection using video-derived motion estimates. *Biomedical Physics & Engineering Express* 4, 3 (2018), 035035.
- [79] MANFREDI, C., BANDINI, A., MELINO, D., VIELLEVOYE, R., KALENGA, M., AND ORLANDI, S. Automated detection and classification of basic shapes of newborn cry melody. *Biomedical Signal Processing and Control* 45 (2018), 174–181.
- [80] MANFREDI, C., BOCCHI, L., ORLANDI, S., SPACCATERRA, L., AND DONZELLI, G. P. High-resolution cry analysis in preterm newborn infants. *Medical Engineering & Physics* 31, 5 (2009), 528–32.
- [81] MARKEL, J. The SIFT algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics* 20, 5 (1972), 367–377.
- [82] MARSCHIK, P. B., POKORNY, F. B., PEHARZ, R., ZHANG, D., O’MUIRCHEARTAIGH, J., ROEYERS, H., BÖLTE, S., SPITTLE, A. J., URLESBERGER, B., SCHULLER, B., ET AL. A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders. *Current Neurology and Neuroscience Reports* 17, 43 (2017), 1–15.
- [83] MAZZONE, L., MUGNO, D., AND MAZZONE, D. The general movements in children with down syndrome. *Early Human Development* 79, 2 (2004), 119–30.
- [84] MICHELSSON, K., JÄRVENPÄÄ, A., AND RINNE, A. Sound spectrographic analysis of pain cry in preterm infants. *Early Human Development* 8, 2 (1983), 141–149.
- [85] MICHELSSON, K., AND MICHELSSON, O. Phonation in the newborn, infant cry. *International Journal of Pediatric Otorhinolaryngology* 49 Suppl 1 (1999), S297–301.
- [86] MIZRAHI, E. M., AND KELLAWAY, P. Characterization and classification of neonatal seizures. *Neurology* 37, 12 (1987), 1837–44.
- [87] MORIELLI, A., LADAN, S., DUCHARME, F. M., AND BROUILLETTE, R. T. Can sleep and wakefulness be distinguished in children by cardiorespiratory and videotape recordings? *Chest* 109, 3 (1996), 680–7.
- [88] NAITHANI, G., KIVINUMMI, J., VIRTANEN, T., TAMMELA, O., PELTOLA, M. J., AND LEPPÄNEN, J. M. Automatic segmentation of infant cry signals using hidden Markov models. *EURASIP Journal on Audio, Speech, and Music Processing* 2018, 1 (2018), 1–14.

- [89] NTONFO, G. M. K., FERRARI, G., RAHELI, R., AND PISANI, F. Low-complexity image processing for real-time detection of neonatal clonic seizures. *IEEE Transactions on Information Technology in Biomedicine* 16, 3 (2012), 375–382.
- [90] OLSEN, M. D., HERSKIND, A., NIELSEN, J. B., AND PAULSEN, R. R. Model-based motion tracking of infants. In *European Conference on Computer Vision* (2014), Springer, pp. 673–685.
- [91] ORLANDI, S., BANDINI, A., FIASCHI, F., AND MANFREDI, C. Testing software tools for newborn cry analysis using synthetic signals. *Biomedical Signal Processing and Control* 37 (2017), 16–22.
- [92] ORLANDI, S., BOCCHI, L., DONZELLI, G., AND MANFREDI, C. Central blood oxygen saturation vs crying in preterm newborns. *Biomedical Signal Processing and Control* 7, 1 (2012), 88–92.
- [93] ORLANDI, S., DEJONCKERE, P. H., SCHOENTGEN, J., LEBACQ, J., RRUQJA, N., AND MANFREDI, C. Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring. *Biomedical Signal Processing and Control* 8, 6 (2013), 799–810.
- [94] ORLANDI, S., GARCIA, C. A. R., BANDINI, A., DONZELLI, G., AND MANFREDI, C. Application of pattern recognition techniques to the classification of full-term and preterm infant cry. *Journal of Voice* 30, 6 (2016), 656–663.
- [95] ORLANDI, S., GUZZETTA, A., BANDINI, A., BELMONTI, V., BARBAGALLO, S. D., TEALDI, G., MAZZOTTI, S., SCATTONI, M. L., AND MANFREDI, C. AVIM - A contactless system for infant data acquisition and analysis: Software architecture and first results. *Biomedical Signal Processing and Control* 20 (2015), 85–99.
- [96] ORLANDI, S., MANFREDI, C., BOCCHI, L., AND SCATTONI, M. Automatic newborn cry analysis: A non-invasive tool to help autism early diagnosis. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (2012), IEEE, pp. 2953–2956.
- [97] OROZCO-GARCÍA, J., AND REYES-GARCÍA, C. A. A study on the recognition of patterns of infant cry for the identification of deafness in just born babies with neural networks. In *Iberoamerican Congress on Pattern Recognition* (2003), Springer, pp. 342–349.
- [98] ORTIZ, S. D. C., BECEIRO, D. I. E., AND EKKEL, T. A radial basis function network oriented for infant cry classification. In *Iberoamerican Congress on Pattern Recognition* (2004), Springer, pp. 374–380.
- [99] O’SULLIVAN, M., GOMEZ, S., O’SHEA, A., SALGADO, E., HUILLCA, K., MATHIESON, S., BOYLAN, G., POPOVICI, E., AND TEMKO, A. Neonatal eeg interpretation and decision support framework for mobile platforms. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (July 2018), pp. 4881–4884.
- [100] OTSU, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9, 1 (1979), 62–66.

- [101] PAMULA, Y., CAMPBELL, A., COUSSENS, S., DAVEY, M., GRIFFITHS, M., MARTIN, J., MAUL, J., NIXON, G., SAYERS, R., TENG, A., ET AL. ASTA/ASA addendum to the AASM guidelines for the recording and scoring of paediatric sleep. In *Journal of Sleep Research* (2011), vol. 20, Wiley-Blackwell Publishing, pp. 4–4.
- [102] PEARCE, S., AND TAYLOR, B. Energy distribution in the spectrograms of the cries of normal and birth asphyxiated infants. *Physiological Measurement* 14 (1993), 263–68.
- [103] PEARCE, S., AND TAYLOR, B. Time-frequency analysis of infant cry: Measures that identify individuals. *Physiological Measurement* 14 (1993), 253–62.
- [104] PEDIADITIS, M., TSIKNAKIS, M., AND LEITGEB, N. Vision-based motion detection, analysis and recognition of epileptic seizures—a systematic review. *Computer Methods and Programs in Biomedicine* 108, 3 (2012), 1133–48.
- [105] POKORNY, F. B., BARTL-POKORNY, K. D., EINSPIELER, C., ZHANG, D., VOLLMANN, R., BÖLTE, S., GUGATSCHKA, M., SCHULLER, B. W., AND MARSCHIK, P. B. Typical vs. atypical: Combining auditory Gestalt perception and acoustic analysis of early vocalisations in Rett syndrome. *Research in Developmental Disabilities* 82 (2018), 109–119.
- [106] POKORNY, F. B., PEHARZ, R., ROTH, W., ZÖHRER, M., PERNKOPF, F., MARSCHIK, P. B., AND SCHULLER, B. W. Manual versus automated: The challenging routine of infant vocalisation segmentation in home videos to study neuro (mal) development. In *Interspeech* (2016), pp. 2997–3001.
- [107] POPPE, R. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding* 108 (2007), 4–18.
- [108] PORÉE, F., SIMON, A., CABON, S., COROLLEUR, A., NARDI, N., PLADYS, P., AND CARRAULT, G. Traitement de vidéos de polysomnographie pour l'estimation de l'état des yeux chez le nouveau-né prématuré. In *XXVe Colloque GRETSI* (2015), pp. 1–4.
- [109] PRECHTL, H. F. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Human Development* 23, 3 (1990), 151–8.
- [110] PRECHTL, H. F., EINSPIELER, C., CIONI, G., BOS, A. F., FERRARI, F., AND SONTHEIMER, D. An early marker for neurological deficits after perinatal brain lesions. *Lancet* 349, 9062 (1997), 1361–3.
- [111] RABOSHCHUK, G., NADEU, C., JANČOVIČ, P., LILJA, A. P., KÖKÜER, M., MAHAMUD, B. M., AND DE VECIANA, A. R. A knowledge-based approach to automatic detection of equipment alarm sounds in a neonatal intensive care unit environment. *IEEE journal of Translational Engineering in Health and Medicine* 6 (2018), 1–10.
- [112] RABOSHCHUK, G., NADEU, C., PINTO, S. V., FORNELLS, O. R., MAHAMUD, B. M., AND DE VECIANA, A. R. Pre-processing techniques for improved detection of vocalization sounds in a neonatal intensive care unit. *Biomedical Signal Processing and Control* 39 (2018), 390–395.

- [113] RAHMATI, H., AAMO, O. M., STAVDAHL, Ø., DRAGON, R., AND ADDE, L. Video-based early cerebral palsy prediction using motion segmentation. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE* (2014), IEEE, pp. 3779–3783.
- [114] RAHMATI, H., DRAGON, R., AAMO, O. M., ADDE, L., STAVDAHL, Ø., AND VAN GOOL, L. Weakly supervised motion segmentation with particle matching. *Computer Vision and Image Understanding* 140 (2015), 30–42.
- [115] RECHTSCHAFFEN, A., AND KALES, A. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Los Angeles: UCLA Brain Information Service/Brain Research Institute, 1968.
- [116] REGGIANNINI, B., SHEINKOPF, S. J., SILVERMAN, H. F., LI, X., AND LESTER, B. M. A flexible analysis tool for the quantitative acoustic assessment of infant cry. *Journal of Speech, Language, and Hearing Research* 56, 5 (2013), 1416–1428.
- [117] REYES-GALAVIZ, O. F., TIRADO, E. A., AND REYES-GARCIA, C. A. Classification of infant crying to identify pathologies in recently born babies with anfis. In *International Conference on Computers for Handicapped Persons* (2004), Springer, pp. 408–415.
- [118] REYES-GALAVIZ, O. F., VERDUZCO, A., ARCH-TIRADO, E., AND REYES-GARCÍA, C. A. Analysis of an infant cry recognizer for the early identification of pathologies. In *Nonlinear Speech Modeling and Applications*. Springer, 2005, pp. 404–409.
- [119] ROSALES-PÉREZ, A., REYES-GARCÍA, C. A., GONZALEZ, J. A., REYES-GALAVIZ, O. F., ESCALANTE, H. J., AND ORLANDI, S. Classifying infant cry patterns by the genetic selection of a fuzzy model. *Biomedical Signal Processing and Control* 17 (2015), 38–46.
- [120] RUNEFORS, P., ARNBJÖRNSSON, E., ELANDER, G., AND MICHELSSON, K. Newborn infants' cry after heel-prick: Analysis with sound spectrogram. *Acta Paediatrica* 89, 1 (2000), 68–72.
- [121] SAMI, A., KARAYIANNIS, N. B., FROST, J. D., WISE, M. S., AND MIZRAHI, E. M. Automated tracking of multiple body parts in video recordings of neonatal seizures. *Building* (2004), 312–315.
- [122] SCHÖNWEILER, R., KAESE, S., MÖLLER, S., RINSCHIED, A., AND PTOK, M. Neuronal networks and self-organizing maps: New computer techniques in the acoustic evaluation of the infant cry. *International Journal of Pediatric Otorhinolaryngology* 38, 1 (1996), 1–11.
- [123] SHEINKOPF, S. J., IVERSON, J. M., RINALDI, M. L., AND LESTER, B. M. Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder. *Autism Research* 5, 5 (2012), 331–339.
- [124] SHIMIZU, A., ISHII, A., AND OKADA, S. Monitoring preterm infants' body movement to improve developmental care for their health. In *Consumer Electronics (GCCE), 2017 IEEE 6th Global Conference on* (2017), IEEE, pp. 1–5.

-
- [125] SHINYA, Y., KAWAI, M., NIWA, F., AND MYOWA-YAMAKOSHI, M. Preterm birth is associated with an increased fundamental frequency of spontaneous crying in human infants at term-equivalent age. *Biology Letters* 10, 8 (2014).
- [126] SIKDAR, A., BEHERA, S. K., DOGRA, D. P., AND BHASKAR, H. Contactless vision-based pulse rate detection of infants under neurological examinations. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015), IEEE, pp. 650–653.
- [127] SIVAN, Y., KORNECKI, A., AND SCHONFELD, T. Screening obstructive sleep apnoea syndrome by home videotape recording in children. *European Respiratory Journal* 9, 10 (1996), 2127–31.
- [128] SO, K., BUCKLEY, P., ADAMSON, T. M., AND HORNE, R. S. C. Actigraphy correctly predicts sleep behavior in infants who are younger than six months, when compared with polysomnography. *Pediatric Research* 58 (2005), 761–765.
- [129] SPITTLE, A. J., BROWN, N. C., DOYLE, L. W., BOYD, R. N., HUNT, R. W., BEAR, M., AND INDER, T. E. Quality of general movements is related to white matter pathology in very preterm infants. *Pediatrics* 121, 5 (2008), e1184–9.
- [130] STAHL, A., SCHELLEWALD, C., STAVDAHL, O., AAMO, O. M., ADDE, L., AND KIRKEROD, H. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20, 4 (2012), 605–14.
- [131] STEVENS, B. J., JOHNSTON, C. C., AND HORTON, L. Factors that influence the behavioral pain responses of premature infants. *Pain* 59, 1 (1994), 101–9.
- [132] SUASTE-RIVAS, I., REYES-GALAVIZ, O. F., DIAZ-MENDEZ, A., AND REYES-GARCIA, C. A. A fuzzy relational neural network for pattern classification. In *Iberoamerican Congress on Pattern Recognition* (2004), Springer, pp. 358–365.
- [133] SUN, D., ROTH, S., AND BLACK, M. J. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision* 106, 2 (2014), 115–137.
- [134] SUNG, M., ADAMSON, T. M., AND HORNE, R. S. C. Validation of actigraphy for determining sleep and wake in preterm infants. *Acta Paediatrica* 98 (2009), 52–57.
- [135] TENOLD, J. L., CROWELL, D. H., JONES, R. H., DANIEL, T. H., MCPHERSON, D. F., AND POPPER, A. N. Cepstral and stationarity analyses of full-term and premature infants' cries. *The Journal of the Acoustical Society of America* 56, 3 (1974), 975–80.
- [136] THACH, B. T. Maturation of cough and other reflexes that protect the fetal and neonatal airway. *Pulmonary Pharmacology & Therapeutics* 20, 4 (2007), 365–370.
- [137] THARP, B. R. Neonatal seizures and syndromes. *Epilepsia* 43 Suppl 3 (2002), 2–10.

- [138] THODÉN, C.-J., JÄRVENPÄÄ, A.-L., AND MICHELSSON, K. Sound spectrographic cry analysis of pain cry in prematures. In *Infant Crying*. Springer, 1985, pp. 105–117.
- [139] TORRES, R., BATTAGLINO, D., AND LEPAULOUX, L. Baby cry sound detection: A comparison of hand crafted features and deep learning approach. In *International Conference on Engineering Applications of Neural Networks* (2017), Springer, pp. 168–179.
- [140] VAN GASTEL, M., BALMAEKERS, B., OETOMO, S. B., AND VERKRUYSSE, W. Near-continuous non-contact cardiac pulse monitoring in a neonatal intensive care unit in near darkness. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics* (2018), vol. 1050114, International Society for Optics and Photonics, pp. 1–9.
- [141] VÁRALLYAY, G. Future prospects of the application of the infant cry in the medicine. *Periodica Polytechnica Electrical Engineering* 50, 1-2 (2006), 47–62.
- [142] VÁRALLYAY, G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology* 71, 11 (2007), 1699–1708.
- [143] VÁRALLYAY, G., BENYÓ, Z., ILLÉNYI, A., FARKAS, Z., AND KOVÁCS, L. Acoustic analysis of the infant cry: Classical and new methods. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE* (2004), vol. 1, IEEE, pp. 313–316.
- [144] VERDUZCO-MENDOZA, A., ARCH-TIRADO, E., REYES-GARCÍA, C. A., LEYBON-IBARRA, J., AND LICONA-BONILLA, J. Spectrographic cry analysis in newborns with profound hearing loss and perinatal high-risk newborns. *Cirugía y Cirujanos* 80, 1 (2012), 3–10.
- [145] VILLARROEL, M., GUAZZI, A., JORGE, J., DAVIS, S., WATKINSON, P., GREEN, G., SHENVI, A., MCCORMICK, K., AND TARASSENKO, L. Continuous non-contact vital sign monitoring in neonatal intensive care unit. *Healthcare Technology Letters* 1, 3 (2014), 87–91.
- [146] WAHID, N., SAAD, P., AND HARIHARAN, M. Automatic infant cry pattern classification for a multi-class problem. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 8, 9 (2016), 45–52.
- [147] WASZ-HÖCKERT, O., MICHELSSON, K., AND LIND, J. Twenty-five years of Scandinavian cry research. In *Infant Crying*. Springer, 1985, pp. 83–104.
- [148] WERMKE, K., AND MENDE, W. Musical elements in human infants' cries: In the beginning is the melody. *Musicae Scientiae* 13, 2_suppl (2009), 151–175.
- [149] WERMKE, K., MENDE, W., MANFREDI, C., AND BRUSCAGLIONI, P. Developmental aspects of infant's cry melody and formants. *Medical Engineering & Physics* 24, 7-8 (2002), 501–14.
- [150] YAMAMOTO, S., YOSHITOMI, Y., TABUSE, M., KUSHIDA, K., AND ASADA, T. Recognition of a baby's emotional cry towards robotics baby caregiver. *International Journal of Advanced Robotic Systems* 10, 2 (2013), 86.

- [151] ZAKER, N., MAHOOR, M. H., MATTSON, W. I., MESSINGER, D. S., AND COHN, J. F. A comparison of alternative classifiers for detecting occurrence and intensity in spontaneous facial expression of infants with their mothers. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013 12* (2013).
- [152] ZAKER, N., MAHOOR, M. H., MESSINGER, D. S., AND COHN, J. F. Jointly detecting infants' multiple facial action units expressed during spontaneous face-to-face communication. *2014 IEEE International Conference on Image Processing (ICIP) 80208* (2014), 1357–1360.
- [153] ZAMZAMI, G., RUIZ, G., GOLDFOF, D., KASTURI, R., SUN, Y., AND ASHMEADE, T. Pain assessment in infants: Towards spotting pain expression based on infants' facial strain. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on* (2015), vol. 5, IEEE, pp. 1–5.
- [154] ZESKIND, P. S. Adult responses to cries of low and high risk infants. *Infant Behavior and Development 3* (1980), 167–177.
- [155] ZESKIND, P. S., AND LESTER, B. M. Acoustic features and auditory perceptions of the cries of newborns with prenatal and perinatal complications. *Child Development 49* (1978), 580–589.
- [156] ZESKIND, P. S., AND MARSHALL, T. R. The relation between variations in pitch and maternal perceptions of infant crying. *Child Development 59* (1988), 193–196.
- [157] ZESKIND, P. S., PARKER-PRICE, S., AND BARR, R. G. Rhythmic organization of the sound of infant crying. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology 26*, 6 (1993), 321–333.

METHODOLOGICAL CONTEXT: METHODS FOR CLASSIFICATION

In many fields, research teams aim to model data for purposes going from better understanding of our world to prediction of the future. Although, historically, these problems were only tackled through statistical modeling, in the last decade, machine learning gained popularity. Indeed, machine learning is the branch of computer science that uses past experiences to take future decisions without a complete knowledge of all influencing elements [2]. Machine learning techniques can be divided in three main categories: supervised learning, unsupervised learning and reinforcement learning. In supervised approaches, such as classification or regression, the relationships between data and a targeted output are taught beforehand whereas for unsupervised techniques (e.g., clustering) relations and hidden patterns in data are independently found. Afterwards, both approaches can be included in reinforcement learning where the model will continue to learn from environment feedback.

The world of machine learning is wide and is still in expansion. It would be difficult to go through all underlying concepts of artificial intelligence, and thus, this chapter mainly focuses on one aspect of machine learning: supervised learning for classification.

Indeed, we saw in Sections 2.2.3 and 3.2.5 of Chapter 2, that classification is a valuable tool to clinically characterize newborn conditions either to recognize motion patterns or to classify cries. During this thesis, methods for classification were applied in order to tackle three problems: sleep stages classification, motion segmentation and cry extraction.

1 Problem formulation

The term "classification" covers all techniques that classify data into a given number of classes. Being part of the supervised branch of machine learning, classification requires a learning phase on labeled data to identify in which class a new data sample belongs to [16]. Therefore, for classification problems, data has to be formed by the pair (\mathbf{X}, Y) ¹ with the data description matrix \mathbf{X} of size $N \times p$ and the set of labels $Y \in \{c_1, \dots, c_k\}$, where k is the number of targeted classes, N is the number of labeled samples and p is the number of features which characterize each sample. More precisely, each sample i of the data is described through the feature set $X_i = [x_{i1}, x_{i2} \dots, x_{ip}]$. This is depicted and illustrated by an example in Figure 3.1.

1. Matrices are noted in bold capital letters and vectors with capital letters

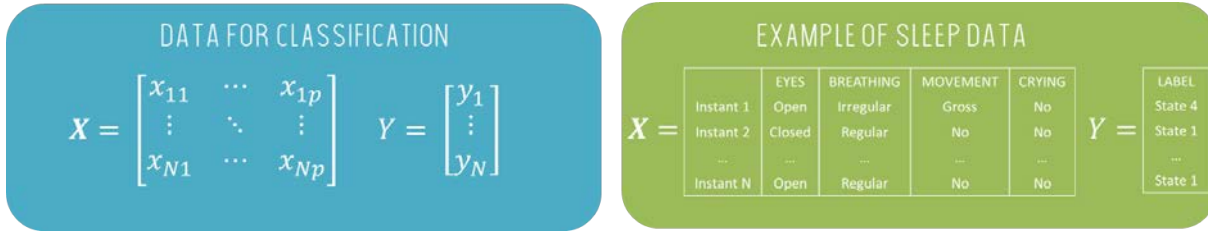


Figure 3.1 – Data for classification summary (left), supplemented by a sleep state dataset example (right).

Visually, each line of X contains the feature set of dimension p that describe a sample. A sample is associated with an output class which is reported in Y at the corresponding line. Hence, in the example, sleep data is formed by the pair (X, Y) , where X is composed by N instants. Four qualitative features are used to describe each instant (= sample): Eyes, Breathing, Movement and Crying ($p = 4$). Associated sleep states are contained in Y . Sleep states are labeled from 1 to 5 ($k = 5$), as defined in Chapter 1. In the sleep dataset example, only qualitative features (i.e., descriptive) are presented although quantitative features (i.e., numeric) can also be integrated into X .

To construct an accurate classification model, the process, depicted by Figure 3.2 and described below, is generally followed [7]. It implies five steps: collection of the data, feature engineering, learning, testing and deployment.



Figure 3.2 – Overview of the classification process from data collection to deployment.

In fact, it is necessary to keep in mind these steps to avoid two major problems of machine learning: underfitting and overfitting. Underfitting characterizes a model which fails to generalize the data, usually, due to a lack of training samples. Reversely, overfitting occurs when a classifier corresponds too closely or exactly to a particular set of data and fails to fit future data.

Collection of data The data collection is an important step to construct an accurate classification model. The more various are the training samples, the better will be the classification when deploying the model. It is important to notice that data used to train are the only knowledge of the model and thus, if a given event is too much represented in the data, the model may overfit.

Feature engineering One of the main challenges of classification is to provide an informative set of features regarding the targeted outputs of the model. In some cases, data may have to be processed first to extract describing features. One can think that the more features are extracted the better will be the model. However, a large big set of features can also lead to overfitting. This effect is known as the curse of dimensionality [28]. To overcome this issue, it is sometimes necessary to reduce the feature set dimension p . To do so, several techniques exist and are presented in Section 2.

Learning phase Once data have been prepared, the learning phase can be initiated. Most of the machine learning techniques depend on parameters and/or hyper-parameters² that need to be tuned to better fit the data (see Section 3). During the learning phase two datasets are usually used: the training and the validation dataset. We will see in Section 4 that several approaches exist to divide the data to obtain both sets. In machine learning, parameters are tuned by covering a large scale of possible values and integrating them into the learning phase. Then, trained model is applied on the validation set and the set of parameters that will be retained is the one giving the best performances regarding the objective of classification. As the way of looking performances can change regarding the application, further details on this question are provided in Section 4.1.

Testing the algorithm on unseen data In order to ensure the quality of the model predictions, it is common in machine learning to apply the model on an additional test set of unseen (but labeled) data. This way, if high performances are observed at the learning phase but a poor generalization is observed on the test set, it is a direct indicator of overfitting.

Deployment of the algorithm The last step of the process is to deploy the algorithm in order to make predictions on unlabeled real data. Normally, if the model has been correctly evaluated on a test set independently of the training phase, future classifications would be accurate. However, it can happen that the performances may be altered. Indeed, the annotated data, inherently to conditions of collection, can be unreliable to infer the whole population. Although it is a tough question to ensure a totally random and independent collection of data, it is important to keep in mind this limitation [11].

In the following sections, further details are provided about dimensionality reduction methods, machine learning algorithms and evaluation techniques used for classification. An overview of the methods mentioned hereafter is proposed by Figure 3.3.

2 Dimensionality reduction

Dimensionality reduction is the process of reducing the size of the feature set. Although this step is not mandatory, when the data are described by many features, performing dimensionality reduction is a good trick for, among others, the following reasons [4]:

- Reduce the risk of overfitting;
- Speed up the learning phase;
- Lower the model computational complexity;
- Visualize data using a limited number of dimensions.

2. To avoid confusion, the use of "parameter" and "hyper-parameter" has been dedicated to mention the setting of methods whereas the term "feature" is exclusively associated with data either when it is question of raw features or after computation.

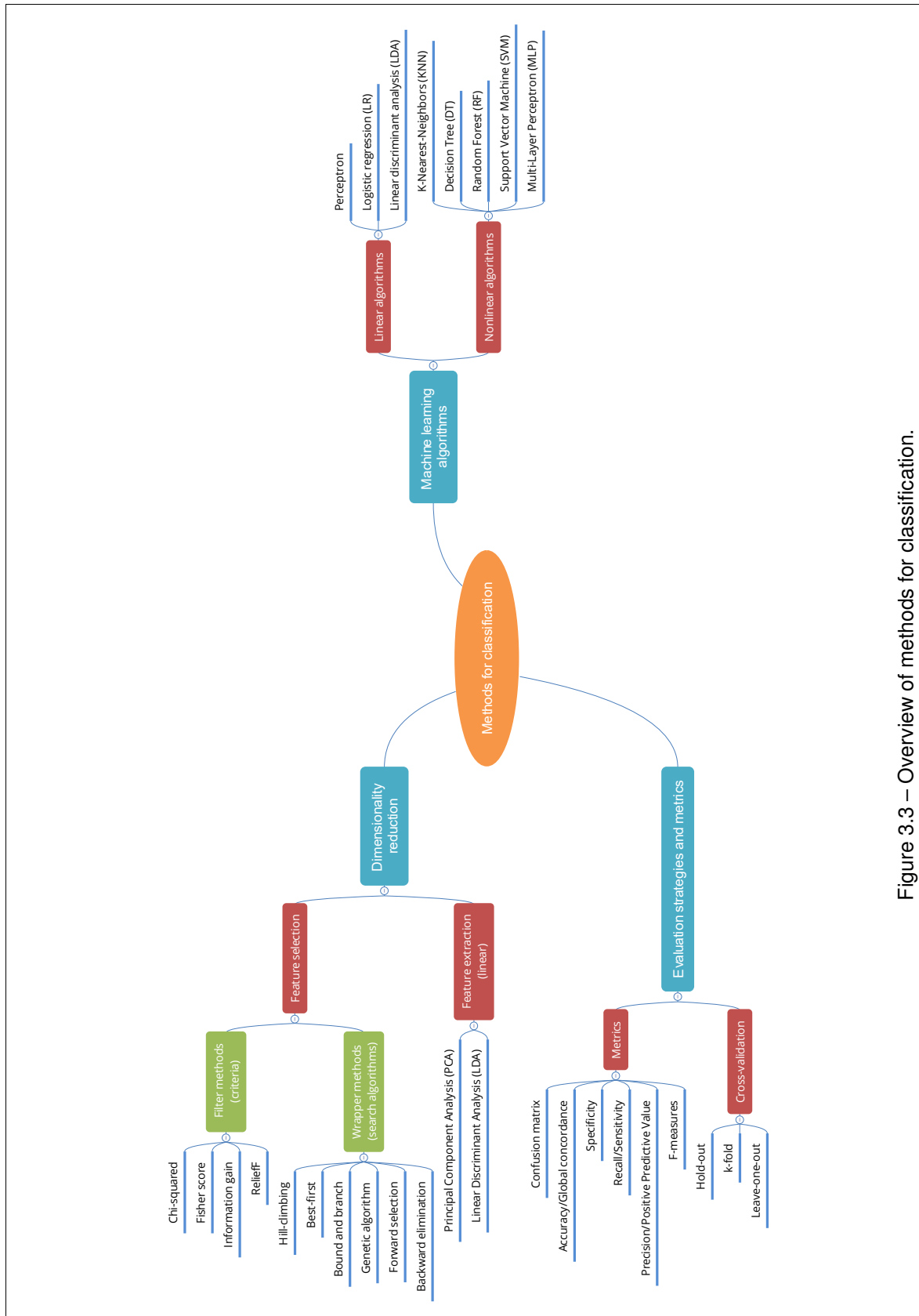


Figure 3.3 – Overview of methods for classification.

In practice, two main groups of dimensionality reduction methods are used: feature selection and feature extraction, sometimes also called feature projection. The main difference between these approaches is that feature extraction maps the original feature space into a lower-dimensional space while, in feature selection, a subset of the original feature set is selected. The choice to use either selection or extraction methods can depend on the objective of the model. If the objective is to better understand the influence of features on classes, feature selection is more suited. Indeed, in feature extraction, the new feature set obtained by projection is generally difficult to link with physical meaning [28].

2.1 Feature selection

Feature selection methods are used to select features that are the most suitable to discriminate samples that belong to different classes. Hence, the goal is to find the best subset of features among the 2^p candidate subsets [8].

Generally, the selection approach, depicted by Figure 3.4, combines feature subset evaluation and search algorithm. Several iterations are performed to search the best subset. For each iteration, a subset is evaluated and its "goodness" is returned.

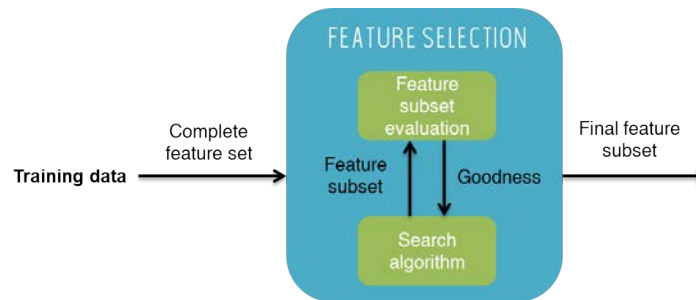


Figure 3.4 – Framework of feature selection methods.

When the original feature set is large, looking for all possibilities is greedy. Hence, computational complexity of the search can be reduced by heuristic or randomized methods to prevent an exhaustive search. These methods are associated with a stopping criterion on the "goodness".

The "goodness" criterion depends on the applied method and objectives. Two main categories of methods stand out: filter models or wrapper models. In the filter approach, the "goodness" is related to information content of the subset while, in wrapper, it is the predictive performances, obtained on the validation set with the subset of features that is evaluated.

2.1.1 Filter methods

Filter models rely on the characteristics of the data without using any classification method [20]. Typically, features are ranked and the highest ranked ones are selected. This can be done in two ways: univariate or multivariate. In the univariate scheme, each feature is ranked independently from others whereas all features are considered simultaneously in the multivariate approach. Several criteria have

been applied to rank data. Among them, we can cite: quality (Fisher score [9, 12]), independency (Chi-squared [2]), redundancy (information gain [24]) or separation of class instances (ReliefF [17]).

Since filter models are easy to understand and implement, it is a popular technique. However, features are selected independently from classification and thus, filter models totally ignore the effects of the selected subset on the performance of the classification algorithm [14].

2.1.2 Wrapper methods

Wrapper models were developed to overcome the limitation of filter models. The subset evaluation is performed by integrating the classifier. Hence, the subset is adapted to the inherent particularities and bias of a predefined classifier [28]. To find the best subset, a wide range of search strategies exists including hill-climbing, best-first, branch-and-bound, and genetic algorithms [13]. Two other popular methods are forward selection and backward elimination where features are respectively added or removed one by one.

Wrapper models provide better predictive performances than filter models [18]. Nevertheless, they are more computationally expensive than filter models.

2.2 Feature extraction

The concept under feature extraction is to project the original feature set of dimension p into a lower-dimensional space of dimension m .

Two types of feature extraction methods can be distinguished depending on the way to combine the features: linear and non-linear. In linear feature extraction methods, new features in the lower dimensional space are given by a linear combination of the original feature set. Reversely, non-linear combinations are sought with non-linear approaches. Linear techniques being principally used, this section mainly focuses on this approach. The main challenge of linear feature extraction methods is to find a transformation W , such as:

$$Z = W^T X \quad (3.1)$$

where Z is the projected data set of size $m * N$, with $m < p$ and N is the number of samples. Dimensionality reduction by the mean of linear feature extraction is depicted by Figure 3.5 with an example of projection from a space of dimensions $p = 2$ to a space of dimension $m = 1$. This example shows that an infinite number of solutions exists to find a new axis to project data points. However, in the literature, two methods stand out: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [28], described below.

2.2.1 Principal Component Analysis

Principal Component Analysis is a well-known technique of data transformation [15]. In its standard formulation, it finds the most discriminant projection of the original feature set by maximizing the variance between data points. In practice, the eigenvectors of the covariance matrix, calculated from the original feature set, are computed. Then, the ones with the largest eigenvalues (principal components) are used to reconstruct a part of the variance of the original dataset.

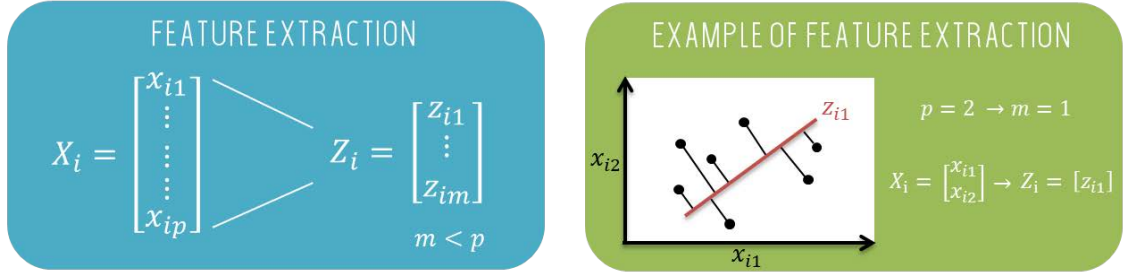


Figure 3.5 – Feature extraction concept (left) and associated example (right). X_i is the original feature set describing the sample i and Z_i is the extracted feature set. In the example, black dots represent data in a two-dimensional space and the red line represents an example of axis for projection in a one-dimensional space.

After that, there are two ways of using principal components regarding the objective of the analysis. The first one is data visualization. In that case, only two or three principal components, carrying most of the variance, will be retrieved in order to plot the data in a two to three dimensional graph. This way, further analyses can be conducted regarding the distribution of data within the objective of class separation. Another purpose of PCA is to feed machine learning algorithms for classification purpose. With PCA, the reduction of dimension can be regulated by the total variance wanted for the feature subset. In other words, the number of principal components that are kept depends on the percent of variance information wanted by the user.

The main limitation of PCA is that there is no guarantee that a small number of principal components with the highest variance will contain the information needed for the classification [21]. Hence, relevant information can be lost and the resulting projected features Z can lead to weak classification performances. Additionally, PCA is only suited for quantitative set of feature. Other factor methods exists. Among them we can cite Multiple correspondence analysis [19] for qualitative feature set and factor analysis of mixed data for mixed feature set [10]. Moreover, a version of PCA, called kernel PCA, has been proposed to make non-linear projections [26]. Briefly, an initial step is first performed to find a particular space where the dataset becomes linearly separable [2].

2.2.2 Linear Discriminant Analysis

Contrary to PCA, in Linear Discriminant Analysis, labels are integrated in the process of dimension reduction. LDA finds the most discriminant projection by maximizing between-class distance and minimizing the within-class distance [1].

In practice, the eigenvectors of the between-class and within-class covariance matrices are computed and the best projection is found using Fisher's criterion.

A comparative example between PCA and LDA is given by Figure 3.6. For a two-class data example, the resulting axis that preserves the variance on the whole data set (PCA) and the one that preserves the distance between both classes (LDA) are drawn. We can see, on each axis, the resulting data projection. LDA offers, in that case, a better projection for discriminating purpose. In fact, besides dimensionality reduction, LDA can also be used for classification (see Section 3.1.3).

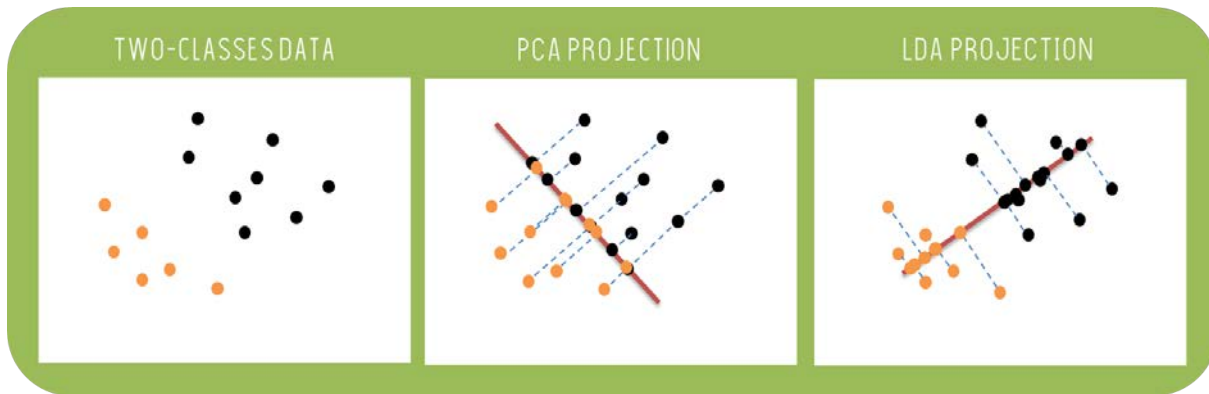


Figure 3.6 – Comparison of data projection between Principal Components Analysis and Linear Discriminant Analysis for a two-classes problem. Orange dots represent data of the first class and data belonging to the second class are reported in black. The red line represents the resulting axis for projection obtained for each method.

3 Machine learning algorithms

As mentioned in Section 1, a supervised machine learning technique is an algorithm that learns from past experiences (training set) a model to make future predictions.

Classifiers³ aim to find the best decision boundaries to discriminate between classes. As illustrated by Figure 3.7, there are two types of classification cases: the ones where classes can be separated with linear boundaries and the ones where classes are not linearly separable.

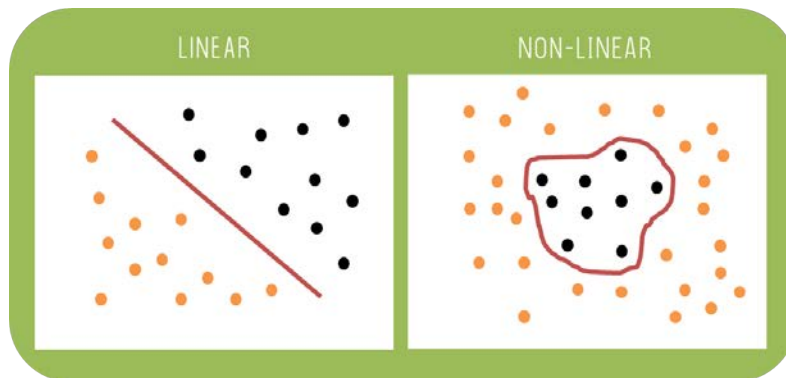


Figure 3.7 – Illustration of a linearly separable case (left) and a non-linearly separable case (right). Red lines represent examples of boundary decisions.

Hence, classifiers can be divided in two groups: linear algorithm and non-linear algorithms. To illustrate the different approaches used to solve classification problems, in this section, we will go through eight commonly used classifiers.

3. In classification, machine learning algorithms are also called classifiers.

3.1 Linear algorithms

Linear classifiers aim to find a linear decision boundary between classes. It can either be a line, a plane or a hyperplane, depending on the dimension of the problem.

3.1.1 Perceptron

The most basic linear model is the perceptron [25], depicted in Figure 3.8.

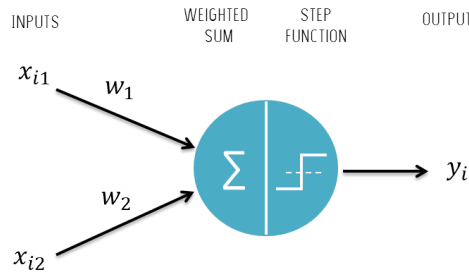


Figure 3.8 – Illustration of a perceptron with two inputs.

In this example, the perceptron output decision y_i , which can either be -1 or $+1$, is computed for each sample i as:

$$y_i = f(w_1 x_{i1} + w_2 x_{i2}) \quad (3.2)$$

where f is a step function, also called threshold or activation function. This can be generalized for a larger dimension p by:

$$Y = f(W^T \mathbf{X} + b) \quad (3.3)$$

where W is the weight vector defined as $W = \{w_1, w_2, \dots, w_p\}$ and b is the bias.

To summarize, to make a prediction, the perceptron computes two quantities. First, the weighted sum of the input features is calculated. Then, this sum is thresholded by the function f in order to retrieve a prediction equal to -1 or $+1$.

During the training phase, W is firstly randomly initialized. Then, weights are settled by considering all the samples of the training set and looking at the output decision [27]. It is performed in three steps which are repeated until all training samples are correctly classified:

1. Make a prediction for an input sample;
2. Compute the error ϵ between the prediction and real label;
3. Adjust the new vector of weights W' accordingly to the error, such as:

$$W' = W + \Delta W \quad (3.4)$$

where $\Delta W = \epsilon \times \eta \times X_i$, with η , called the learning rate⁴ and X_i , the input features of the sample.

4. It is used to regulate the scale of the weight modifications.

So far, we presented a two-class classification approach. In case of multiclass problems, several perceptrons can be trained in order to predict each class versus all others (one-versus-the-rest).

The power of perceptron classifiers resides in the fact that if a linear boundary exists to discriminate classes, the model will converge perfectly. However, in most classification problems, an overlap exists between classes and perceptron models will necessary present misclassification.

3.1.2 Logistic Regression

Logistic Regression (LR) is quite similar to perceptron except that instead of a class prediction, it returns a probability of belonging to the positive class. Hence, in some cases, it may be more robust to class overlap. To compare between perceptron and logistic regression lets restart from Equation (3.3). In logistic regression, b is commonly renamed w_0 , the model intercept. Then, for each sample, a score $S(X_i)$ is computed as:

$$S(X_i) = f(W^T X_i + w_0) \quad (3.5)$$

In logistic regression, f is fixed and is called the logistic (or sigmoid) function. The sigmoid function, depicted by Figure 3.9, is used to associate each score to a probability bounded between 0 and 1, such as:

$$P(Y_i) = f(S(X_i)) \quad (3.6)$$

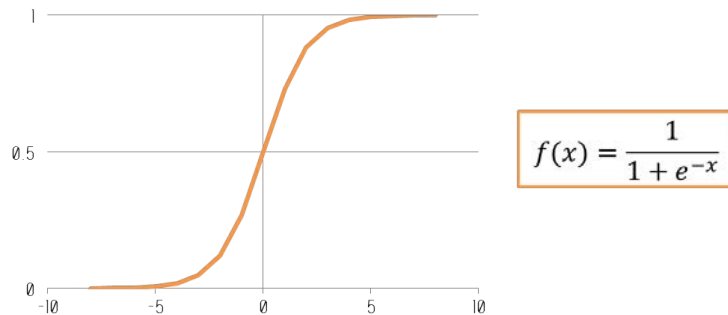


Figure 3.9 – Sigmoid function.

After that, a cut-off value is usually applied on the probability to obtain the class output.

In the learning phase, the vector of weights W has to be estimated. Here, the best W is the one that maximizes the conditional probabilities $P(Y|X, W)$ on the training set. This is commonly done by Maximum Likelihood Estimation (MLE), based on the assumption that weights are normally distributed [6].

Logistic regression is usually applied for binary classification. However, as well as perceptron, it is possible to extend it to multiclass problems by training several one-versus-the-rest LR classifiers.

3.1.3 Linear Discriminant Analysis

As mentioned in Section 2.2.2, Linear Discriminant Analysis can be used for feature extraction. It can also be applied for classification purpose. Contrary to previous linear methods, LDA is more suited

for multiclass analysis [23]. We saw that LDA aims to find the best projection by maximizing the mean between-class distances while minimizing the within-class variance. For multiclass problems, a initial step is added to compute the overall mean (= center) of the data. Then, it is the distance between each class and this center that is maximized while minimizing the within-class variance.

In this way, distributions (means and variance) of each class are estimated. Thus, by the use of Bayes' theorem, the probability of a sample to belong to each class can be estimated. For future predictions, the class associated with the higher probability will be returned.

Nevertheless, LDA presents two main limitations. The first one is due to the number of samples for each class in the training set. If a class is under-represented, the estimated distribution for this class will be corrupted. Secondly, as previous methods, LDA is more suitable on linearly separable multiclass problems.

3.2 Non-linear algorithms

In case of non-linear problems, where class boundaries cannot be approximated by hyperplanes, non-linear algorithms may be more reliable than linear classifiers.

3.2.1 K-Nearest Neighbors

K-Nearest Neighbors is a basic method used for classification. Basically, it computes all the distances between a new sample and the ones of the training set. Then, the majority class of its neighbors is assigned to it. The number of neighbors that contributes to the vote is determined by k .

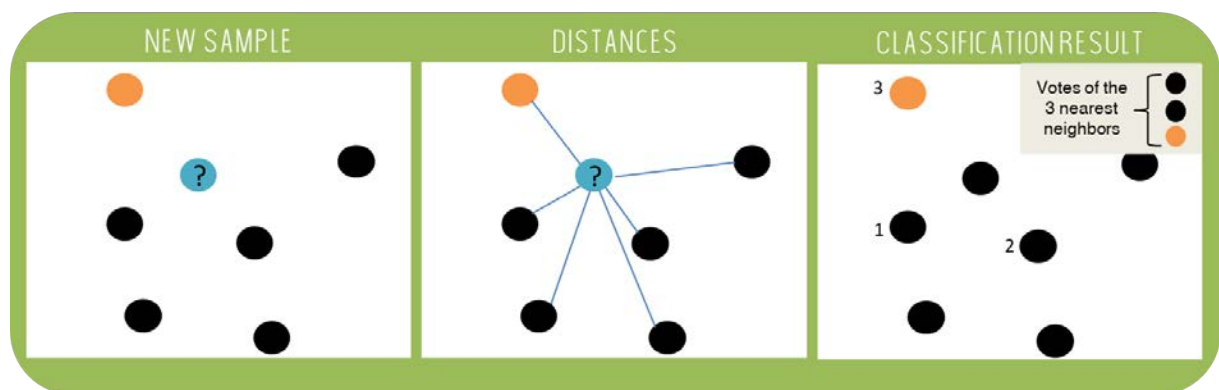


Figure 3.10 – Example of K-nearest neighbors classification with $k = 3$.

An example of KNN classification, with $k = 3$, is given by Figure 3.10. In this example, the class of a new sample (in blue) is sought. The distance with all others points of the learning set is computed. Labels of the 3 nearest neighbours are checked. With two neighbors belonging to class "Black" and one to class "Orange", the new sample is classified as a "Black" sample.

Although Hamming, Manhattan or Minkowski distances can be used, KNN classifier is commonly based on the Euclidean distance. The Euclidean distance between a sample of the training set X_i and

a new sample X_n , is computed as:

$$d(X_i, X_n) = \sqrt{(x_{i1} - x_{n1})^2 + (x_{i2} - x_{n2})^2 + \dots + (x_{ip} - x_{np})^2} \quad (3.7)$$

To compute an accurate distance, it is necessary to work on homogeneous features (i.e., with the same scale). Indeed, absolute differences in features must weight the same to avoid the computation of a meaningless distance.

KNN is a popular algorithm in classification due to its simplicity. In fact, the algorithm is intuitive and easy to implement. Additionally, it can perform well for both linear and non-linear cases.

However, KNN has some drawbacks. It is easily subject to overfitting, especially when working with an imbalanced training dataset⁵. In addition, it can require a lot of memory for future predictions since it needs the training data samples to compute distances.

3.2.2 Decision tree

Decision tree is a predictive model approach which is constructed as a tree-shaped diagram, composed by nodes, branches and leaves. It provides the statistical probability of a class to occur. An example of tree architecture, for a four classes problem and a feature set of three components, is provided in Figure 3.11.

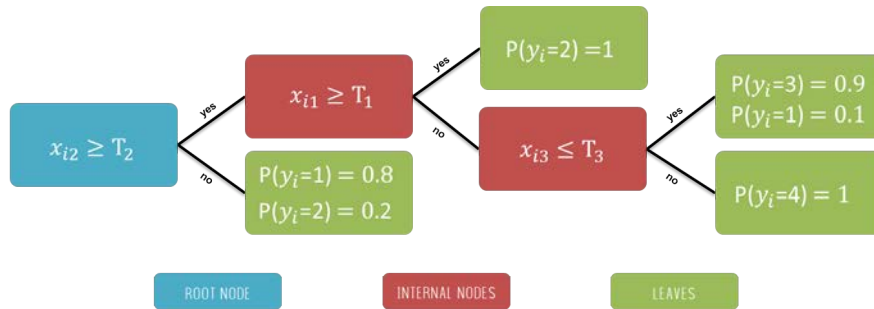


Figure 3.11 – Tree architecture for a four classes examples, where x_{i1} , x_{i2} and x_{i3} are the feature set of a sample i and y_i is the predicted class. T_1 , T_2 , T_3 are three thresholds.

The prediction for a new sample is given using a succession of small tests. Each node corresponds to a test until a leaf, giving the overall output decision, is reached. Hence, in the tree example, if a new data sample n has:

- its feature x_{n2} is superior to T_2 ;
- its feature x_{n1} is inferior to T_1 ;
- its feature x_{n3} is inferior to T_3 .

Then, there is a probability of 0.9 that it belongs to class 3 and a probability of 0.1 to belong to class 1. Generally, the highest probability is retained, making "class 3" the final decision for this sample.

5. when the proportion of samples for a class is high or low in comparison with others

In the training phase, two elements are determined: the architecture (e.g., the order of the considered features, the number of nodes and leaves) and threshold for each feature. A decision tree is constructed following these steps:

1. Split the training set by thresholding a feature;
2. Compute a measure of the quality of the split;
3. Repeat Step 1 and Step 2 until all the splitting possibilities (all thresholds for all features that wasn't used in previous nodes) are associated with a quality measurement;
4. If the split with the best quality measurement improves the classification of the previous node, it is a new node. Otherwise, the previous node was better and is turned into a leaf.
5. Repeat all the previous steps until only leaves can be reached.

To measure the quality of a split, a criterion based on impurity (e.g., Gini's) or information gain (e.g., entropy) can be used [7]. In practice, it is often the Gini's impurity that is applied:

$$Gini = 1 - \sum_{c=1}^C P(c)^2 \quad (3.8)$$

where C is the number of classes and $P(c)$ is the fraction of samples of class c observed after the split. The lower is the Gini's criterion, the better is the split.

Decision trees are easy to understand and interpret. Additionally, they require very few data preparation (such as scaling in KNN) since splits are independently made for each feature. However, they can easily lead to overfitting if the set of features is too wide. As KNN, they are also sensitive to imbalanced dataset since the number of samples in each class impacts the quality criterion.

3.2.3 Random Forests

Random Forest is a part of "ensemble methods" of machine learning. Ensemble methods are based on the assumption that diversified and independent models tend to give better classification results. Therefore, Random Forest is a combination of tree predictors [3]. The classification result comes from the majority vote of a collection of decision trees (also called bagging). In fact, it aims to enhance the generalization by averaging multiple decision trees trained on different parts of the same training set. Figure 3.12 shows the workflow for a new prediction with a RF model of E trees.

In the learning phase, each tree is growing, as described in Section 3.2.2, from a randomly selected subset of the feature set. In this way, each tree is growing with different features. Indeed, if the whole feature set was used, significant features would always come first in the top nodes of splitting which would make all trees be more or less similar.

Random forest is a very popular learning method that can reach really high performances. Contrary to previous methods, it can be robust to imbalanced dataset since a prior knowledge of class occurrences can be integrated in the algorithm. However, results are difficult to interpret since it can be constructed with hundreds of trees. Additionally, sometimes, RF can overfit due to a too noisy dataset (i.e., with a lot of outliers/extreme cases).

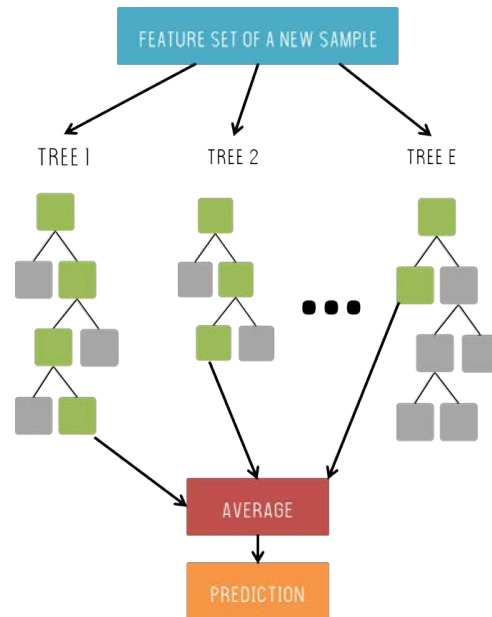


Figure 3.12 – Prediction workflow with a Random Forest model composed by E trees. Successive test results for each tree are reported in green.

3.2.4 Support Vector Machines

Support Vector Machines are suited for both linear and non-linear problems [5]. The aim of SVM is to find the hyperplane boundary that leaves the maximum margin between two classes. Figure 3.13 illustrates the SVM terminology on a linear example.

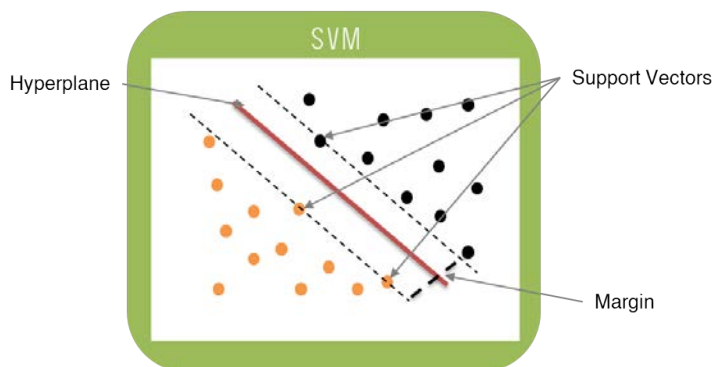


Figure 3.13 – Example of a linear boundary estimated by SVM for a two class problem.

In the learning phase, the basis of SVM is the perceptron. This time, the vector of weights W is estimated ensuring that the margin is maximized between the support vectors of both classes. Support vectors are the data points of both classes which are near the hyperplane. Generally, a parameter g defines the quantities of points that will be taken into account while estimating W . The highest is g , the less support vectors will be used. For multiclass problem, the one-versus-the-rest strategy is often used.

The main strength of SVM is what is commonly called the "kernel trick". The idea is to apply a transformation ϕ to the dataset in order to work on a space where data are linearly separable. An example of kernel transformation is given by Figure 3.14.

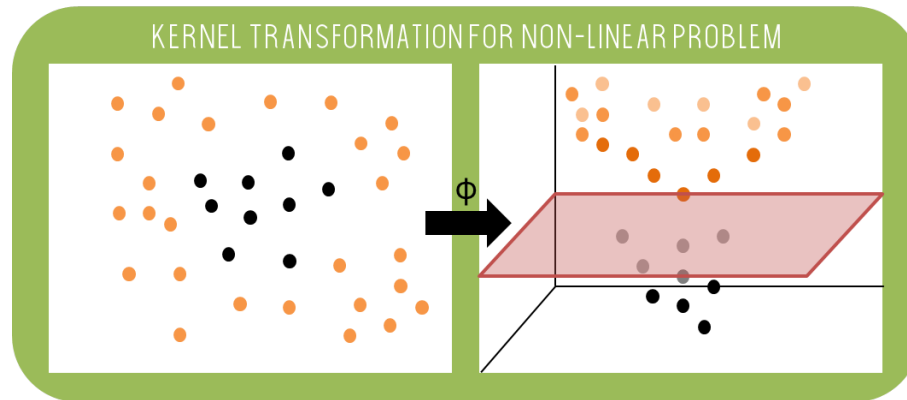


Figure 3.14 – Example of kernel transformation ϕ and boundary estimation (in red) with SVM.

A wide variety of kernel transformations can be used such as polynomial, radial basis function or sigmoid as well as customized kernels.

Support Vector Machines is a widely used machine learning algorithm. Contrary to RF, it is robust to extreme cases since the hyperplane boundary is computed from support vectors. The main drawback of SVM is that a lot of parameters, depending on the kernel, has to be tuned and finding the best set of parameters can be computationally expensive.

3.2.5 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a feedforward artificial neural network. Although a sole perceptron is limited to linear situations, Artificial Neural Networks (ANN) have been constructed in order to solve non-linear problems. There are composed by one input layer, at least one hidden layer and one output layer. Hidden layers and output layers are made up of perceptrons that are all connected from a layer to another. Figure 3.15 depicts an example of MLP with two hidden layers.

Predictions are made by feeding the feature set of a sample to the network. Then, perceptrons are progressively activated (or not) until reaching an output node which gives the predicted class. Contrary to the linear case, the activation function of each perceptron is non-linear. Among them, we can cite hyperbolic tangent, rectified linear unit function or sigmoid⁶.

For the learning phase, MLP uses a supervised technique called backpropagation. As for the linear case, all samples are passed through the network and weights W are updated according to the output error [22]. Weights of each perceptron are updated one-by-one, starting from the ones of the output layer.

6. Note that in LR, a sigmoid was also used and LR was classified as a linear model. In fact, the output of LR can be written in a linear form whereas, because of the multiple perceptrons, it is not possible to write a linear equation to summarize neural networks outputs. Thus, Neural Networks are classified as non-linear models.

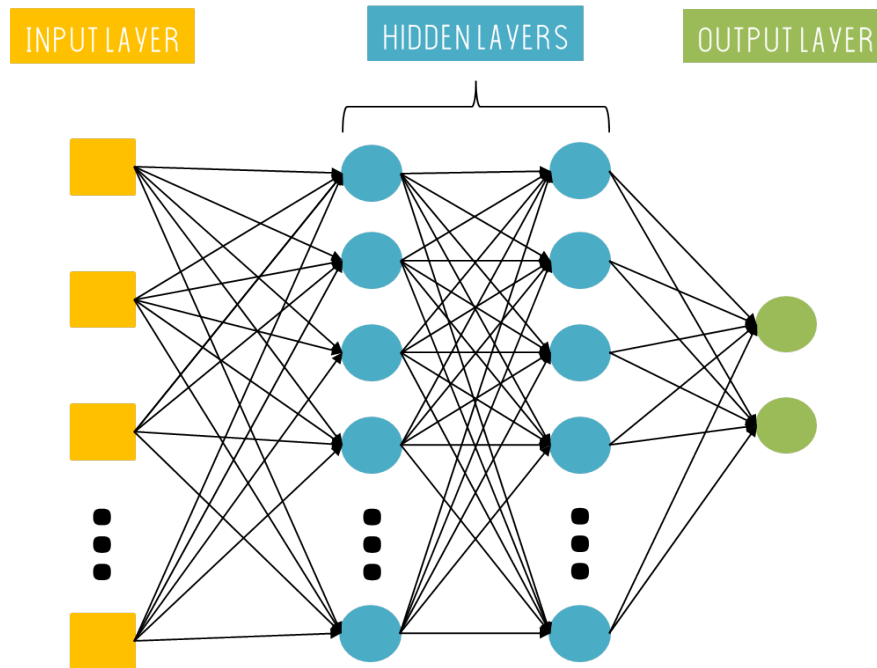


Figure 3.15 – Example of multi-layer perceptron with 2 hidden layers and two output classes.

Neural network is a field in expansion notably because of improvement in computing capacities. Nowadays, neural networks can be composed of a large amount of hidden layers, increasing their depth. This increase of the computing capabilities leads to an evolution of Neural Networks (e.g., Convolutional Neural Networks, Recurrent Neural Networks) to what is today commonly called deep learning. The main drawback of neural networks is that it involves a lot of parameters to tune (e.g., number of layers, number of perceptrons per layer, connections between layers). In addition, it is also subject to overfit if the training set is too small or not representative of the whole population. Although today, various complex problems can be solved with deep learning, it is important to remind that it exists alternatives (e.g., SVM, RF) that are faster, easier to train and can provide better performances regarding the classification objective.

4 Techniques for performance evaluation

As briefly mentioned in Section 1, evaluation strategies can change regarding the classification objectives, as well as considering the available data. This section firstly focuses on the metrics that are used to evaluate the performances. Secondly, we will go through different techniques used to split an annotated dataset in order to ensure good results for future predictions.

4.1 Metrics

Parameters and hyper-parameters of machine learning algorithms are tuned by maximizing the performances metrics. Additionally, performance metrics are mandatory to compare the results of different

classifiers. In practice, most of performance metrics are based on the confusion matrix, reporting the number of good classifications and misclassifications by comparing the predicted with actual labels. A confusion matrix for a two class problem is reported in Figure 3.16. It is composed of four numbers:

- True Positive TP : number of samples correctly classified as yes;
- True Negative TN : number of samples correctly classified as no;
- False Positive FP : number of samples classified as 'yes' instead of no;
- False Negative FN : number samples classified as 'no' instead of yes.

	PREDICTED YES	PREDICTED NO
ACTUAL YES	TP	FN
ACTUAL NO	FP	TN

Figure 3.16 – Confusion matrix for a two class classification.

From there, the overall classification accuracy Acc measures the total of good classification over the whole data set:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.9)$$

Measuring the accuracy of a classification is important to give an insight on what the algorithm is capable of. However, it is not always the best way of looking. In fact, for example, in case of imbalanced dataset, a class can take the lead on the accuracy metric and a very high value can be observed even if a less represented class is never detected. Actually, it is often the case in biomedical engineering, notably when working on rare incident diseases. For example, for disease screening, it may be better to hand up with some false positive cases than missing one patient. Reversely, in some cases, it may be more important to be sure that samples predicted as belonging to a class really belongs to this class.

For these reasons, it exists a bunch of metrics which analyze the performances in different ways. A list of the more common ones is provided hereafter.

Historically, sensitivity and specificity are used to precise classification results:

- Sensitivity Se , which measures the proportion of actual positives that are correctly identified as such:

$$Se = \frac{TP}{FN + TP} \quad (3.10)$$

- Specificity Sp , which measures the proportion of actual negatives that are correctly identified as such:

$$Sp = \frac{TN}{FP + TN} \quad (3.11)$$

Recently, with the gain of popularity of machine learning, recall, precision and F-scores are commonly used:

- Recall R , corresponds to Se :

$$R = \frac{TP}{FN + TP} \quad (3.12)$$

- Precision P , also known as the Positive Predictive Value (PPV), which measures the fraction of actual positives among the retrieved positive cases:

$$P = \frac{TP}{TP + FP} \quad (3.13)$$

- F1-score F_1 is the harmonic mean of precision and recall and allows a single measure of performance:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \quad (3.14)$$

- F_β score, which is the weighted (according to β) harmonic average of precision and recall:

$$F_\beta = \frac{(1 + \beta)^2 \cdot (P \cdot R)}{(\beta^2 + P + R)} \quad (3.15)$$

All these metrics are bounded between 0 and 1, where a value of one means a perfect score. Until there, it may seem that these metrics are only applied on two class problems. However, all these performance measurements can be generalized to multiclass problems, notably by constructing a confusion matrix of one-versus-the-rest for each class.

4.2 Cross-validation

In order to validate the classifier abilities to make future predictions, several evaluation strategies can be followed. Cross-validation is the most popular method that helps to ensure the robustness of a classifier since it allows the detection of underfitting or overfitting [7]. In this section, three popular techniques of cross-validation are presented: Hold-out, K-fold cross-validation and Leave-one-out cross-validation.

4.2.1 Hold-out

The hold-out method is the simplest cross-validation technique. Data is divided into two parts: the training set and the testing set, as depicted by Figure 3.17. This way, the classifier is trained on the

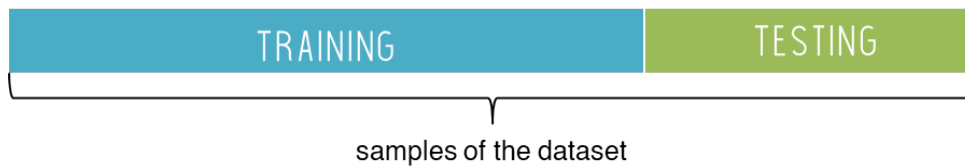


Figure 3.17 – Illustration of the hold-out strategy.

training set and can be evaluated on unseen data of the testing set. Usually, a larger part of the data is used for training. There is two different ways to divide the data: by randomly selecting a number of samples over the whole dataset or by randomly selecting a number of samples of each class, in order

to ensure a representation of all classes in the training and in the testing set. This method is usually applied for small datasets.

4.2.2 k-fold cross-validation

One way to improve the hold-out strategy is to perform a k -fold cross-validation to tune the parameters and train the model. The idea is to train the model on k different data splits to ensure its robustness. For each iteration, a training set and a validation set are constructed. An example of 3-fold cross-validation is provided by Figure 3.18.

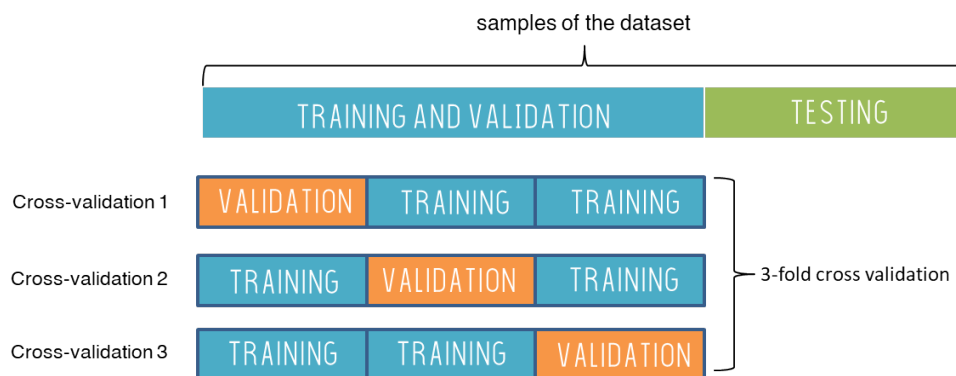


Figure 3.18 – Illustration of the 3-fold cross-validation strategy.

For better control on the classifier performances, it is also recommended to work with an additional set, that wasn't seen during the learning phase and the tuning of the parameters. This set is called the testing set. Hence, in Figure 3.18, the 3-fold cross-validation step is only performed on a part of the data.

4.2.3 Leave-one-out cross-validation

Leave-one-out cross-validation (LOOCV) is the extreme k -fold validation, where k is equal to the number of samples in the training/validation dataset. An example of LOOCV is depicted by Figure 3.19.

Another way to perform LOOCV exists in biomedical engineering. In fact, to evaluate the capacity of classifiers to work with a new patient, a leave-one-patient-out strategy is often performed by working on all patients except one at each time. Performances for each draw indicate the generalization of the model to make future predictions for a new patient.

5 Conclusion

In this chapter, the key concepts of machine learning for classification were presented. Hence, the process of designing a new classifier was described. It goes from the data collection to the deployment of a new model to make future predictions. We saw that at each step of the design there are some

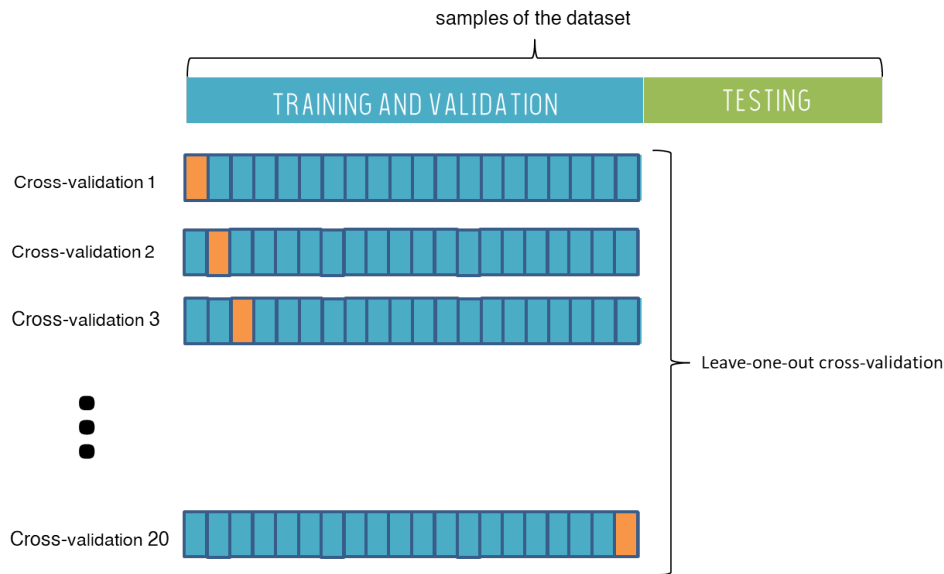


Figure 3.19 – Illustration of the leave-one-out cross-validation strategy for a training/validation dataset of 20 samples.

considerations to keep in mind to avoid the construction of non-generalized classifiers. Ensuring the robustness of a model can be done either by carefully selecting or extracting relevant features or by correctly tuning classifier regarding our objective as well as using an accurate evaluation strategy.

However, the main issue remains the construction of an informative database. In fact, if the dataset is not representative of the reality, all necessary precautions can be taken to avoid underfitting or overfitting, it will be impossible to get to a generalized model.

In the following chapters, several classification problems are tackled for: automatic detection of sleep stages, segmentation of motion and cry extraction.

Bibliography

- [1] BALAKRISHNAMA, S., AND GANAPATHIRAJU, A. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing 18* (1998), 1–8.
- [2] BONACCORSO, G. *Machine learning algorithms*. Packt Publishing Ltd, 2017.
- [3] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.
- [4] COELHO, L. P., AND RICHERT, W. *Building machine learning systems with Python*. Packt Publishing Ltd, 2015.
- [5] CORTES, C., AND VAPNIK, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.
- [6] CZEPIEL, S. A. Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf* (2002).

-
- [7] DANGETI, P. *Statistics for machine learning*. Packt Publishing Ltd, 2017.
- [8] DASH, M., AND LIU, H. Feature selection for classification. *Intelligent data analysis 1*, 1-4 (1997), 131–156.
- [9] DUDA, R. O., HART, P. E., AND STORK, D. G. *Pattern classification*. John Wiley & Sons, 2012.
- [10] ESCOPIER, B., AND PAGÈS, J. *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation*. Dunod, 2008.
- [11] GRIMSON, E., GUTTAG, J., AND BELL, A. Introduction to computational thinking and data science., 2016.
- [12] GU, Q., LI, Z., AND HAN, J. Generalized fisher score for feature selection. *arXiv preprint arXiv:1202.3725* (2012).
- [13] GUYON, I., AND ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research 3*, Mar (2003), 1157–1182.
- [14] HALL, M. A., AND SMITH, L. A. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In *FLAIRS conference* (1999), vol. 1999, pp. 235–239.
- [15] JOLLIFFE, I. *Principal component analysis*. Springer, 2011.
- [16] JOSHI, P. *Artificial intelligence with python*. Packt Publishing Ltd, 2017.
- [17] KIRA, K., AND RENDELL, L. A. The feature selection problem: Traditional methods and a new algorithm. In *Aaai* (1992), vol. 2, pp. 129–134.
- [18] KOHAVI, R., AND JOHN, G. H. Wrappers for feature subset selection. *Artificial intelligence 97*, 1-2 (1997), 273–324.
- [19] LE ROUX, B., AND ROUANET, H. *Multiple correspondence analysis*, vol. 163. Sage, 2010.
- [20] LIU, H., AND MOTODA, H. *Computational methods of feature selection*. CRC Press, 2007.
- [21] NEAL, R. M., AND ZHANG, J. High dimensional classification with bayesian neural networks and dirichlet diffusion trees. In *Feature Extraction*. Springer, 2006, pp. 265–296.
- [22] PARIZEAU, M. Le perceptron multicouche et son algorithme de rétropropagation des erreurs. *département de génie électrique et de génie informatique, Université de laval* (2004).
- [23] RAO, C. R. Tests of significance in multivariate analysis. *Biometrika 35*, 1/2 (1948), 58–79.
- [24] ROOBAERT, D., KARAKOULAS, G., AND CHAWLA, N. V. Information gain, correlation and support vector machines. In *Feature extraction*. Springer, 2006, pp. 463–470.
- [25] ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review 65*, 6 (1958), 386.

- [26] SCHÖLKOPF, B., SMOLA, A., AND MÜLLER, K.-R. Kernel principal component analysis. In *International conference on artificial neural networks* (1997), Springer, pp. 583–588.
- [27] SHIFFMAN, D. The nature of code: Chapter 10. *Neural Networks* (2012).
- [28] TANG, J., ALELYANI, S., AND LIU, H. Feature selection for classification: A review. *Data classification: algorithms and applications* (2014), 37.

PRELIMINARY BEHAVIORAL SLEEP STATES ESTIMATION

We saw that several audio and video descriptors can be extracted to characterize newborn development. Here, even if we are not in the real time monitoring context, we managed to exemplify a new and very promising approach where video and audio monitoring can be combined. The evaluation of sleep behavior is addressed, based on the extraction of eye state, motion and vocalizations. This work must be seen as a first attempt to propose a semi-automatic sleep states estimation method based on machine learning. This work has been published in Biomedical Signal Processing and Control journal[8]. The content of this paper is reported hereafter and is supplemented by a conclusion section that introduces the priority developments made as part of this thesis.

1 Introduction

Preterm birth, defined as birth before 37 weeks of gestation, is concerning 15 million babies per year or 11% of all live births worldwide and this number tends to increase every year [6].

Premature babies have several immature functions such as digestive, immunological, cardio-respiratory or neurological functions and begin their life in a Neonatal Intensive Care Unit (NICU), generally in an incubator, under high medical supervision. Their health status is monitored, as well as their maturation, in order to program the incubator exit and then the discharge home. This monitoring relies on several vital signs (cardiac activity, breathing, blood pressure...), and may be extended by a sleep analysis, leading to a sleep stage sequence as a function of time, also called hypnogram [23, 50].

Since the sleep behavior differs across PostMenstrual Age (PMA), its analysis may give a good indication of the degree of brain maturation. In particular, the duration of sleep/wake cycles is supposed to increase with PMA and the sleep organization to evolve with more time spent in quiet sleep [12]. For neonates, two sleep scoring techniques exist: the polysomnography based on the analysis of the ElectroEncephaloGram (EEG) and the direct behavioral observation, which is most commonly used. Based on the rules of Prechtl [43], the sleep scoring is performed in the presence of the baby, by observing body activity levels, eye state (open or closed), respiration regularity, vocalizations... This technique, contactless and without constraint for the baby, is particularly recommended in the NIDCAP (Newborn Individualized Developmental Care and Assessment Program) [2], centered on the comfort of the baby. However, sleep analysis (polysomnography or behavioral observation) is difficult to install, time consuming and cannot systematically be used. In this context, development of new ways to automatically

monitor the neonates, using contactless modalities, is necessary.

This work is a part of the Digi-NewB project, funded by the European Union programme for Research and Innovation Horizon2020. Its objective is to reduce mortality, morbidity and health costs of hospitalized newborns by assisting clinicians in their decision-making related to sepsis risk and neurobehavioral maturation. For this purpose, the project aims to develop a new generation of monitoring systems in NICU, using clinical and signal data from different sources (electrophysiological, audio and video).

First studies using video as a support of sleep analysis appeared in 1969, when the Association of the Psychophysiological Study of Sleep (APSS) highlighted the importance to develop a guide for assessing infant sleep. In fact, contemporary criteria (Rechtschaffen and Kales [44]) were not applicable to the infants, who present unique behavioral features of development. Two years later, this manual was proposed by Anders et al. and recommended to supply polysomnographic recordings by behavioral observations [3]. From there, Anders et al. proposed to study infants using only behavioral observations, sometimes supplemented by time-lapse video recordings, an alternative method for long-term recordings. In [4], a study was performed with full-term infants, at two and eight weeks of age. Behavioral states were scored from video considering eye state, vocalizations and movements. The polygraphic scoring was based on EEG, electrooculogram, electromyogram and respiratory signals. A correlation of 0.79 was obtained between both scorings of the three states (Active REM Sleep, Quiet REM Sleep, Wakefulness). Fuller et al. proposed a similar approach with premature newborns, but only focused on the eye state and the body movements. Furthermore, only sleep stages were considered and vocalizations were not included [18].

The automatic sleep stage classification has been much less addressed in newborns, full-term or preterm, than in adults. However, several modalities were studied including EEG [5, 13, 14, 16, 41, 42], cardiorespiratory signals [20] and facial expressions [21]. Though, these methods offer a sleep stage classification more or less specific regarding the PMA. In fact, EEG can only be investigated to distinguish quiet sleep from all other sleep stage under the age of 32 weeks PMA, whereas facial expression assessment can provide a more specific sleep stage classification since 26 weeks PMA. For their part, cardiorespiratory analyses can be reliable for particular sleep stage qualification before 32 weeks PMA [50].

This paper proposes to estimate behavioral sleep states from audio and video acquisitions, which is suitable with a contactless and non invasive monitoring, an approach never envisaged before in the context of NICU [9]. It is based on two main steps (Figure 4.1):

- Audio and video processing, leading to the characterization of information of three types: i) vocalizations, ii) motion and iii) eye state.
- Combination of the three signals in order to obtain an estimation of the behavioral sleep states.

Paper is organized as follows. In section 2, methods for the extraction of vocalizations, motion and eye state and the automatic estimation of sleep stages are described. Section 3 is devoted to the results concerning the eye state estimation algorithm and the classification of sleep stages.

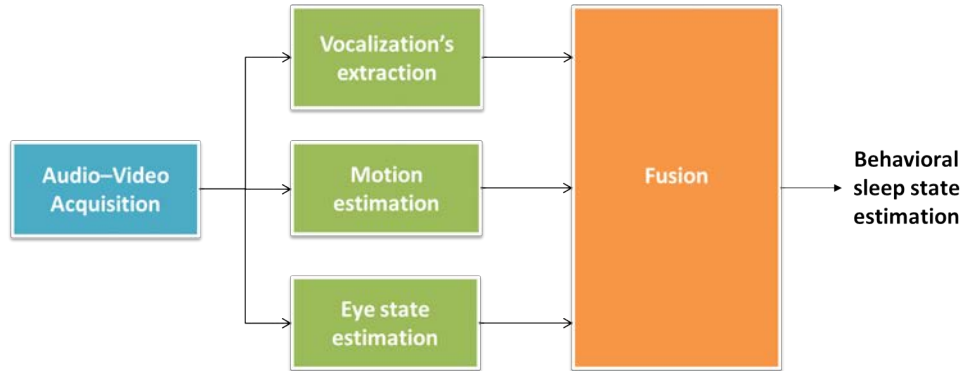


Figure 4.1 – Workflow of the methodological steps for the behavioral sleep state estimation.

2 Methods

In this section, methodology for the extraction of information from audio and video is first presented. A larger part is devoted to the description of the method we developed for the eye state estimation. Then, we propose to estimate sleep stages from these extracted signals by the use of different classifiers.

2.1 Database

Videos were acquired during a project conducted at the University Hospital of Rennes, approved by the Committee on Protection of Individuals (CPP Ouest 6-598) and complying with standards established by the Declaration of Helsinki. Ten newborns were included in this study and the signing of an informed consent, was obtained from parents for each of them.

During the experiments, a camera was set up in the room of the babies to record the scene. It was installed near the bed in order to observe the major part of the body. Recordings were performed in moderate obscurity. The camera had a resolution of 720x756 pixels and recorded 25 frames per second. Sound was acquired by a microphone integrated in the camera with a frequency sampling of 8 kHz. The choice of this low sampling rate has been motivated by our objective to simply detect periods with sound activity while keeping a fast computation time. To consider conditions compatible with a monitoring context, no specific setup was imposed.

Recordings were performed between the 7th and the 11th day of life of premature newborns having a GA ranged from 26 to 32 weeks and, consequently, a PMA comprised between 28 and 33 weeks (see Table 4.2). Each video duration was between 10 and 32 minutes, leading to a total duration of more than 4 hours (242 minutes and 14 seconds).

For each video recording, a frame was randomly selected and reported on Figure 4.2 to illustrate the complexity of the database.

One can observe that all infants lay on one side and are most of the time covered by a blanket. Four of them are equipped with a ventilatory support and six of them are intubated. Differences can also be noticed in luminosity conditions and camera distances. In audio recordings, only background noises of low energy, for example emitted by the ventilator, were reported.

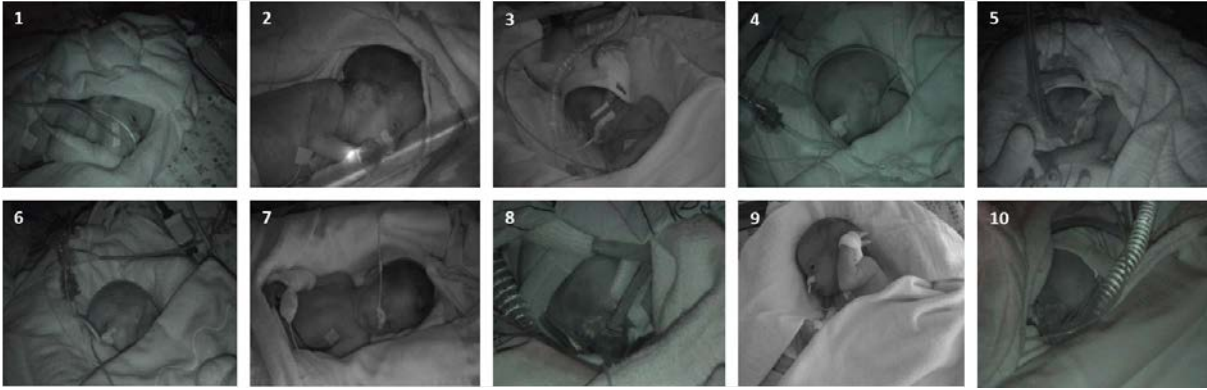


Figure 4.2 – Overview of 10 videos of premature newborns with a PMA comprised between 28 and 33 weeks. All infants lay on one side and are most of the time covered by a blanket. Four of them are equipped with a ventilatory support and six of them are intubated.

A scoring of sleep stages, based on a direct behavioral observation [43], was synchronously carried out by a NIDCAP expert during the recording, considering five stages: Quiet Sleep (QS), Active Sleep (AS), Drowsiness (D), Quiet Alert (QA) and Active Alert (AA).

2.2 Vocalizations' extraction

The development of automated sound processing began in the 1960s (see [49] for an historical review). In the context of monitoring, the main issue is to extract newborns vocalizations, also called Voiced/UnVoiced (V/UV) detection. Several strategies were recently proposed to perform this detection automatically. A V/UV detection procedure was proposed in [35], where an interval was selected as voiced if the maximum of the autocorrelation function is greater than a fixed threshold. Several techniques based on the thresholding of the Short-Term Energy (STE) were also investigated [15, 26, 31–34].

Here, baby vocalization detection is performed by applying the methodology proposed in [32]. It is based on the computation of the STE in 20 ms length windows, with 50% overlap between adjacent windows. Then, the highest STE intervals, corresponding to baby vocalizations, are detected using two thresholds, automatically selected using Otsu method [36]. Finally, to construct $V(t)$ signal, values of unvoiced frames are set to 0. An example of a raw sound signal and the resulting vocalization signal $V(t)$ is given in Figure 4.3.

2.3 Motion estimation

Many methods have been proposed in the literature to estimate and characterize motion in videos [29]. In paediatrics, specific topics such as general movement assessment or detection of neonatal seizures were addressed [27, 38, 46]. In this work, the goal is not to estimate the local motion of the baby, which would be very challenging because of the unconstrained acquisition setup, but to globally characterize its activity. For this purpose, we consider the modifications between two successive frames by computing their difference [30].

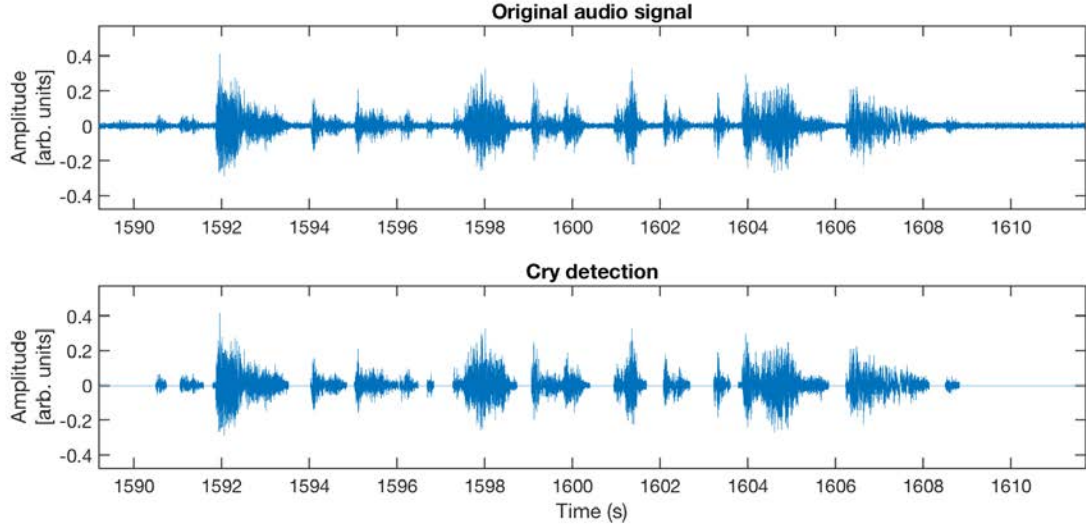


Figure 4.3 – Example of a soundtrack processing (video 9): Top: Raw sound signal - Bottom: Vocalization signal $V(t)$.

In order to limit the influence of noise, the resulting difference image is thresholded with a value T_M (typically low). The amount of activity at a time t , or "motion signal" $M(t)$, is obtained by counting the number of pixels above the threshold. An example is given in Figure 4.4.

2.4 Eye state estimation

As for adults, where the conditions are usually controlled (full face, front view and open eyes, good luminosity, no occlusion. . .) [1], eye tracking of infants has been only addressed in a few specific studies. For instance, they were always located in front of the camera and seated either on a parent's lap [24], in an infant chair [22] or in a baby car seat [17, 19].

These conditions being not fulfilled, a specific algorithm was developed for the estimation of the eye state $E(t)$. As the baby may move during the recordings, the algorithm relies on a tracking step of the region of an eye associated with a detection step in this area. It is a semi-automatic procedure, with a limited number of user interactions.

2.4.1 Initialization

The region of the eye has to be tracked since the baby moves during the video acquisition. However, since the state of the eye is changing (open, closed, or in-between), its appearance is often modified. Thus, we decided to perform the tracking of another region of the face, supposed to keep the same appearance and to be at a constant distance from the eye. This region is called the "reference" region of interest (R_{Ref}), and may for example include the nose or an ear. Depending on the acquisition characteristics, the choice is left to the user and performed during the initialization of the processing, i.e. on the first frame ($R_{Ref}(0)$). The user has also to select the region of the eye ($R_{Eye}(0)$). The

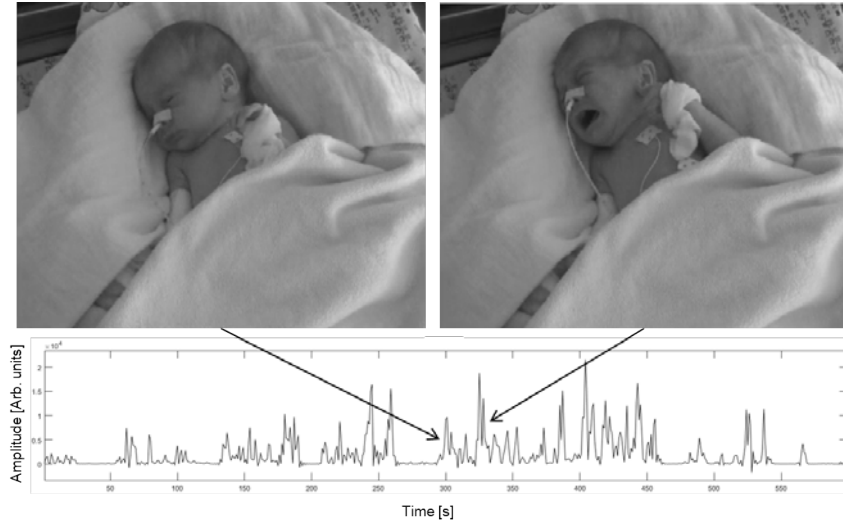


Figure 4.4 – Example of a motion signal $M(t)$ (video 9). Frames acquired before and during movements are presented.

link between both regions is defined by the relative position between the regions' centers, called δ_{ROI} (Figure 4.5(a)).

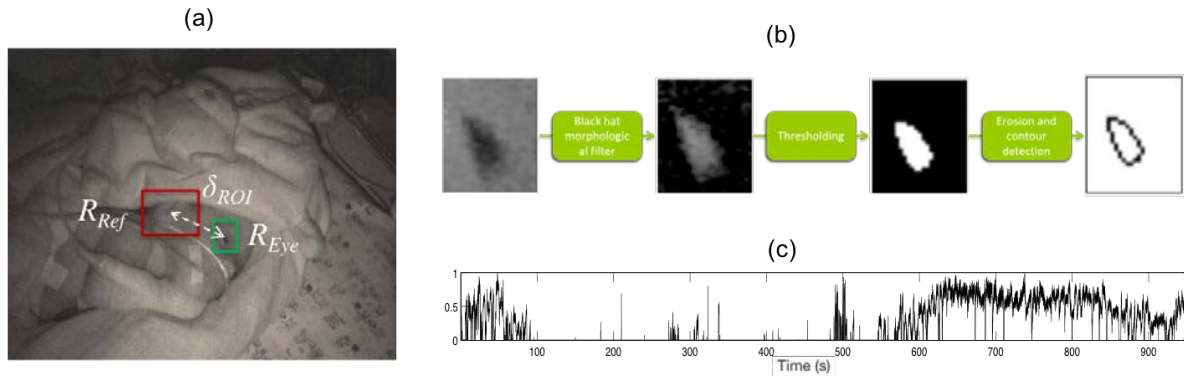


Figure 4.5 – Illustration of the eye state estimation algorithm (video 1): (a) Initialization; (b) Processing steps for the eye segmentation; (c) Example of an eye state signal $E(t)$ with normalized values.

2.4.2 Tracking of the reference region

For each new frame, the reference region is firstly tracked using the template matching approach. It is based on the comparison between the template and each possible position in the new frame. Since the motion amplitude is limited between two successive frames, the search is restricted to a region centered on the previous position of the reference. For each position in the search region, a metric is computed to evaluate the correspondence between the template and the new frame's region centered on this position. The reference region being supposed to keep its appearance, the considered metric

was the Sum of Squared Differences (SSD) between the pixels' intensities.

For each new frame at time t , the considered template is, firstly, the initial reference region ($R_{Ref}(0)$). Then, the minimal value of the metric $SSD(t)$ obtained by the template matching is compared to two threshold values D_1 and D_2 , as follows:

- $SSD(t) < D_1$: the reference region is considered to have been found in the frame t ; $ROI_{Ref}(t)$ is defined as the corresponding position;
- $D_1 \leq SSD(t) \leq D_2$: the resulting position has to be refined; a new template matching is applied, using $R_{Ref}(t-1)$ as the template;
- $SSD(t) > D_2$: the reference region has been lost, the tracking is stopped and the system waits for user's interactions.

At each frame t , after the estimation of $R_{Ref}(t)$, the eye region $R_{Eye}(t)$ is retrieved thanks to the relative position δ_{ROI} .

2.4.3 Eye detection

Once the eye region $R_{Eye}(t)$ has been found in the frame t , a segmentation process is used to extract the eye contour. It includes the following steps (Figure 4.5(b)):

- A "black hat" morphological transformation using a structuring element of size 15x15 to enhance the contrast between the darkest regions and their neighborhood;
- A thresholding of the resulting image by a value T_E ;
- A morphological erosion using a structuring element of size 5x5 to remove small regions;
- An edge detection of the extreme outer contours, using Green's theorem.

Then, the eye state $E(t)$ is defined by its surface, depending on the number k of detected contours, as follows:

- $k = 0$: the eye is considered closed and $E(t) = 0$;
- $k = 1$ or $k = 2$: the eye is considered as open, $E(t)$ is the sum of the surfaces of the k detected edges and δ_{ROI} is updated with the center of the eye area. The case $k = 2$ occurs when the contour is divided in two areas separated by the pupil;
- $k > 2$, this result corresponds to noisy detections. The eye is taken for not detected, the tracking is stopped and the system waits for a user interaction.

2.4.4 User interactions

In the case of uncertainty concerning the tracking ($SSD(t) > D_2$) or concerning the detection ($k > 2$), the algorithm stops and the user has to select again both regions of interest (as in the first frame), possibly after forwarding the video if one occlusion occurs.

2.4.5 Smoothing of the eye state signal

Once the video recording has been processed, a sliding median filter is applied on the eye opening values to limit the brief incoherent changes. Since an eye blanking has been observed to last less than five successive frames, the median filter window length has been set up at 5.

An example of a eye state signal is given in Figure 4.5(c).

2.5 Sleep stage estimation

In this section, we propose a strategy to characterize newborn sleep organization based on the fusion of the extracted descriptors.

In this objective, data are first standardized by applying a set of post-processing to the three signals $V(t)$, $M(t)$ and $E(t)$:

- A Hilbert transform is applied to the vocalizations signal $V(t)$ to recover the signal envelope and proceed next with a positive signal;
- It is also downsampled to 25 Hz, the sampling frequency of motion and eye state signals;
- The three signals are smoothed using a median filter on 1-second length windows;
- The three signals are normalized to the range $[0, 1]$, relatively to the global maximum of the database (separately for each type of signal).

The resulting signals are called $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$.

Then, a model to estimate sleep stages on the whole population, based on machine learning, can be built. For this purpose, each t is considered as a sample described by three features $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$, associated with its sleep stage label (QS, AS, D, QA or AA). We selected five commonly known approaches that cover a large scope of classification hypotheses: K-Nearest Neighbors (KNN) [40], Linear Discriminant Analysis (LDA) [28], Support Vector Machine (SVM) [48], Random Forest (RF) [7] and Multi-layer Perceptron (MLP) [37]. Since some of these methods are more reliable on balanced dataset, a random under-sampling method was first applied to equalize the number of elements of each sleep stage class. Then, an hold-out evaluation is performed. The dataset is randomly split into a training and a testing part containing respectively 60% and 40% of the balanced dataset. These operations are repeated 30 times.

3 Results

This section is dedicated to the validation of our approach. First, software and platforms that were used to produce these results are reported. As sound segmentation and motion estimation have been already evaluated by their authors, their conformity was only confirmed by visual assessment. Nevertheless, an original strategy was defined for the evaluation of our eye state estimation method.

Then, performances concerning sleep stage estimation are given by comparing the results of the five different machine learning approaches to an expert annotation.

3.1 Software and platforms

Several software and platforms have been involved in this project. Video processing (motion and eye state estimation) was developed in C++ with OpenCV 3.0 library whereas vocalization extraction and statistical analyses were performed with Matlab R2018a. Machine learning approaches were implemented in Python 3.6 using scikit-learn 0.20.0 [39].

3.2 Tuning of the parameters

Eye state detection Some parameters in motion as well as in eye state detection algorithms had to be tuned to fit the database properties.

Regarding motion analysis, the threshold T_M has been defined by studying the cumulative histogram of video sequences with empty rooms (without baby or adult), resulting to a low value of 10, initial intensity ranging in $[0, 255]$.

Three thresholds were used in the eye state estimation algorithm. Thresholds D_1 and D_2 (intensity ratios), that defined the accepted appearance modifications of the reference region, were empirically set to 0.02 and 0.04, respectively. The threshold T_E was manually selected between 15 and 25, depending on the video luminosity. In fact, the contrast is less pronounced in low luminosity videos and consequently, a lower threshold is necessary.

Classifier parameters In section 3.4.1, five classifiers are compared. For each of them, the set of parameters resulting to the highest performances was first identified. For that purpose, several parameters and hyper-parameters have been tuned. A summary of the tests is reported in Table 4.1.

Table 4.1 – Parameters testing summary. Final selecting sets of parameters are marked in bold.

Method	Parameters
KNN	Number of neighbors $\in [1, 3, 5, 10, 20]$ Distance: Manhattan or Euclidean
LDA	Solver \in [singular value decomposition, least squares solution , eigenvalue decomposition]
SVM	Kernel \in [linear, Gaussian , polynomial] Hyper-parameters depending on the kernel: → linear: no additional parameter → Gaussian: margin $\in [0.01, 0.1, 1, 10, 100, 10^3, 10^4, \mathbf{10^5}, 10^6, 10^7]$ gamma $\in [0.01, 0.1, 1, 10, 100, 10^3, \mathbf{10^4}, 10^5, 10^6, 10^7]$ → polynomial: degree $\in [1, 2, 3, 4]$
RF	Number of trees $\in [5, 10, \mathbf{50}, 100, 200]$ Quality split criterion: gini or entropy
MLP	Number of hidden layers $\in [1, \mathbf{2}, 5]$ Number of perceptrons per layer $\in [1, \mathbf{2}, 5, 10, 20]$ Activation function \in [identity, logistic sigmoid, hyperbolic tan , rectified linear unit]

From there, the best set of parameters (in bold in Table 4.1) was retained for each method. It is important to note that some parameters had little or no influence on the performances. When several

values were suitable, we chose to keep the one with the lowest computational time.

3.3 Performances of the eye state estimation method

The eye state estimation algorithm was evaluated by comparing its results with a manual analysis of the videos. On the one hand, video durations and sampling rate implied that a visual scoring frame by frame was not possible. A scoring with another resolution (for example one value per second) was also rejected because it could avoid some short events (as eye blinking). For these reasons, we chose to perform a scoring of 5% of randomly selected frames. On the other hand, intermediate states being difficult to objectively determine, the user decided if the eyes were 'Open' (=1) or 'Closed' (=0). In this context, the values of the surfaces provided by the algorithm (Figure 4.5(c)) were binarized i.e. all the non-zero values were set to 1 (i.e. "Open").

Considering the manual scoring as the reference, the Sensitivity (Se), Specificity (Sp) and Accuracy (Acc) of the proposed method were computed for the 10 videos and are reported in Table 4.2.

Table 4.2 – Newborn data (number, GA and PMA in weeks+days). Video data (duration in min'sec, number of frames visually scored and their repartition 'Open'/'Closed'). Algorithm's performance (sensitivity, specificity, accuracy in %). Number of user interactions (total number and regardless hidden eye).

N°	NEWBORNS		VIDEOS				PERFORMANCES			NB OF INTERACTIONS	
	GA (w+d)	PMA (d)	Duration (min'sec)	Nb of frames visually scored	Nb of frames 'Open'	Nb of frames 'Closed'	Se (%)	Sp (%)	Acc (%)	Total number	Regardless hidden eye
1	28+4	29+6	17'27	1384	40	1344	98.21	99.68	99.64	2	2
2	28+4	29+4	31'50	2367	343	2024	99.71	99.75	99.75	1	1
3	28+4	29+4	30'58	2357	132	2225	97.73	99.87	99.75	2	2
4	28+6	30+0	27'10	2105	56	2049	98.21	100.00	99.11	0	0
5	27+0	28+4	10'11	831	0	831	-	100.00	100.00	25	2
6	28+6	29+6	24'04	1634	146	1478	99.32	100.00	99.66	0	0
7	30+6	32+0	15'07	1198	28	1170	78.57	99.74	99.25	43	37
8	26+0	27+1	41'02	3064	4	3060	100.00	100.00	100.00	0	0
9	31+3	32+4	16'03	1191	584	607	95.38	97.69	96.56	54	25
10	29+5	30+6	28'22	2156	4	2152	100.00	100.00	100.00	0	0

Results show that accuracies range from 96.56% to 100% (99.4% on average). More precisely, sensitivity and specificity are always greater than 95% and 97% respectively, except in the case of the video 7 where Se is equal to 78.57%. In this video, the baby had very rapid motions that led to tracking errors.

The total number of user interactions (Table 4.2) has also been quantified and is supplemented by the number of them not due to hidden eye. We can also notice that most of time only few interactions were required (often none and up to two) except for three videos (5, 7 and 9). For video 5, the total number of interactions is consistent since it appeared mostly in case of hidden eye. However, processing of videos 7 and 9 requested irrelevant manual interactions due to the change of appearance of the ROI_{Ref} , for example after a rotary motion of the head. Nevertheless, performances on video 7 and 9 are high with a global concordance of 96.56% and 99.25%, respectively.

Computational time of our approach is attractive since the algorithm takes 0.047 second to process one frame, in its current version. As an example, the video 8 (duration: 41'02), that required no interaction (except the initialization step), was processed in 48 minutes and 12 seconds. In case of interaction,

time to resume the algorithm is equivalent to the initialization step duration, meaning a few seconds. In addition, several videos can be processed simultaneously optimizing considerably the time required to assess the sleep of different babies.

3.4 Results of sleep stage classification from extracted descriptors

3.4.1 Descriptor analysis

As a first step, the distribution of the values (mean \pm std) of signals $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$, obtained on the whole database, as functions of the sleep stages provided by the expert, are reported in Figure 4.6.

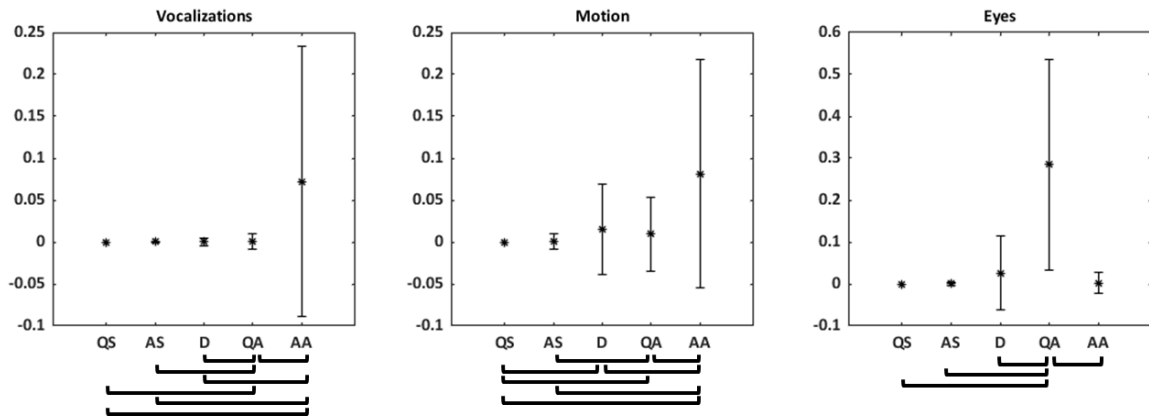


Figure 4.6 – Values (mean \pm std) of signals $\bar{V}(t)$, $\bar{M}(t)$ and $\bar{E}(t)$ as functions of the sleep stages: Quiet sleep (QS), Active Sleep (AS), Drowsiness (D), Quiet Alert (QA) and Active Alert (AA). Pairwise comparison among sleep stages that revealed statistically significant differences ($p < 0.05$) by Mann-Whitney U test are identified by brackets.

We can observe that there are no vocalization in QS and AS, very low amplitudes in D and QA, and higher values and dispersion in AA. Motion values are null in QS, very low in AS, moderate in D and QA, and very high in AA.

Results are different for eye state. Eyes are coherently closed in sleep stages QS and AS. In D, values have a low mean amplitude with a moderate standard deviation, corresponding to short openings. Highest values and dispersion of the values are observed in QA since infant eyes are most of the time opened, but can also be closed. In AA, values are low, because a newborn closes most of the time the eyes when he is nervous or while he is crying (see right picture in Figure 4.4).

Statistical analyses were also conducted by Mann-Whitney U test to discuss pairwise differences between sleep stages for each descriptor. Bootstrap method [51] was applied to minimize the repetitive effect of our dataset. Hence, 100 draws of 24 random samples (representing only 0.2% of the less represented class) have been achieved. The resulting median p -values were studied. Only statistically significant differences ($p < 0.05$) are identified by brackets in Figure 4.6. These results confirm that our set of descriptors is valuable to characterize sleep since most of sleep states can be differentiated

from others by at least one descriptor. We can also note that vocalization and motion features are discriminating in more cases (7 over 10) than eyes (4 only). However, no descriptor showed statistical differences for QS vs AS.

Figure 4.6 shows the complementarity of the three informations, since the value repartition according to the sleep stages is different from a modality to another. Moreover, they are in accordance with the stage definitions for the newborns [43] and give a qualitative validation of the approach. However, they, as of now, augur potential difficulties to differentiate Active Sleep and Quiet Sleep.

3.4.2 Classification results

Performances of the sleep stage classification are evaluated taking as reference the manual scoring performed by the expert. The results of the 30 repeated operations were averaged, which led to a mean accuracy and standard deviation for each sleep stage.

Results presented in Figure 4.7(a) show first that the five classification methods have greater accuracy values for the alert stages (QA and AA). Results with KNN and LDA are fluctuating with high standard deviations observed for some stages (e.g., D, QS and AS for KNN or QS and AS for LDA). To a lesser extent, the same observation can be made for MLP in QS and AS. Although they are closed to SVM results, best performances are obtained with Random Forest for QA and AA, with 93.5% and 99.0% of accuracy respectively, while the results for calm stages (QS, AS and D) are weaker (under 84.1%).

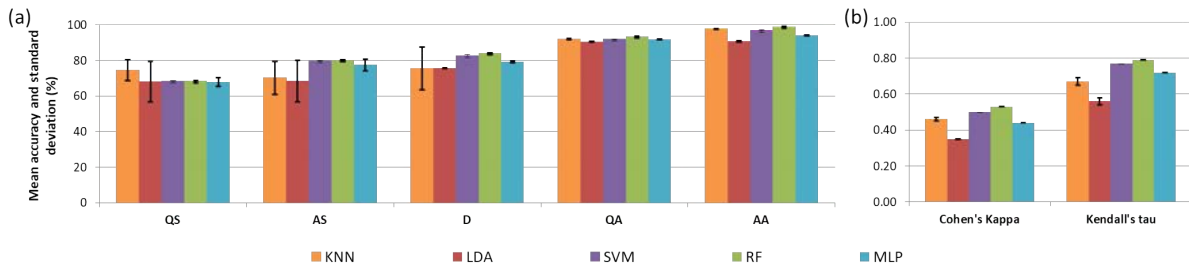


Figure 4.7 – Five machine learning methods are compared: KNN, LDA, SVM, RF and MLP. (a) Classification mean accuracy (%) and standard deviation for each sleep stage: Quiet sleep (QS), Active Sleep (AS), Drowsiness (D), Quiet Alert (QA) and Active Alert (AA); (b) Cohen's Kappa and Kendall's tau coefficients (mean \pm std).

To complement these results, Cohen's Kappa [11] and Kendall's tau [25] coefficients have been computed and reported in Figure 4.7(b).

Cohen's Kappa coefficient, that measures a ratio-scaled degree of disagreement between two approaches, shows, with greater values than 0.44 for all methods except LDA, a moderate concordance between sleep stages provided by the expert and the ones automatically estimated [45]. Kendall tau coefficient aims to measure the association between two quantities, assessing the similarity between ordered data. For this purpose, each stage was affected with a value from 1 (QS) to 5 (AA). For each machine learning approach, excluding LDA, high degrees of concordance are observed (above 0.67), meaning that the estimation errors are mainly made from one stage to another close one (e.g., AS was

estimated instead of QS). As proof, a test considering only two classes, QS+AS+D versus QA+AA (in other terms calm vs alert stages), led to a higher accuracy of 94.8% with Random Forest classifier.

In conclusion, all these results suggest that the estimation of alert stages (QA and AA) is correctly performed, but that the differentiation between the three calm stages (QS, AS and D) remains more difficult. These results are not surprising considering that Drowsiness is an intermediate state by definition, and that QS and AS are close in terms of behavior, as shown by Figure 4.6.

4 Discussion

In this paper, a whole process was defined to semi-automatically and contactless monitor premature newborns using audio-video acquisitions. It includes different processing to extract baby's vocalizations, motion and eye state. For the last one, a specially developed algorithm based on a two-step approach was evaluated with manual annotations of 10 videos and led to a mean accuracy of 99.4%. Weaknesses have been pointed out in the presence of rapid motion and/or complex transformation of the reference region.

Then, the three descriptors were used in order to obtain an estimation of the behavioral sleep states. Five classifiers (KNN, LDA, SVM, RF and MLP) were compared to a NIDCAP expert annotation. Best results were obtained with Random Forest for the two alert stages QA and AA.

Results presented in this paper are new since no similar approach was proposed in the literature in the context of NICU. In fact, no automatic classification of Prechtl sleep stages was ever conducted in preterm infants. All studies dealing with early days of preterm newborns were based on EEG analyses and only quiet sleep detection was performed [5, 13, 14]. Only two studies dealing with preterm and full-term newborns proposed to identify four sleep states with EEG, but was focused on newborns at 38 to 42 weeks PMA [41, 42]. Alternatively, an automatic classification of three states (sleep, awake and crying) was proposed from face analysis [21]. However, authors reported that its application in a realistic hospital environment was not directly possible. In addition, to date, no automated video analysis of sleep has been conducted on preterm infants [50]. The same observation can be made concerning audio analyses. Furthermore, regardless of the clinical target, the combination of audio and video descriptors is also innovative. Despite a wide variety of publications about audio or video processing in paediatrics, only one study integrating both automatically, were, to our knowledge, published [34]. Nevertheless, they were investigated separately.

If this preliminary study shows encouraging results, they will have to be confirmed on a larger database. Although, it is worthwhile to remind that the constitution of such a database is difficult, notably because the annotation of sleep stages by an expert is time consuming.

In the present study, algorithms are applied off-line, on video recordings. The manual interactions for eye state estimation are only required when needed (e.g., occlusions). In that case, the processing is paused. Thus, there is no need for the user nor to perform the analysis in newborn rooms nor to continuously supervise the processing. Consequently, in a heavy workload context for nurses, the sleep analysis can be deferred and thus more newborns may benefit from this follow-up. However, refinements can be envisaged to enhance performances and move towards a fully automated solution. For example, the eye state detection algorithm robustness may be improved by tracking several regions of interest. In

addition, an automatic selection of region(s) of interest by the use of a deep learning approach could be considered on a larger database.

Moreover, it is important to note that the level of discomfort induced to the baby by such a strategy is lower or equivalent to actual techniques but its impact, in both forms (semi-automatic/deferred or automatic/continuous), on daily care routine will have to be studied.

Additionally, a better differentiation between Quiet Sleep and Active Sleep may be achieved by adding the cardio-respiratory information, since it has been shown to be discriminative in Quiet Sleep [20]. However, rather than using additional sensors, it would be doubtless preferable to pursue a non-invasive strategy. Indeed, heart rate and respiration were recently estimated by automatic video processing in NICU in real conditions [10, 47].

Nonetheless, these results augur well for the automatic sleep organization assessment to improve newborn care, but also infant well-being and development. Indeed, this work shows the relevance of our approach to estimate sleep stages by the means of non-invasive techniques such as audio and video processing. These results are directly linked to Digi-NewB objectives and suggest the possibility to monitor sleep in premature newborns and, thus, to quantify their neuro-behavioral development *ex-utero*.

5 Conclusion

This work showed the great interest to combine audio and video analyses for a clinical purpose and mainly for maturation quantification in preterm newborns. From this initial and positive experience, we decided to focus our attention on the remaining part of this dissertation on motion and cry analyses. In fact, the next step to ensure continuous newborn monitoring feasibility is to propose a first fully-automated solution and, as it was confirmed by the methodological state of the art, motion and cry analyses emerge as an obvious basis to this end.

However, as mentioned previously, solutions need to be developed to propose a robust and reproducible behavioral sleep monitoring. In fact, in order to reach continuous monitoring, unpredictable events such as the presence of adults (e.g., parents, nurses) in the camera field, equipment alarms or adults voices on the audio track have to be detected and discarded to extract relevant periods of analyses.

Besides, in this chapter sleep states estimation was performed on each sample although providing estimation on longer period, could, as an example, enhance our capacity to discriminate Quiet Sleep and Active Sleep. In that case, set of parameters to characterize periods, here regarding motion and cry activities, would have to be computed.

In addition, integration of an audio/video system in NICU has to be tackled. In fact, the present study was only performed in 10 different configurations for newborns with a PMA comprised between 28 and 33 weeks, mostly lying in radiant warmer. Nevertheless, in the clinical context chapter, we have seen that NICUs regroup a variety of configurations and equipment. Thus, a new monitoring and adaptable solution based on a new type of care equipment will have to deal with this diversity. This last point is the object of the next chapter where our audio/video acquisition system and its integration in NICU are presented.

Bibliography

- [1] AL-RAHAYFEH, A., AND FAEZIPOUR, M. Eye tracking and head movement detection: A state-of-art survey. *IEEE Journal of Translational Engineering in Health and Medicine* 1 (2013).
- [2] ALS, H. *Program guide: Newborn individualized developmental care and assessment program (NIDCAP): An education and training program for health care professionals*. Boston, MA: Children's Medical Center Corporation, 2002.
- [3] ANDERS, T. F., EMDE, R. N., AND PARMELEE, A. H. *A manual of standardized terminology, techniques and criteria for scoring of states of sleep and wakefulness in newborn infants*. Los Angeles: UCLA Brain Information Service, NINDS Neurological Information Network, 1971.
- [4] ANDERS, T. F., AND SOSTEK, A. M. The use of time lapse video recording of sleep-wake behavior in human infants. *Psychophysiology* 13, 2 (1976), 155–8.
- [5] ANSARI, A. H., DE WEL, O., LAVANGA, M., CAICEDO, A., DEREYMAEKER, A., JANSEN, K., VERVISCH, J., DE VOS, M., NAULAERS, G., AND VAN HUFFEL, S. Quiet sleep detection in preterm infants using deep convolutional neural networks. *Journal of Neural Engineering* 15, 6 (2018), 066006.
- [6] BLENCOWE, H., COUSENS, S., OESTERGAARD, M. Z., CHOU, D., MOLLER, A.-B., NARWAL, R., ADLER, A., GARCIA, C. V., ROHDE, S., SAY, L., AND LAWN, J. E. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet* 379, 9832 (2012), 2162–72.
- [7] BREIMAN, L. Random forests. *Machine Learning* 45, 1 (Oct 2001), 5–32.
- [8] CABON, S., PORÉE, F., SIMON, A., MET-MONTOT, B., PLADYS, P., ROSEC, O., NARDI, N., AND CARRAULT, G. Audio- and video-based estimation of the sleep stages of newborns in neonatal intensive care unit. *Biomedical Signal Processing and Control* 52 (2019), 362–370.
- [9] CABON, S., PORÉE, F., SIMON, A., ROSEC, O., PLADYS, P., AND CARRAULT, G. Video and audio processing in paediatrics: a review. *Physiological Measurement* 40(2) (2019), 1–20.
- [10] CATTANI, L., ALINOV, D., FERRARI, G., RAHELI, R., PAVLIDIS, E., SPAGNOLI, C., AND PISANI, F. Monitoring infants by automatic video processing: A unified approach to motion analysis. *Computers in Biology and Medicine* 80 (2017), 158–165.
- [11] COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.
- [12] CURZI-DASCALOVA, L., AND MIRMIRAN, M. *Manual of methods for recording and analyzing sleep-wakefulness states in preterm and full-term infant*. INSERM, Paris, 1996.
- [13] DE WEL, O., LAVANGA, M., DORADO, A. C., JANSEN, K., DEREYMAEKER, A., NAULAERS, G., AND VAN HUFFEL, S. Complexity analysis of neonatal EEG using multiscale entropy: applications in brain maturation and sleep stage classification. *Entropy* 19, 10 (2017), 516.

- [14] DEREYMAEKER, A., PILLAY, K., VERVISCH, J., VAN HUFFEL, S., NAULAERS, G., JANSEN, K., AND DE VOS, M. An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation. *International Journal of Neural Systems* 27, 06 (2017), 1750023.
- [15] DÍAZ, M. A. R., GARCÍA, C. A. R., ROBLES, L. C. A., ALTAMIRANO, J. E. X., AND MENDOZA, A. V. Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis. *Biomedical Signal Processing and Control* 7, 1 (2012), 43–49.
- [16] FRAIWAN, L., LWEESY, K., KHASAWNEH, N., FRAIWAN, M., WENZ, H., AND DICKHAUS, H. Time frequency analysis for automated sleep stage identification in fullterm and preterm neonates. *Journal of Medical Systems* 35, 4 (2011), 693–702.
- [17] FRANKLIN, A., PILLING, M., AND DAVIES, I. The nature of infant color categorization: Evidence from eye movements on a target detection task. *Journal of Experimental Child Psychology* 91 (2005), 227–248.
- [18] FULLER, P. W., WENNER, W. H., AND BLACKBURN, S. Comparison between time-lapse video recordings of behavior and polygraphic state determinations in premature infants. *Psychophysiology* 15, 6 (1978), 594–8.
- [19] GREDEBACK, G., AND VON HOFSTEN, C. Infants' evolving representations of object motion during occlusion: A longitudinal study of 6- to 12-month-old infants. *Infancy* 6 (2004), 165–184.
- [20] HARPER, R. M., SCHECHTMAN, V. L., AND KLUGE, K. A. Machine classification of infant sleep state using cardiorespiratory measures. *Electroencephalography and Clinical Neurophysiology* 67, 4 (1987), 379–387.
- [21] HAZELHOFF, L., HAN, J., BAMBANG-OETOMO, S., AND DE WITH, P. H. N. Behavioral state detection of newborns based on facial expression analysis. In *International Conference on Advanced Concepts for Intelligent Vision Systems* (2009), Springer, pp. 698–709.
- [22] HUNNIUS, S., AND GEUZE, R. H. Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. *Infancy* 6 (2004), 231–255.
- [23] HUVANANDANA, J., THAMRIN, C., TRACY, M., HINDER, M., NGUYEN, C., AND MCEWAN, A. Advanced analyses of physiological signals in the neonatal intensive care unit. *Physiological Measurement* 38, 10 (2017), R253.
- [24] JOHNSON, S. P., SLEMMER, J. A., AND AMSO, D. Where infants look determines how they see: Eye movements and object perception performance in 3-month-olds. *Infancy* 6 (2004), 185–201.
- [25] KENDALL, M. G. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
- [26] MANFREDI, C., BANDINI, A., MELINO, D., VIELLEVOYE, R., KALENGA, M., AND ORLANDI, S. Automated detection and classification of basic shapes of newborn cry melody. *Biomedical Signal Processing and Control* 45 (2018), 174–181.

-
- [27] MARCROFT, C., KHAN, A., EMBLETON, N. D., TRENELL, M., AND PLOTZ, T. Movement recognition technology as a method of assessing spontaneous general movements in high risk infants. *Frontiers in Neurology* 5 (2014), 284.
- [28] MIKA, S., RATSCH, G., WESTON, J., SCHOLKOPF, B., AND MULLERS, K.-R. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.* (1999), IEEE, pp. 41–48.
- [29] MOESLUND, T. B., HILTON, A., AND KRÜGER, V. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding* 104, 2-3 (2006), 90–126.
- [30] OKADA, S., OHNO, Y., KATO-NISHIMURA, K., MOHRI, I., TANIKE, M., ET AL. Examination of non-restrictive and non-invasive sleep evaluation technique for children using difference images. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2008), IEEE, pp. 3483–3487.
- [31] ORLANDI, S., BOCCHI, L., DONZELLI, G., AND MANFREDI, C. Central blood oxygen saturation vs crying in preterm newborns. *Biomedical Signal Processing and Control* 7, 1 (2012), 88–92.
- [32] ORLANDI, S., DEJONCKERE, P. H., SCHOENTGEN, J., LEBACQ, J., RRUQJA, N., AND MANFREDI, C. Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring. *Biomedical Signal Processing and Control* 8, 6 (2013), 799–810.
- [33] ORLANDI, S., GARCIA, C. A. R., BANDINI, A., DONZELLI, G., AND MANFREDI, C. Application of pattern recognition techniques to the classification of full-term and preterm infant cry. *Journal of Voice* 30, 6 (2016), 656–663.
- [34] ORLANDI, S., GUZZETTA, A., BANDINI, A., BELMONTI, V., BARBAGALLO, S. D., TEALDI, G., MAZZOTTI, S., SCATTONI, M. L., AND MANFREDI, C. AVIM - A contactless system for infant data acquisition and analysis: Software architecture and first results. *Biomedical Signal Processing and Control* 20 (2015), 85–99.
- [35] ORLANDI, S., MANFREDI, C., BOCCHI, L., AND SCATTONI, M. Automatic newborn cry analysis: A non-invasive tool to help autism early diagnosis. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (2012), IEEE, pp. 2953–2956.
- [36] OTSU, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9, 1 (1979), 62–66.
- [37] PAL, S. K., AND MITRA, S. Multilayer perceptron, fuzzy sets, classification.
- [38] PEDIADITIS, M., TSIKNAKIS, M., AND LEITGEB, N. Vision-based motion detection, analysis and recognition of epileptic seizures—a systematic review. *Computer Methods and Programs in Biomedicine* 108, 3 (2012), 1133–48.

- [39] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [40] PETERSON, L. E. K-nearest neighbor. *Scholarpedia* 4, 2 (2009), 1883.
- [41] PILLAY, K., DEREYMAEKER, A., JANSEN, K., NAULAERS, G., VAN HUFFEL, S., AND DE VOS, M. Automated eeg sleep staging in the term-age baby using a generative modelling approach. *Journal of neural engineering* 15, 3 (2018), 036004.
- [42] PIRYATINSKA, A., TERDIK, G., WOYCZYNSKI, W. A., LOPARO, K. A., SCHER, M. S., AND ZLOTNIK, A. Automated detection of neonate EEG sleep stages. *Computer Methods and Programs in Biomedicine* 95, 1 (2009), 31–46.
- [43] PRECHTL, H. F. The behavioural states of the newborn infant (a review). *Brain research* 76, 2 (1974), 185–212.
- [44] RECHTSCHAFFEN, A., AND KALES, A. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Los Angeles: UCLA Brain Information Service/Brain Research Institute, 1968.
- [45] SCATENA, M., DITTONI, S., MAVIGLIA, R., ET AL. An integrated video-analysis software system designed for movement detection and sleep analysis. Validation of a tool for the behavioural study of sleep. *Clinical Neurophysiology* 123, 2 (2012), 318–23.
- [46] STAHL, A., SCHELLEWALD, C., STAVDAHL, O., AAMO, O. M., ADDE, L., AND KIRKEROD, H. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20, 4 (2012), 605–14.
- [47] VAN GASTEL, M., BALMAEKERS, B., OETOMO, S. B., AND VERKRUYSSE, W. Near-continuous non-contact cardiac pulse monitoring in a neonatal intensive care unit in near darkness. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics* (2018), vol. 1050114, International Society for Optics and Photonics, pp. 1–9.
- [48] VAPNIK, V., AND MUKHERJEE, S. Support vector method for multivariate density estimation. In *Advances in neural information processing systems* (2000), pp. 659–665.
- [49] WASZ-HÖCKERT, O., MICHELSSON, K., AND LIND, J. Twenty-five years of Scandinavian cry research. In *Infant Crying*. Springer, 1985, pp. 83–104.
- [50] WERTH, J., ATALLAH, L., ANDRIESSEN, P., LONG, X., ZWARTKRUIS-PELGRIM, E., AND AARTS, R. M. Unobtrusive sleep state measurements in preterm infants—a review. *Sleep Medicine Reviews* 32 (2017), 109–122.
- [51] ZOUBIR, A. M., AND BOASHASH, B. The bootstrap and its application in signal processing. *IEEE signal processing magazine* 15, 1 (1998), 56–76.

VOXYVI, A NEW AUDIO-VIDEO ACQUISITION SYSTEM FOR NEONATAL INTENSIVE CARE UNIT

1 Introduction

One of the main objectives of the Digi-NewB project is to construct a large database of synchronized signals of three different sources: electro-physiological (e.g, heart or respiratory rate), video and audio, in order to identify the best features that characterize the newborn condition regarding infection and neuro-behavioral development. Therefore, during the four years of the project, a large number of babies have to be included and recorded at different PMA. In particular, a rare incidence rate of sepsis is observed in newborns, confirming that a high number of babies will have to be recorded to reach a good representation of sepsis. To achieve this, six french hospital centers are involved in the project: Rennes, Brest, Nantes, Angers, Poitiers and Tours. However, each NICU has its special characteristics regarding room configurations, beds brands but also care habits.

The main challenge, addressed in this chapter, is then to design an audio and video acquisition system that fits all these characteristics while being as non-invasive as possible, within the objectives, first, to construct a homogeneous database and later, to be integrated in daily care as an equipment for monitoring. In the literature, studies dealing with preterm newborns were mainly conducted on short recordings on relatively small databases and thus, solutions to settle in NICU were not clearly addressed. Most of the authors decided to place the acquisition system above one corner of the foot end of the bed in order to record the full body of the baby. However, no author recorded newborns lying neither in closed incubators nor in a dark environment, particularly common in NICU.

In addition, several features extracted from audio and video have been proven to be relevant indicators of newborn conditions regarding a large scope of clinical situations. Hence, Black and White (B&W) cameras were used for motion analyses in order to detect cerebral palsies [12] or neonatal seizures [8] but also to estimate physiological signals such as cardiac pulse [13] or respiratory [4] rates. The last two points were also investigated with color cameras [3, 14]. Concerning audio analyses, two types of microphones have been used: unidirectional for cry analysis [9] or omnidirectional to study NICU acoustic environments [11]. In addition, thermal cameras were also employed to investigate temperature of the baby [2] or respiratory rate [1]. However, among them, no study dealt with neonatal sepsis.

Thus, it was difficult to know, in advance, which descriptor will be revealed more relevant regarding this objective. Reversely, considering behavioral developments, assessing baby cries and movements from microphones and B&W camera were shown as important elements to predict different brain disorders. Consequently, to construct an exhaustive and informative database for research purpose, a variety of audio and video modalities had to be considered. Thus, different constraints had to be fulfilled by the system both in terms of considered modalities (B&W, color, and thermal cameras, microphones) and acquisition conditions (very long recordings in very constrained and varying environments). Since no existing devices were meeting these conditions, it was decided to develop a specific system.

This chapter is organized in two main sections. First, a material and methods section presenting the acquisition system (Voxyvi) in terms of hardware and software, solutions to integrate NICU and the data acquisition protocol are proposed. Then, in a second section, the impact of the new material in NICU is studied through the inclusion progress and clinician feedback.

2 Materials and methods

2.1 Description of the acquisition system

The acquisition system is composed of an hardware and a software parts: the audio-video device and the local computer unit which provides an interface, computing capabilities and data recording. The whole system will deliver, after pre-processing, a set of descriptors that can be integrated in a Decision Support System (DSS) in order to monitor newborn health. This architecture is depicted by Figure 5.1.

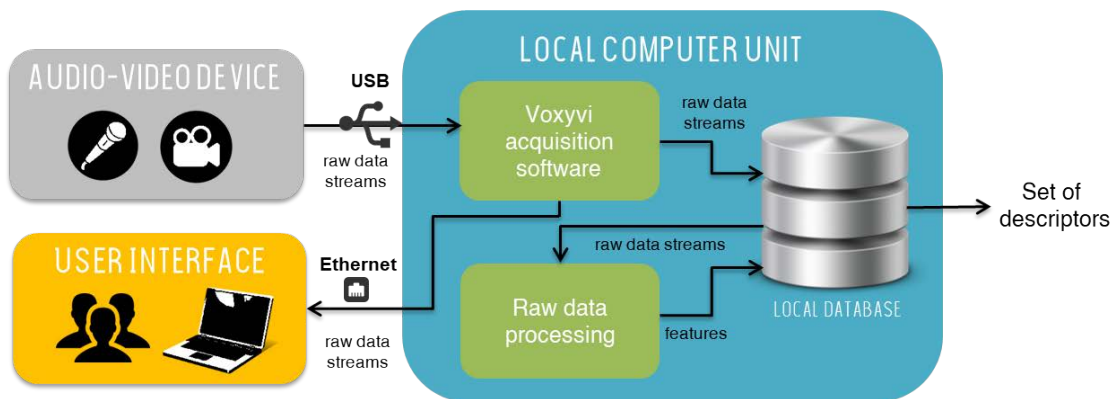


Figure 5.1 – Main framework design of the Voxyvi acquisition system.

2.1.1 Audio-video device

Three video components marketed by FLIR have been selected. First, a B&W infrared camera that presents a resolution of 752x480 pixels (FMVU-03MTM-CS) was chosen. Associated with an infrared illumination by LEDs, it provides clear images day or night up to 30 frames per second (fps). Its small size (44 mm x 34 mm x 24.4 mm) makes it very easy to integrate. This first component was considered

as susceptible to give information on newborn motion. The second video component is the equivalent color camera (FMVU-03MTC-CS). However, images are clear only by day or when lights are on. At the beginning of the project, this camera was considered as pertinent to bring information on the skin color. The last one is a thermal camera (60x80 pixels) with up to 8.6 fps capturing capability. There too, its size (8.5 x 11.7 x 5.6 mm) motivated its selection. This camera was expected to bring information about the presence of the baby in the bed, eventually his/her motion or other parameters like respiration or temperature.

Regarding acoustic, an omni-directional microphone (FG-23329-P07) marketed by Knowles Acoustics was chosen. It allows an acquisition sampling rate up to 44 KHz. This microphone was judged as robust enough for very long acquisitions in real time context and susceptible to capture relevant information about the cry of the newborns.

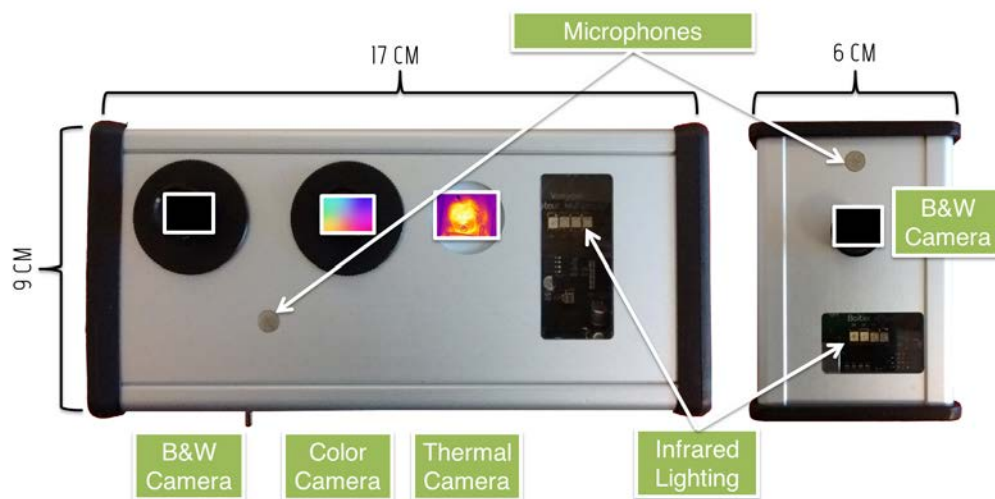


Figure 5.2 – Composition of the acquisition device.

The different components have been embedded in a single acquisition device (see Figure 5.2) that provides only two connections: one connection to transfer all the data and another one for power supply. The device is composed by two housings: a main housing that is self-dependent and a secondary housing provided to enable B&W acquisition with a second view.

The main housing contains all the above mentioned components meaning an infrared B&W camera, a color camera, a thermal camera and a microphone. Moreover, other components such as infrared illumination and informative sensors (e.g., temperature, infrared illumination, ambient light photo sensor) have been added. The second housing only embeds an infrared B&W camera with its associate infrared illumination and a microphone. The composition of each housing is summarized in Table 5.1.

The NICU environment implies different conception constraints. In fact, the room being cluttered by a variety of medical equipment, the size of acquisition device has been minimized. Moreover, in some cases (cf. below), it has to be positioned inside the incubator. The total dimensions of the main housing are 17x9x4cm whereas the secondary housing is smaller (6x9x4cm).

An intense infrared illumination can be dangerous, even more because of its invisibility. The thermal

Table 5.1 – Description of the composition of each housing

Component	Main housing	Secondary housing
Dimensions	17x9x4cm	6x9x3cm
B&W infrared camera	✓	✓
Infrared lighting	✓	✓
Color camera	✓	-
Thermal camera	✓	-
Microphone	✓	✓
Infrared sensor	✓	-
Ambient light photo sensor	✓	-
Temperature sensor	✓	-
Humidity indicator	✓	✓

effect linked to infra-red illumination can lead to ocular or skin lesions. To protect the preterm infant's eyes, the maximum Near-Infrared (NIR) radiation must be limited to 10 mW/cm^2 [15]. In our case, with the newborns' eyes placed at least at 30 centimeters from the housing, the power for one steradian will be allocated on a surface of 1225 cm^2 , resulting to a total illumination of $29.4 \mu\text{W/cm}^2$. Thus, the total radian illumination is well below the 10 mW/cm^2 recommendation.

In addition, acquisition have to be performed into incubators. In this type of bed, the temperature can reach 40°C and the humidity rate up to 90%. This configuration implied a particular attention regarding the choice of the sensors but also to the conception of the protection of the electrical board. Besides, the housings are waterproof and machined aluminum. To control humidity infiltration, an indicator has been integrated in each housing.

From our instructions, the housings have been manufactured by a company specialized in electronic integration.

2.1.2 Local computer unit

The software requirements are fulfilled by an Intel Next Unit of Computing (NUC), depicted by Figure 5.3, a mini-computer of $10 \times 10 \text{ cm}$.



Figure 5.3 – Intel Next Unit of Computing.

The audio-video device is directly connected to the NUC using a USB connection. Two main tasks, described below, are managed by the Voxyvi acquisition software: interfacing and data recording. Both parts had been developed by the engineers of Voxygen. In the second phase of the project (continuous monitoring), it will also integrate the computation of features from raw data.

User interface The graphic configuration interface allows the visualization of the raw streams captured by the audio-video device and the launching of a recording session.

To access the interface, a laptop is connected by Ethernet to the NUC. All video streams can be visualized (Fig. 5.4(1)) by selecting the corresponding check box (Fig. 5.4(2)).



Figure 5.4 – Functionalities of the graphic configuration interface: camera images location (1), check boxes to switch between cameras (2), sensors status (3) and start recording button (4).

Other information such as the global luminosity, the temperature or sensor connection status are transmitted in order to control the device before launching the acquisition (Fig. 5.4(3)). Then, it is possible to begin the recording by clicking on the "Start Recording" button (Fig. 5.4(4)). A pop-up window appears and the research identifier (anonymous unique ID of the newborn within Digi-NewB) need to be entered. From there, no additional manipulation is needed and the laptop can be disconnected. In fact, this solution limits the quantity of devices in the room while recording. To end the recording session, accessing the interface is not required, it is sufficient to simply stop the NUC.

Data recording Recordings are stored locally but Voxyvi also allows saving data on an external hard drive disk connected to the NUC. Each stream is recorded under an unique session directory. Moreover, in order to limit file corruption and data loss, recordings are split into multiple records of 30 minutes. Due to the high number of integrated components, a lot of data is collected. Therefore, the encoding formats were carefully chosen in order to keep a good compromise between data quality and file sizes. As a result, all video streams are recorded at 25 fps with MPEG-4 encoding, under AVC container, except for the thermal images. Indeed, in order to keep the integrity of raw measurements, thermal values are directly saved in binary files at 4 fps. Each audio stream is recorded as a channel of a stereo WAV file at 24 KHz. In the end, the resulting amount of data to store per hour is about 1.6 Go.

2.2 Integration in Neonatal Intensive Care Unit

The main challenge of recording newborn in NICU is to integrate Voxyvi in the care routine. Besides, the principal will of this project is to propose monitoring solution that is non-invasive regarding the newborn, as well as medical staff and parents. The second challenge is to build a standardized database to automatize audio and video analyses. All of this is even more challenging in the multicentre study context. In fact, as mentioned previously, each NICU is different in term of service organization, bed brands or care habits. In this section, we only focus on the positioning and the orientation of the audio-video device since the local unit computer is easily placed in an accessible location, usually on a shelf near the bed or with the actual physiological monitor.

2.2.1 Positioning of the audio-video device

The newborn safety is the most important requirement. In fact, the system must be designed to be integrated, in the safer manner, in the newborn environment. Moreover, to ensure a non-invasive integration, it must not require direct interaction with the newborn meaning that it should fit to the environment without modifying it (i.e. no additional electrodes, no need to move the newborn from a location to another). Meanwhile, the device should not disturb the medical staff work in terms of manipulation and congestion. This has an impact on the localization of the device and the way to position it.

In addition, premature newborns are placed in NICU from their birth to their discharge from hospital. This specific care implies different kinds of beds (e.g., closed beds like incubators or open beds such as radiant warmer beds and cradles) corresponding to the newborn needs. Figure 5.5 illustrates three types of bed within three different room configurations in Rennes Neonatal Intensive Care Unit.

This diversity of configuration is also observed in the other hospital centers.

Two types of mounting In response, two types of mounting have been proposed: support bracket and clamps.

The support relies on a base made of a stainless steel acting as a counterweight. This base is slipped underneath the mattress and fits any kind of bed. In order to have an optimized angle of observation, the support integrates a rail that enables moving both housings. To adapt the system position regarding height, the rail is also fixed on two telescopic tubes (Max. height 250mm, min. height 154mm, diameter 27/20mm).

The second strategy is to hang the housings as the other medical devices in the room (i.e., using bars or masts). For that purpose, a solution with clamps (Manfrotto REF035) is proposed. The operational domain of the clamp is on sections comprised between 15 and 55 mm. In that context, the wide variety of possibilities of hanging can led to safety issues if clamps are poorly anchored. To manage this critical situation, a security lanyard is provided.

In addition, the viewing angles of the recording devices are likely to change. Thus, the housings are mounted on ball joints which can be locked down to block the position. These mountings can be fixed on the rail of the support with clamping bolts or directly on the clamp.



Figure 5.5 – Overview of three different types of bed and room configurations in Rennes Neonatal Intensive Care Unit: an incubator (1), a radiant warmer bed (2) and a cradle (3).

A combined solution In order to adapt safely and efficiently the device to the environment, a protocol depending on the type of bed and environment was constructed (Figure 5.6).

- Closed beds

An outside positioning is irrelevant for closed beds, specially for audio acquisition since the baby sounds will be attenuated. To deal with this type of beds, the support bracket located at the feet of the newborn is considered as the best position. This decision has been taken firstly, regarding the technical constraints and for its adaptation to standard care routine. In fact, medical staff and parents rarely accede to the newborns by this side of the incubator. This positioning has the benefit to make sure that the newborn will stay in the field of view if the bed is moved.

- Open beds

In case of open beds, clamps fixed on masts at the head of the newborns have been defined as the best positioning. In that way, the integration in the room is better than a support placed at the feet of the newborn. In fact, to achieve the best view of the newborn, the housings placed at the feet should be very high to see his/her face and it would affect nursing. Some brands of opened incubators have integrated masts as seen in Figure 5.5(2). For other brands of open beds, the clamps can be hanged on one of the masts that may be already present in the room. However, our experience showed that in some rooms none of each is present. In that case, the housing can be hang on a mobile and ballasted mast usually used for subcutaneous infusion. The main issue implied by this solution is to be sure that the bed will remain in the field of view. To overcome this difficulty, the location of the bed is marked on

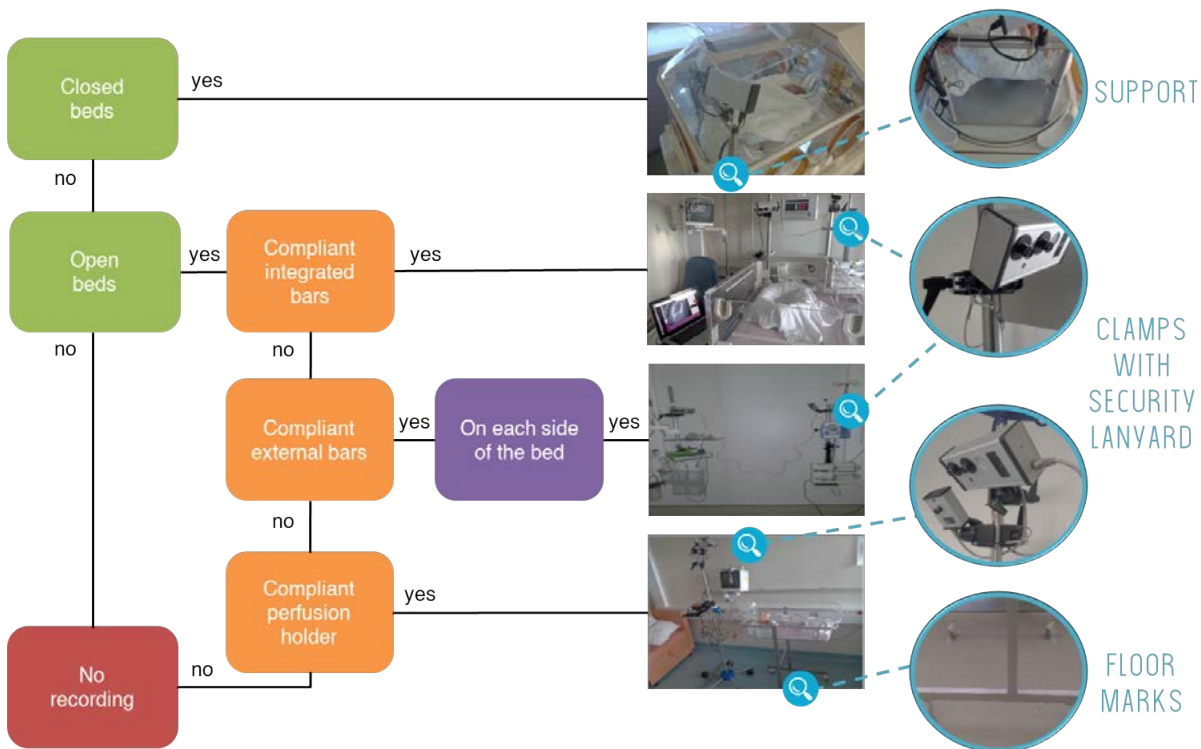


Figure 5.6 – Audio-video device protocol of positioning for NICU.

the floor and medical staff and parents are aware to replace the bed in this position. For security, data standardization and infrared lighting reasons the distance between the bed and the housings must be kept between 30 and 80 centimeters.

2.2.2 Standardization of the device orientation

In order to build a homogeneous database, recommendations regarding the camera orientation were proposed and are illustrated by Figure 5.7. The main one is to orient the camera to visualize the whole baby. For that purpose, we proposed to let a border, in the image, around the newborn. Secondly, the newborn must be located at the center of the image. These recommendations are necessary to be sure that the newborn will stay in the field of view in case of position change in the bed while nursing.

2.2.3 Training

In a heavy workload context for the medical staff, the system has to be user-friendly, meaning, among other things, that this new equipment has to be easy to understand and use. Thus, a training and technical support were introduced.

First of all, an exhaustive user manual has been provided to each research team in charge of the study conduction in the six hospitals. The user guide is reported in Appendix A. It was presented during a day of training conducted in each center. It contains all the previous mentioned requirements (position-

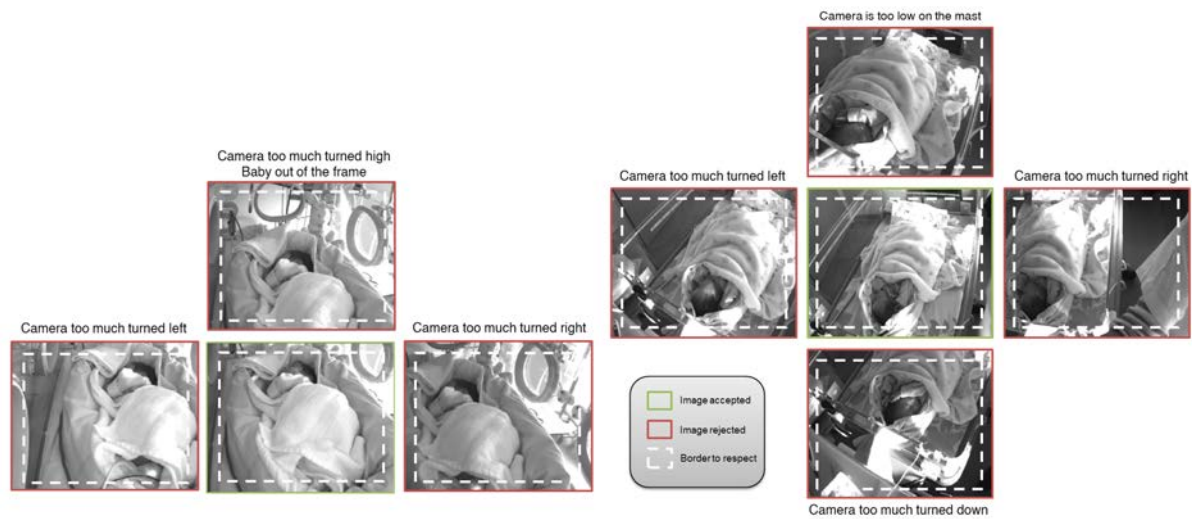


Figure 5.7 – Camera orientation recommendations for open (left) and closed (right) beds.

ing, orientation, access to the interface) but also a step by step description of the connection between the different components of the system. To facilitate the installation of the system in rooms, summarized sheets have been given and wires have been marked with different colors (Figure 5.8). Technical



Figure 5.8 – Color marked wire connection to the computer unit (left) and to the device (right).

support details have also been communicated and allowed a personalized follow-up of difficulties.

Moreover, signage has been added to the system in order to prevent each person (e.g., parents, nurses or cleaners) to the security precautions. As an example, in case of mobile elements (e.g., beds, perfusion holder) the device should not in any way be placed over the newborn. The corresponding signage, positioned on the device, is depicted by Figure 5.9.

2.3 Data acquisition protocol

One goal of the Digi-NewB project is to construct an extensive database regarding two clinical targets: the evaluation of the neurobehavioral development and of the sepsis risk in newborns.



Figure 5.9 – Security signage.

For that purpose, 780 patients should be recorded along the four years of the project, among them 480 preterm newborns susceptible to contract sepsis. Three populations have been retained:

1. 180 premature newborns with a gestational age inferior to 32 weeks;
2. 180 full-term newborns with high sepsis risk
3. 120 premature newborns born between 26 and 34 GA, with birth followed a Premature Rupture of Membranes (PRM) since this population also carries a high risk of infection [10].

This number is supplemented by 300 newborns, for the maturation objective, equally distributed between five categories of GA: extremely, very and late preterm, early term (37-38 weeks GA) and full-term (>39 weeks GA) newborns.

Regarding the infection target, a recording protocol has been planned for each population: 1) 10 consecutive days of recordings in the early days of life, 2) one recording of 6 hours started in the first 24 hours after birth and 3) one recording of three days at birth, followed by recordings of 17 hours every 10 days. In case of maturation, a recording of 24 hours every 10 days until the discharge is required.

To this end, 19 acquisition devices have been allocated between six centers: 4 in Rennes, 4 in Angers, 3 in Nantes, 3 in Brest, 3 in Tours and 2 in Poitiers. Inclusions have been started in November 2016 in Rennes, in April 2017 in Angers, in June 2017 in Nantes, Brest and Tours and in December 2017 in Poitiers.

3 Evaluation of the acquisition system integration in NICU

3.1 Inclusion progress

Figure 5.10 summarizes the inclusions along the period of November 2016 to February 2019 while Figure 5.11 resumes the number of inclusions for each category. In February 2019, 414 newborns have been recorded in the study. Among them, 137 newborns were included both in infection and in maturation protocols.

Presently, 43% of the expected total number of inclusions is reached in infection. The most advance category is the extremely/very preterm with 150 inclusions followed by PRM and high risk fullterm with rates of 47.5% and 17% respectively.

Considering maturation, a total rate of inclusions of 75% is attained. In fact, the extremely, very and late preterm objectives are completed. Early term and full term categories are behind with respectively 23% and 50%.

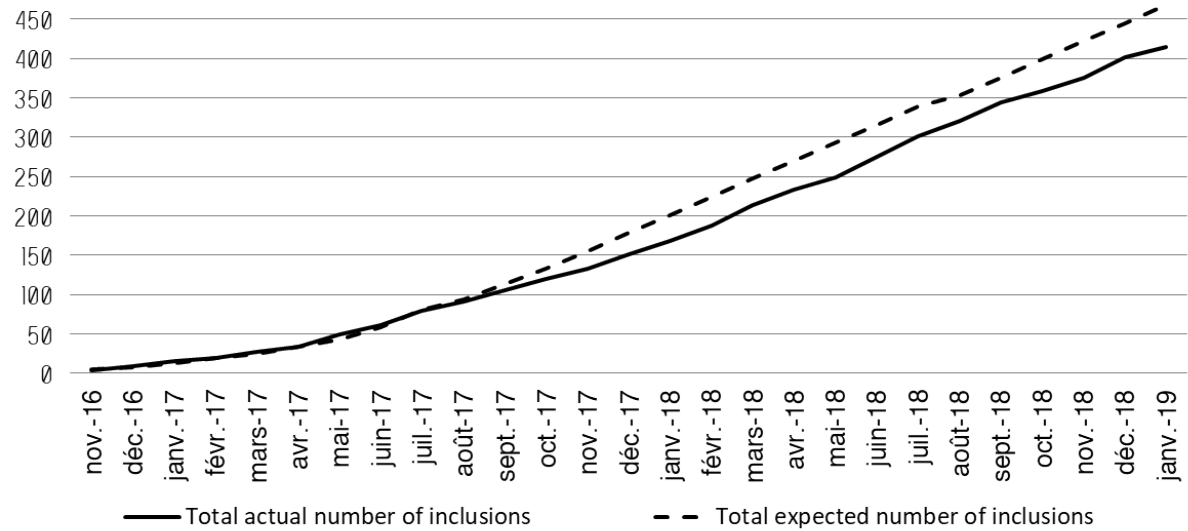


Figure 5.10 – Progress of the total inclusions in Digi-NewB.

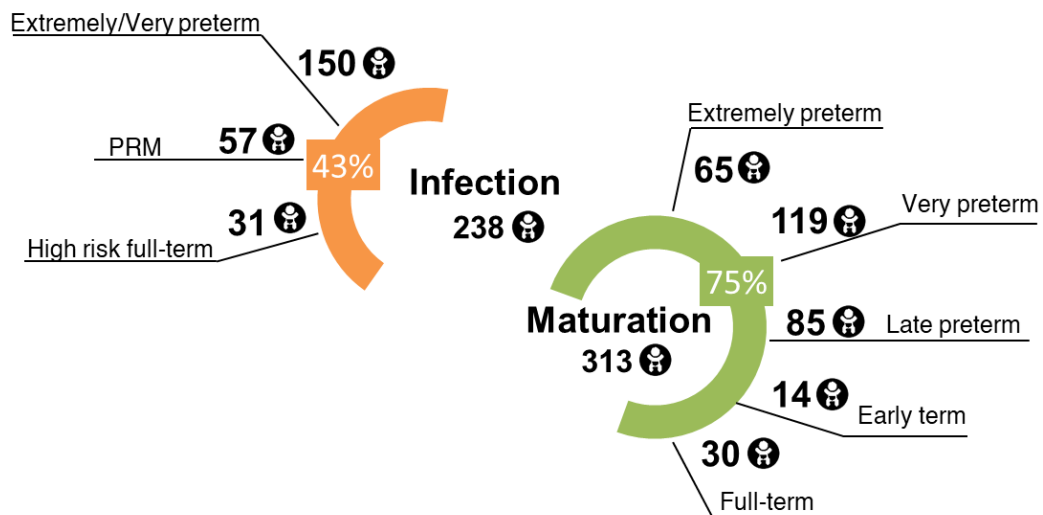


Figure 5.11 – Newborns included regarding each targeted category.

The whole database is composed by 1513 recording sessions. It represents about 37 500 hours of data acquisition giving a total amount of data of 60 To, disseminated between the six centers from 7 To in Angers to 19 To in Rennes.

3.2 Clinicians feedback

A survey was proposed to nine clinicians, with different profiles (five pediatric nurses, two research nurses, one nurse and one clinical research associate), responsible of the acquisition progress in each hospital center. The objective was to gather their feedback regarding four aspects of the system: installation, interface, training and care routine integration. For each point, overall questions were first asked (e.g., how long do you take to install the system?). Then, agreements to diverse propositions were retrieved on a 4-level scale (Agree, Quite Agree, Quite Disagree and Disagree). They are reported hereafter.

3.2.1 Installation

As a first result, clinicians reported that the installation of the system lasted between 15 and 30 minutes and was mostly achieved by one person. Then, collected answers about installation on the 4-level scale are compiled in Figure 5.12.

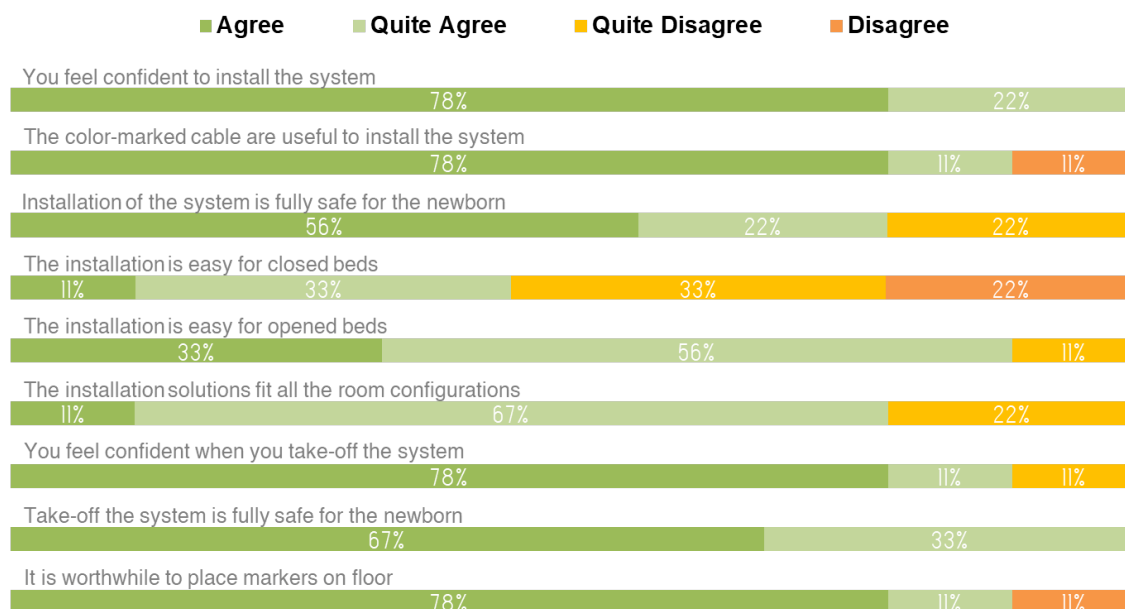


Figure 5.12 – System installation evaluation results on a 4-level scale (Agree, Quite Agree, Quite Disagree and Disagree).

Clinicians mostly consider that the installation solutions fit all room configurations of their units. In

addition, all clinicians feel confident to install the system and mostly consider its installation as fully safe for the newborn except for two of them. Reversely, one clinician don't feel confident to take off the system although all others report that this operation is fully safe for the newborn. The idea to color marked wires is well received since most of the clinicians recognized that this is useful for the installation of the device. However, installation reveals complicated on closed beds for 55% of the clinicians. The opposite trend is seen for open beds. They reported that it was mainly caused by the difficulty to orient the housings in incubators.

3.2.2 User interface

The user interface is fully satisfactory since, as reported in Figure 5.13, all clinicians are comfortable with the proposed use of the laptop to recover the interface.

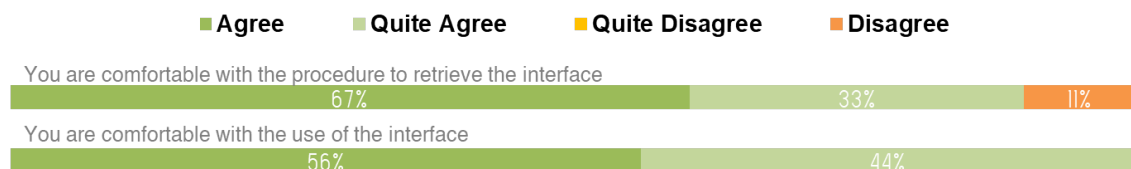


Figure 5.13 – System interface evaluation results on a 4-level scale (Agree, Quite Agree, Quite Disagree and Disagree).

In addition, they are satisfied with all functionalities of the interface (e.g., stream visualizations, start/stop a recording session).

3.2.3 Training

The training evaluation answers of clinicians are summarized in Figure 5.14.

Once again, the overall evaluation is positive. In fact, respondents mostly declare that the one-day training was sufficient to take control on the system and that the supplied manual user was useful and understandable. In fact, they globally do not need to refer it anymore. In addition, the importance of providing a technical support all along the acquisition progress is underlined by all participants. Reversely, mixed reactions are collected about the usefulness of the summarized sheets. Indeed, nearly an half of the clinicians used them at the beginning of the project and almost all of them no more use them.

3.2.4 Care integration

Finally, and undoubtedly the most important point for future considerations, integration of the system in the care routine was evaluated by participants. Answers are reported in Figure 5.15 and show that 78% of clinicians think that the system, in its current form, is quite well integrated to the daily care routine. Until now, the system did not hinder the parents. The signage reveals useful to give the safety

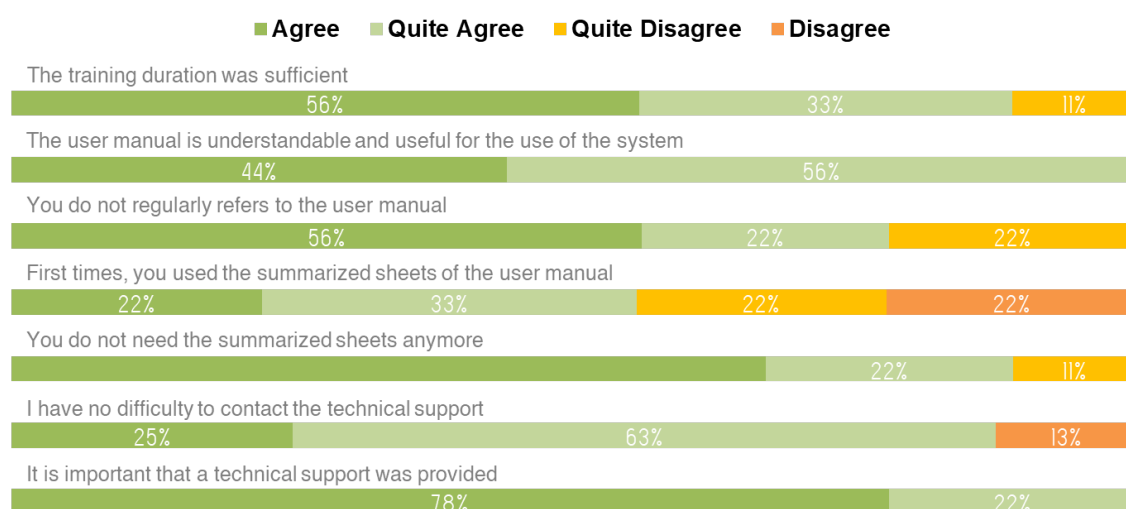


Figure 5.14 – System training evaluation results on a 4-level scale (Agree, Quite Agree, Quite Disagree and Disagree).

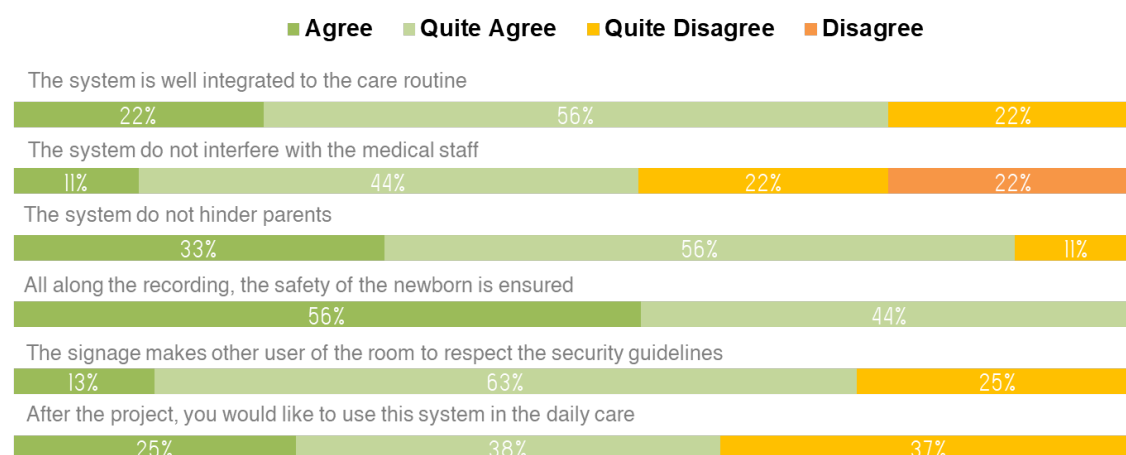


Figure 5.15 – Care routine integration of the system evaluation results on a 4-level scale (Agree, Quite Agree, Quite Disagree and Disagree).

directives to all people and thus, newborns safety is ensured all along recording sessions. However, clinicians raised concerns regarding the system interference with medical staff. First, the system can be cumbersome in case of sensitive medical gestures such as the setup of a central catheter, especially in closed beds. Secondly, medical staff questions the legal rights of using of video and audio in case of trial proceedings. Consequently, mixed feedback are obtained regarding their will to integrate such a system in their daily practice after the project.

3.3 Difficulties

In this section, difficulties observed in our database are presented. Three sources of difficulties have been noticed: hardware dysfunction, non-compliance to the protocol and real conditions.

Hardware dysfunction Several malfunctions during acquisition have been raised up by clinicians to the technical support. Further investigations conducted by the housing manufacturer revealed that the power supply showed weakness in a stressful environment such as in NICU. Unfortunately, this results into unexpected stops of acquisition sessions. Although this problem is now corrected, several recordings were impacted and data, especially around the infected episodes, is missing regarding the inclusion objectives.

Non-compliance to the protocol In one hospital, the protocol regarding the positioning of the device for open beds was not followed. In fact, acquisitions were performed with the support instead of using clamps. In this configuration, the height that can be reached is not sufficient to see the newborn entirely, especially because babies are bigger/older when there are placed in open bed.

In addition, we sometimes observed that the bed was either belatedly or not replaced at the marked spot on the floor leading to unusable periods of recording.

Real conditions Since long duration recordings were slightly investigated in the literature, several difficulties induced by the real conditions of NICU were revealed during the project.

The first point is that babies are usually well covered, sometimes so much so that it is difficult to see where the newborn is under the blanket. We have also observed that they can be very congested by respiratory assistance equipment. In addition, most of the time, the environment is really dark. This results into a very few periods of usability of the color video.

Secondly, babies experiences a lot of manipulations either from medical staff or from parents. They are notably taken out bed over quite long periods for feeding or "skin-to-skin" with parents as well as for bed cleaning. As provided for in the protocol, some recordings were stopped by the medical staff while beds were cleaned or when the baby was moved to another bed (e.g., in case of recordings over days for the infection target).

Two types of flashing were also observed on video recordings. The first one is due to photo-therapy that is used to treat babies which contracted jaundice and impacts the entire image making unusable these periods of recordings. Photo-therapy sessions last a few hours and recordings were usually performed outside of those periods except in case of recording of very long duration such as for the sepsis

population. The second one is due to pulse-oximetry, used to continuously measure blood oxygenation, but this time, it only impacts a part of the image.

To finish, in some of the recordings, the baby was never alone. He/she could be either with his/her mother sleeping in a bed next to his/her own (especially full-term newborns) or with other babies placed in the same room (depending on the hospital). At the beginning of the project, we also observed that twin babies can be placed in the same bed, for a few hours, but this situation had been rapidly excluded from the protocol.

4 Discussion and conclusion

In this chapter, we showed the ability of our first version of Voxyvi, a new audio-video acquisition system, to fit a wide variety of NICU environments. Indeed, the audio-video device had been successfully integrated in six different hospitals, for several room configurations and more particularly on different kind of beds from incubators to cradles.

Additionally, we provided several supports (e.g., user manual, signage, training) to ensure the acceptance of the system by clinicians. As a consequence, a high number of newborns has been recorded during the past three years. However, our inclusion targets are unequally advanced, if some of the objectives are reached or being finalized, there is still some categories under represented (e.g., early term preterm newborns, PRM births, and full-term newborns with high risk to develop sepsis). In fact, the less common populations (extremely and very preterm newborns) have been firstly targeted in order to enhance the chance to reach the expected number of inclusions along the project period. Acquisition systems were mostly requisitioned to record these populations which required, in one hand, longer recordings duration (sepsis) and, on the other hand, more periodic recordings until the discharge (maturation). Though, the inclusions regarding the infection target turned out to be more difficult than for maturation. In fact, these populations of newborns are usually not stable, making more delicate to quickly convince the parents to integrate the protocol. Another difficulty is that the period to include high risk full-term newborns is shorter, as well as their stay in hospital, and has to coincide with an available system.

Moreover, the evaluation of the system regarding installation, user interface, training and care integration was globally very positive. Clinicians confirmed that the newborn safety is ensured all along the process. In fact, the main challenge, which was to propose a system that fits most of the NICUs configurations, is quite successful although solutions to reduce the device size and the number of wires are needed. For that purpose, once relevant descriptors will be retained, the number of sensors may be significantly lower.

However, this positive picture is marred by the fact that, so far, clinicians are not totally convinced by the idea to integrate such a system in their care routine. As a matter of fact, in this phase of the project, the system was only used to collect data, making difficult to envisage the whys and wherefores of its clinical purpose. Besides, in this chapter, we present the possibility to integrate raw data processing in the second phase of the project without getting into the specifics. We also mentioned several difficulties induced by real conditions that will have to be tackled to provide relevant analyses. This will be the object of the next two chapters where motion and cry analyses within the continuous monitoring objective are

addressed.

From the beginning of the project, a few commercial products have been proposed for video streaming of newborns either in NICU [7] or at home [6]. Although, in NICU, the system is only designed to allow parents to visualize their baby outside the hospital, at home, the system includes by wake/sleep analyses. Additionally, a new multimodal system, including pressure mattress, two color cameras, two kinects and two microphones, has been recently published [5]. In this paper, authors suggest to use motion and cry analyses to assess the neuro-behavioral development. However, to date, it remains in the state of a proof of concept and no results were presented in a real environment.

Bibliography

- [1] ABBAS, A. K., HEIMANN, K., BLAZEK, V., ORLIKOWSKY, T., AND LEONHARDT, S. Neonatal infrared thermography imaging: analysis of heat flux during different clinical scenarios. *Infrared Physics & Technology* 55, 6 (2012), 538–548.
- [2] ANDERSON, E., WAILOO, M., AND PETERSEN, S. Use of thermographic imaging to study babies sleeping at home. *Archives of Disease in Childhood* 65, 11 (1990), 1266–1267.
- [3] BRIEVA, J., AND MOYA-ALBOR, E. Phase-based motion magnification video for monitoring of vital signals using the hermite transform. In *13th International Conference on Medical Information Processing and Analysis* (2017), vol. 10572, International Society for Optics and Photonics, p. 105720M.
- [4] FANG, C.-Y., HSIEH, H.-H., AND CHEN, S.-W. A vision-based infant respiratory frequency detection system. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on* (2015), IEEE, pp. 1–8.
- [5] MARSCHIK, P. B., POKORNY, F. B., PEHARZ, R., ZHANG, D., O’MUIRCHEARTAIGH, J., ROEYERS, H., BÖLTE, S., SPITTLE, A. J., URLESBERGER, B., SCHULLER, B., ET AL. A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders. *Current Neurology and Neuroscience Reports* 17, 43 (2017), 1–15.
- [6] NANIT. Nanit plus camera. <https://store.nanit.com/products/nanit-plus-camera-wall-mount>, 2018.
- [7] NATUS. Nicview web camera system. <https://newborncare.natus.com/products-services/newborn-care-products/live-video-streaming/nicview-web-camera-system>, 2016.
- [8] NTONFO, G. M. K., FERRARI, G., RAHELI, R., AND PISANI, F. Low-complexity image processing for real-time detection of neonatal clonic seizures. *IEEE Transactions on Information Technology in Biomedicine* 16, 3 (2012), 375–382.
- [9] ORLANDI, S., GARCIA, C. A. R., BANDINI, A., DONZELLI, G., AND MANFREDI, C. Application of pattern recognition techniques to the classification of full-term and preterm infant cry. *Journal of Voice* 30, 6 (2016), 656–663.

- [10] PUOPOLO, K. M., DRAPER, D., WI, S., NEWMAN, T. B., ZUPANCIC, J., LIEBERMAN, E., SMITH, M., AND ESCOBAR, G. J. Estimating the probability of neonatal early-onset infection on the basis of maternal risk factors. *Pediatrics* (2011), peds–2010.
- [11] RABOSHCHUK, G., JANČOVIČ, P., NADEU, C., LILJA, A. P., KÖKÜER, M., MAHAMUD, B. M., AND DE VECIANA, A. R. Automatic detection of equipment alarms in a neonatal intensive care unit environment: A knowledge-based approach. In *INTERSPEECH 2015: 16th Annual Conference of the International Speech Communication Association* (2015), pp. 2902–2906.
- [12] STAHL, A., SCHELLEWALD, C., STAVDAHL, O., AAMO, O. M., ADDE, L., AND KIRKEROD, H. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20, 4 (2012), 605–14.
- [13] VAN GASTEL, M., BALMAEKERS, B., OETOMO, S. B., AND VERKRUYSSSE, W. Near-continuous non-contact cardiac pulse monitoring in a neonatal intensive care unit in near darkness. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics* (2018), vol. 1050114, International Society for Optics and Photonics, pp. 1–9.
- [14] VILLARROEL, M., GUAZZI, A., JORGE, J., DAVIS, S., WATKINSON, P., GREEN, G., SHENVI, A., MCCORMICK, K., AND TARASSENKO, L. Continuous non-contact vital sign monitoring in neonatal intensive care unit. *Healthcare Technology Letters* 1, 3 (2014), 87–91.
- [15] WERTH, J., ATALLAH, L., ANDRIESSEN, P., LONG, X., ZWARTKRUIS-PELGRIM, E., AND AARTS, R. M. Unobtrusive sleep state measurements in preterm infants—a review. *Sleep Medicine Reviews* 32 (2017), 109–122.

VIDEO-BASED CHARACTERIZATION OF NEWBORN MOTION ORGANIZATION FOR NEURO-DEVELOPMENTAL MONITORING

In previous chapters, we saw that video processing to characterize motion activity is an important step to assess newborn condition for a wide variety of clinical aspects. In this chapter, our approach to integrate motion analysis within the objective of neuro-behavioral development monitoring is presented. First, the process designed to automatically evaluate the newborn motion organization is described. Then, the different steps of this process are evaluated. In the last section, the whole method is applied on Digi-NewB data in order to present its ability to assess newborn behavioral development.

1 Methods to characterize motion organization in newborns

The objective of the process presented in this section is to estimate and characterize the motion of newborns from video recordings of long duration. First, to ensure that the motion series only contains motion of the newborn, intervals including adults' presence have to be automatically detected and removed from the analysis. From there, a segmentation of motion and non-motion intervals can be performed in order to compute a set of features that characterize the motion organization of the newborn.

Figure 6.1 illustrates the four steps of the process that were developed: motion estimation, adults detection, motion segmentation and feature extraction. These steps are detailed in the following sections.

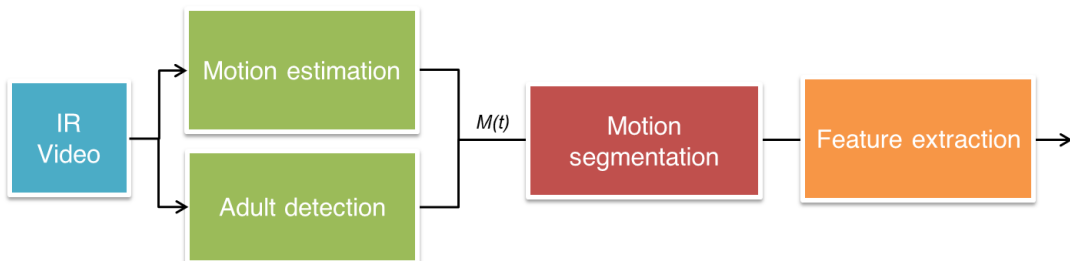


Figure 6.1 – Global framework of motion characterization.

1.1 Estimation of the amount of motion

There are several approaches for motion estimation in a continuous video stream. Three popular motion extraction techniques come out: frame differencing, mainly used to characterize motion globally, optical flow and block matching, used to estimate local displacements [17]. In this work, newborns being well covered by a blanket along recordings, we considered that estimating the global motion information was the most effective manner to characterize the activity of the newborn. Thus, we proposed a simple method based on frame differencing that includes four steps: cropping, frame differencing, morphological opening and quantification of the motion.

Cropping A mask is manually defined and applied on all frames in order to compute motion only in a region of interest. This operation is optional and is mostly used in case of open beds when the camera field is too wide. In fact, parents or nurses may stay in this area without interacting with the newborn. For example in the case of a parent bed. An example of a video frame with the resulting frame after cropping is presented by Figure 6.2.



Figure 6.2 – Example of frame cropping: original frame on the left and cropped frame on the right.

Frame differencing The absolute difference of pixels' intensities is computed between two successive frames $I(t)$ and $I(t - 1)$ (Figure 6.3).



Figure 6.3 – Example of frame differencing (without cropping).

Morphological opening A morphological opening with a 3x3 square structuring element is applied to the resulting difference image in order to remove the impulsive noise.

Quantification of the motion The amount of motion $A(t)$ is obtained by counting the number of pixels with an intensity superior to a threshold T in the resulting image in order to reduce the impact of small intensity variations linked to the camera sensitivity.

The threshold T has been defined by studying the cumulative histogram of video sequences acquired in empty rooms (without baby or adults), resulting to a value of 10 (initial intensity range: [0, 255]).

1.2 Adult detection

Having for objective to monitor the newborns during long periods implies that some intervals will include presence of the parents, rather during the day, but also medical staff, at any time (Figure 6.4).



Figure 6.4 – Examples of adults' presence, parents (left) or nurse (right).

These intervals can have different durations, from few seconds to tens of minutes. Since the goal is to develop a long term monitoring system, the impact of this removal is negligible.

The main difference between adult's motion and baby's motion lies in the fact that, in the first case, it comes from the outside of the scene. This statement led us to develop an algorithm based on the analysis of the change in the image border. Its goal is firstly to detect the adult (medical staff or parent) arrival, and then his/her departure. Successive steps of the process are detailed hereafter.

Initialization The video analysis is started at the first clean frame (i.e., the first frame without medical staff or parents), selected manually. The initial reference border C_{ref} is then defined as the border of 20 pixels thickness from each side of the frame (Figure 6.5(a)). This thickness has been empirically defined according to the area that is supposed to not be crossed by the baby. In case of cropping, the border is automatically adapted to the mask, as depicted by Figure 6.5(b).

In some cases, especially for recordings performed on closed beds, the video device is more difficult to adjust due to a lack of space. Thus, some body parts or blanket can cross the border. In that case, impacted parts of the border are discarded from the analysis. Figure 6.5(c) illustrates an example where newborn's feet are in the border of the image and this area is discarded from the border reference.

Processing Every second, a metric $S(t)$ is computed between the border of the current image $I(t)$ and the reference border C_{ref} :

$$S(t) = \sum_{p \in C} |I(t, p) - C_{ref}(p)| \quad (6.1)$$

where C is the border area and $I(t, p)$ is the intensity of the pixel p in $I(t)$.

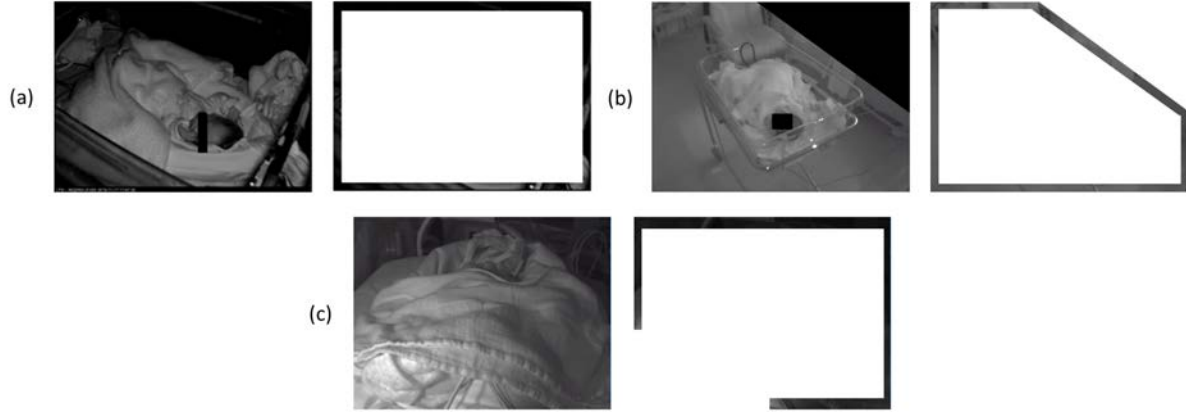


Figure 6.5 – Example of reference border selection: (a) basic case (b) with cropping and (c) with a discarded part on the bottom left.

Based on this metric, the step by step detection is described below and illustrated with an example reported in Figure 6.6.

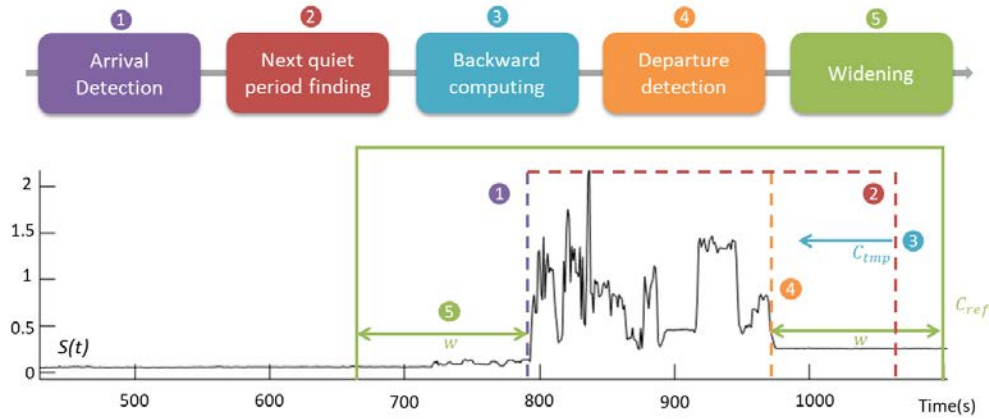


Figure 6.6 – Description of the adults' detection processing.

An arrival is detected if $S(t)$ is superior to a threshold T (Figure 6.6, step 1), that has been empirically set at 6% of the maximal possible value of the metric, i.e. $[\text{number of pixels in } C] \times 255$. However, as the adult can have modified the background of the scene (e.g. cable displacement), the amplitude of $S(t)$ can no more be trusted to detect the departure since it is computed related to C_{ref} . So, the algorithm waits for the next quiet period, in the border, corresponding to a certainty of departure (Figure 6.6, step 2). It is defined as a standard deviation of $S(t)$ (over a period of 80 seconds), lower than 0.02% of the previously defined maximal possible value.

When the quiet period has been detected, the departure time is refined. For this purpose, a temporary border C_{tmp} is firstly defined on the current image. The metric is then computed backward (Figure 6.6, step 3), using C_{tmp} , until the threshold T is reached (Figure 6.6, step 4).

Until there, the algorithm is designed to detect adult presences while avoiding false positive detection coming from movements of the baby in the border. However, in some cases, such as depicted in the example, adults appeared before only in the back of the frame, corresponding to a very small change in the border or reversely, goes slowly out of the frame. In order to counteract this kind of situations, we chose to widen the detected passages on both sides by a value w (Figure 6.6, step 5), which will be set in Section 2.1.3. The reference border C_{ref} is then updated and the analysis continues in the forward direction.

Finally, periods that are artifacted by adult presences are discarded from $M(t)$.

1.3 Motion segmentation

The aim of this process, as depicted by Figure 6.7, is to segment $M(t)$ in order to retrieve motion and non-motion intervals. For that purpose, we designed a three-step strategy. First, a pre-processing step is applied in order to clean the motion series from noisy components. Then, a motion/non-motion classification is performed. To finish, it is supplemented by a cleaning step used to merge and discard short motion and non-motion epochs.

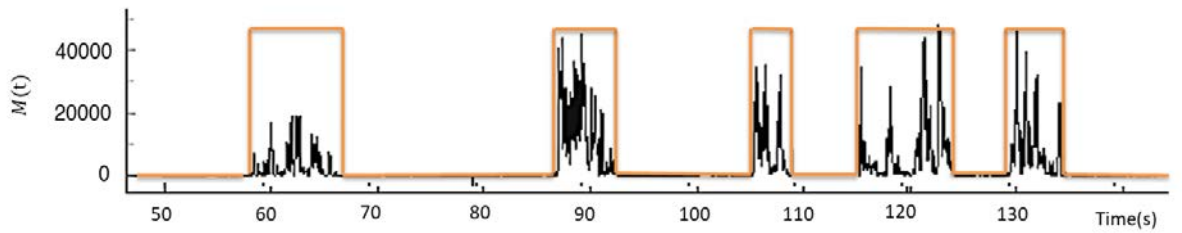


Figure 6.7 – Example of a motion $M(t)$ and the associated automatic motion and non-motion segmentation in orange.

1.3.1 Pre-processing

As a reminder, $M(t)$ is the motion series where adults presence have been discarded. Two pre-processing steps are applied before the segmentation.

The first one was set up in order to correct artifacts resulting from the video encoding that induce localized peaks of high amplitude on $M(t)$. It was corrected by applying a median filter. Since these events are short and only impact a few frames, the size of the windows had been empirically chosen at 11 samples that is a good compromise between removing these peaks without degrading the amplitude of small motion of the baby.

Another difficulty is related to the photo-detector used to measure blood oxygen saturation (also called pulse oximetry) that may intermittently induce noisy components on $M(t)$. Although the flashing is invisible to the naked eye, it is captured by the infrared camera. Its impact is fluctuating between recordings but also along the same video recording. In fact, when the detector is placed on one of the newborns' foot or finger, it may be hidden by the blanket and has no impact on motion series. Reversely,

when the baby moves and that the photo-detector becomes discovered, a noise impacts $M(t)$ more or less significantly depending on the reflection of the flashing on the blanket. An example of artifacted motion is presented in Figure 6.8. It occurs when the detector is visible. In this example, the photo-detector is placed on one of the newborn's foot.

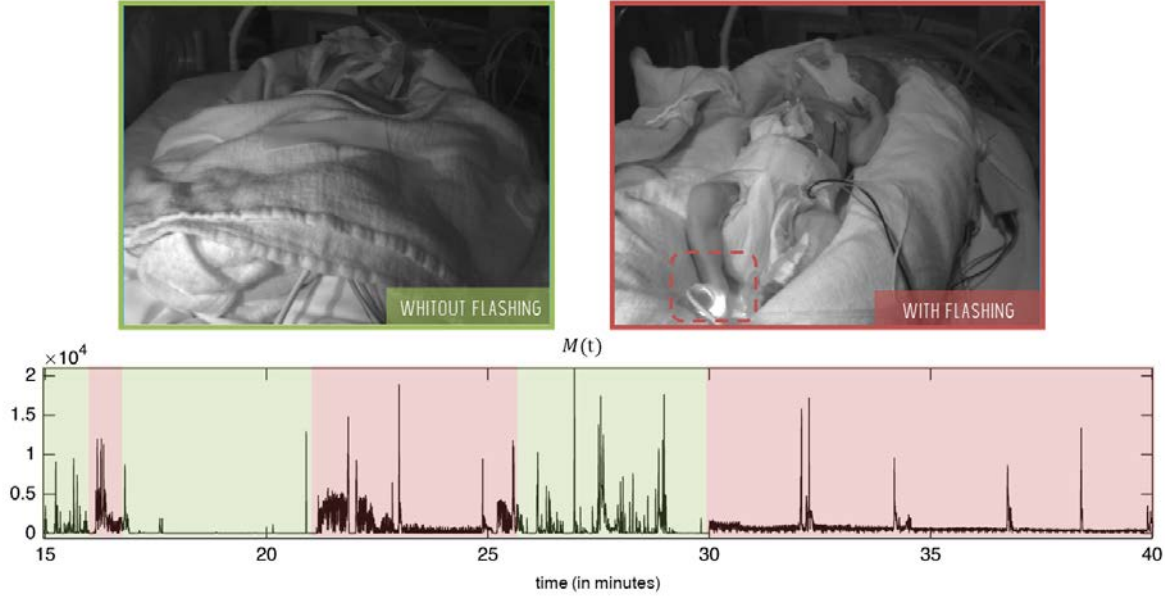


Figure 6.8 – Example of motion $M(t)$ with periods artifacted by pulse-oximetry flashing (in red). In comparison, periods without flashing are reported in green. An illustrative image for each situation is also given.

This noise may interfere with the segmentation and has to be corrected. For this purpose, we applied a Butterworth bandstop filter on $M(t)$ followed by a baseline subtraction. These operations results on a new corrected motion series $M_c(t)$. Both steps rely on parameters, introduced hereafter, that have to be tuned with regard to the segmentation objective (see Section 2.2.3).

Butterworth bandstop filtering Butterworth bandstop filtering is the serial combination of a low-pass and a high-pass filter with f_1 and f_2 being their respective cut-off frequencies (in Hz) and n is the filter order.

Baseline subtraction After filtering, an additional step is needed to remove irrelevant offsets in the baseline during flashing intervals. In regard to the shape of the motion series, we chose to retrieve a flat baseline using the Baseline Estimation and Denoising with Sparsity (BEADS) algorithm [13]. This algorithm is based on the hypothesis that the vector of observations $y(t)$ can be modeled as the combination of a sparse signal $x(t)$, a low-pass baseline f and a stationary white Gaussian process $w(t)$, such as:

$$y(t) = x(t) + f(t) + w(t) \quad (6.2)$$

The model is adjusted by four parameters that are necessary to perform the estimation of each piece of the equation:

- d , the filter order of the baseline $f(t)$;
- f_c , the filter cutoff frequency of the baseline $f(t)$;
- r , the asymmetric ratio of peaks in $x(t)$;
- amp , that regulates the proportionality between the regularization parameters of the optimization problem used to estimate $x(t)$ in the algorithm.

1.3.2 Motion/non-motion classification

In this section, our approach for motion and non-motion classification is presented. First, we described the set of features that was defined. Then, classification algorithms that were investigated are presented.

Feature engineering The aim of this part is to define a valuable set of features for motion and non-motion interval classification. In Chapter 2, Section 2.2.3, we mentioned several groups which tackled classification [9, 14, 15]. However, they focused on seizures or cerebral palsy detection and features that were computed were principally used to describe trajectories, speed or acceleration. Although these features are interesting to describe complex movements of the newborn, they are not appropriate for a motion/non-motion classification. Thus, we decided to look at set of features used for a similar purpose: burst detection in EEG.

In [12], a review of features extracted for burst detection was presented. From there, authors proposed a new set of features of different types (e.g, amplitude, statistical, energy) that we chose to apply to our objective. Nine features compose the set of features X_i for each sample i . They are computed by taking into account adjacent values of $M_c(i)$, in a window W of 2 seconds, as follows:

- M_m, M_{m-1}, M_{m+1} : the differences between the maximal and the minimal values on, respectively, the current, previous and following window;
- DM , the maximum of absolute values of the first order derivative on W , computed as:

$$DM = \max_{k=1, \dots, n_s} \{ |M_c(k) - M_c(k-1)| \} \quad (6.3)$$

where $n_s = 2 * F_s$ samples and $F_s = 25$;

- Sd : Standard deviation on W ;
- Kt : Kurtosis on W ;
- NL : Computation of the Non Linear Energy Operator (NLEO) [16] on W , expressed as:

$$NL(k) = M_c(k)M_c(k-3) - M_c(k-1)M_c(k-2) \quad (6.4)$$

- *RMS*: Root mean square on W , defined as:

$$RMS = \sqrt{\frac{1}{n_s} \sum_{k=1}^{n_s} M_c(k)^2} \quad (6.5)$$

- *MF*: Averaged differentiation on W , computed as:

$$AD = \frac{1}{n_s} \sum_{k=1}^{n_s} |M_c(k) - M_c(k-1)| \quad (6.6)$$

Classifiers We have seen in Section 2.2.3 of Chapter 2 that only Cuppens *et al.* aimed at discriminating movement and non-movement episodes. In [7], two approaches were presented: a) a fixed threshold determined thanks to a ROC curve and b) a variable threshold adapted to the noise and computed from mean and standard deviation of a selected non-movement period. Since the variable threshold approach requires the identification of a non-movement period on each recording, we chose to only focus on the fixed threshold approach, which is more suitable with continuous monitoring purpose. Additionally, two types of classification algorithms were investigated: linear (Logistic Regression) and non-linear (K-Nearest Neighbors). They were retained due to their simplicity (see Chapter 3) and the fact that they gave satisfactory results for burst detection [12].

In our case, two strategies have been defined regarding the dimension of the feature set: $p = 1$ for fixed thresholding whereas $p = 9$ for LR and KNN. In fact, for the thresholding approach, we chose to directly map $M_c(t)$ values according to their labels (motion or non-motion). In case of LR and KNN, we use the feature set described before.

1.3.3 Clustering

In order to work on relevant motion epochs, we added two steps for clustering.

First, a fusion of the detected events is performed. If a motion period appears less than five seconds after another, the two periods are merged. Then, a suppression of short motion events is made. All periods that last less than two seconds are eliminated.

These values were defined based on literature values [7], as well as, from the experience acquired during scoring. This clustering step is illustrated in Figure 6.9.



Figure 6.9 – Example of an original detection and the result after clustering.

1.4 Feature extraction

Once the segmentation is performed, it is possible to characterize the intervals by different features.

Amplitude features, such as the mean, the median or the maximum of motion series, as proposed in [2, 3], can be influenced by external factors like the baby size or the recording conditions (e.g., shadows, zoom). Thus, to characterize the newborn motion activity, a set of 14 features independent of the amplitude of the motion series has been defined. These features were proposed in order to describe motion and non-motion intervals in terms of duration and number. The whole feature set is reported in Table 6.1.

Table 6.1 – Definition of the motion set of features.

Feature description	from motion intervals	from non-motion intervals
Total duration (seconds)	T_m	T_{nm}
Mean duration (seconds)	m_m	m_{nm}
Median duration (seconds)	md_m	md_{nm}
Maximum duration (seconds)	mx_m	mx_{nm}
Standard deviation of duration (seconds)	sd_m	sd_{nm}
Relative standard deviation of duration	rsd_m	rsd_{nm}
Number of intervals (scaled to an hour)	n_m	n_{nm}

These features are computed on sliding windows of 5 minutes with 50% of overlap. Additionally, in order to work with relevant features, only windows with more than 50% of cleaned data was kept. Hence, if an adult presence is detected on the window for a duration of more than 2 minutes and 30 seconds, features are not computed.

2 Evaluation of the process

2.1 Evaluation of the adult detection method

2.1.1 Dataset presentation

In order to evaluate the adult detection method, periods of adults' presence have been manually annotated on ten different videos of nine premature infants, representing 149 hours of recordings. Figure 6.10 illustrates this dataset.



Figure 6.10 – Overview of the annotated dataset for the evaluation of the adult's presence algorithm.

Infants were born between 26 and 36 weeks GA and recorded between 33 and 39 weeks PMA. This database was extracted from a previous protocol conducted at the University Hospital Rennes Sud

(VARINEONAT). In this study, an infra-red black and white camera (600x400 pixels, 25 fps) was installed near the bed of the baby in order to observe most of his body.

2.1.2 Evaluation strategy

We compared the manual and the automatic detections at each second. For both states "Presence" and "Absence", and considering the first one as the positive case, values of sensitivity (Se), specificity (Sp) and accuracy (Acc) have been computed.

2.1.3 Tuning of the widening

As mentioned previously, two parameters of the adult's detection algorithm have been empirically fixed (T and the size of the border). However, the algorithm also includes a widening parameter w to counteract crossing in the back of the frame before and after an intervention on the baby. According to our objective, we particularly need to detect all the parents and medical staff presence, i.e. reach sensitivity as high as possible, without cutting too many analyzable periods. For this purpose, we studied, by the mean of a receiver operating characteristic curve (ROC), the evolution of the sensitivity and specificity after applying different widening durations w (see Figure 6.11).

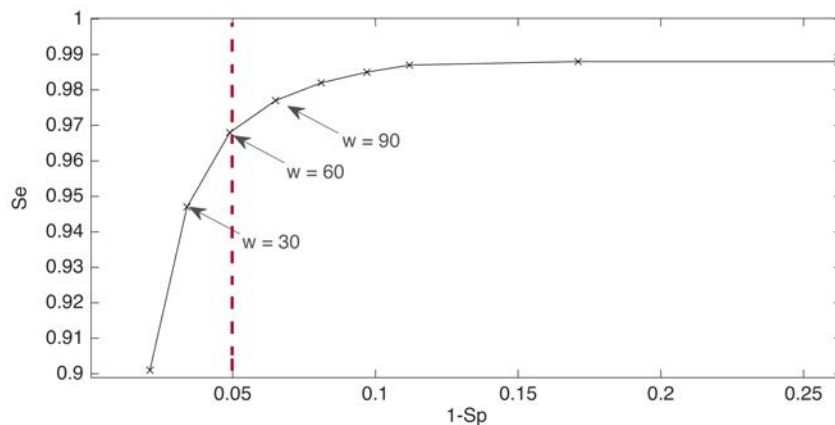


Figure 6.11 – Influence of the widening w on the values of Sp and Se.

In order to increase the algorithm sensitivity without decreasing specificity too much, we chose to apply a widening w of 60 seconds which allows to keep a false positive rate under 5%. An example of processing is presented by Figure 6.12.

In this example, all of the manually annotated passages of adults had been detected and the duration of the detected periods are globally accurate. It is mainly thanks to the widening step. Indeed, if we take a closer look on the third detected epoch, the widening prevented all the false negative detections at the end of the event. Reversely, in the second epoch, a larger widening would have been better to detect a more accurate end of the passage but, in that case, the arrival would have been detected too early. The global impact of this type of errors on the performances had to be evaluated recording per recording and this is the object of the next section.

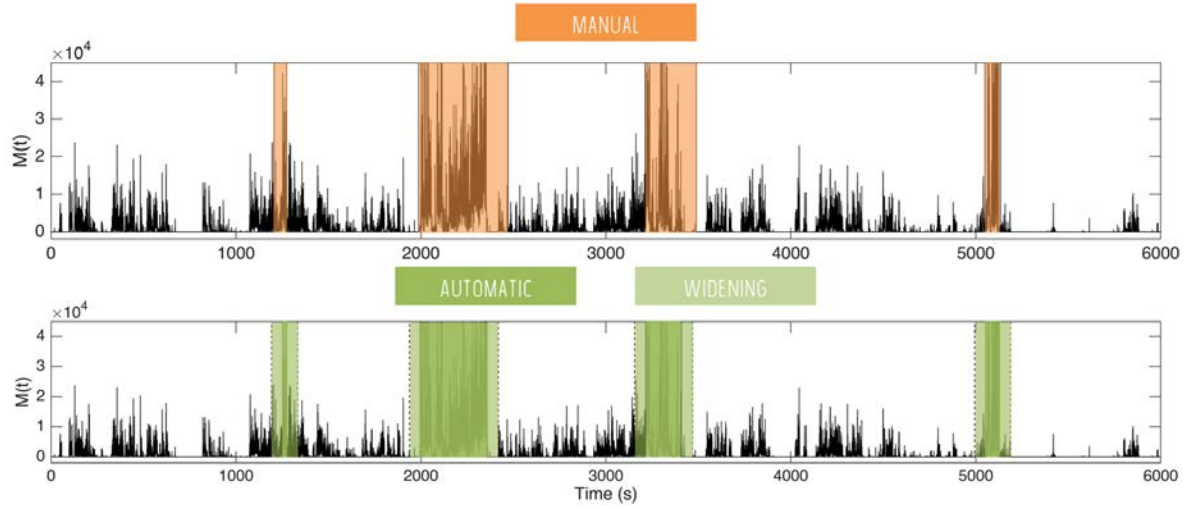


Figure 6.12 – Example of an estimated motion series artfacted by passages of adults. Manual adult's detection (orange) and automatic adult's detection (green) with the widening (light green).

2.1.4 Performances

In this section, performances of the adult detection algorithm are computed for each recording of the database presented in Section 2.1.1. The resulting performances are presented in Table 6.2.

Table 6.2 – Performances of the medical staff and parents detection method. Duration in hours. FN: False Negative, TP: True Positive, FP: False Positive and TN: True Negative. Se : Sensitivity, Sp : Specificity and Acc : Accuracy, in percent.

Video	Duration	FN	TP	FP	TN	Se	Sp	Acc
1	17	30	8463	3929	47128	99.6	92.3	93.4
2	17	555	10135	1068	48138	94.8	97.8	97.3
3	9	3	478	1354	31173	99.4	95.8	95.9
4	15	265	3273	2203	46674	92.5	95.5	95.3
5	14	14	2686	783	46717	99.5	98.4	98.4
6	16	276	5316	2866	49097	95.1	94.5	94.5
7	15	70	3392	2648	49378	98.0	94.9	95.1
8	17	97	7990	2685	51976	98.8	95.1	95.6
9	13	113	806	4702	40949	87.7	89.7	89.7
10	16	74	3338	1892	52123	97.8	96.5	96.6
Total	149	1497	45877	24130	463353	96.8	95.1	95.2

Averaged performances are high with 96.8% of sensitivity, 95.1% of specificity and 95.2% of accuracy. Results per video are also high, ranging from 87.7% to 99.6% for the specificity and from 89.7% to 98.4% for the specificity. We can observe that the performances on Video 9 are below the others. This is explained by a long presence of adults in the back of the frame at the end of the video. It lasted longer than what is handled by the widening of our method. Thus, the main limitation of the detection algorithm is when adults stay in the border without moving too much. However, in that case, induced values on motion series may be negligible in comparison with the motion of the baby.

2.2 Evaluation of the motion segmentation method

2.2.1 Dataset presentation

A sub-dataset of the Digi-NewB maturation protocol, composed by ten recordings, was selected in order to evaluate our segmentation approach (see Figure 6.13).

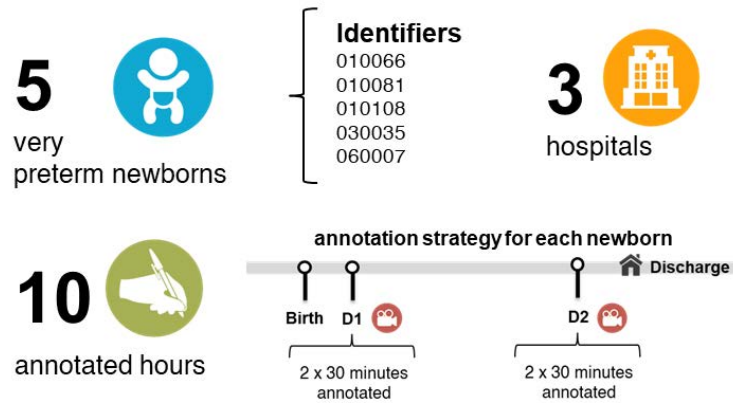


Figure 6.13 – Summary of the annotated dataset for motion segmentation evaluation.

We chose to focus on five very preterm infants recorded in three different hospitals: Rennes, Brest and Tours (baby inclusion numbers respectively begin by 01, 03 and 06). For this purpose, a total of ten hours of video recording were manually scored, using VisiAnnot¹. For each newborn, four periods were annotated in terms of motion ('1') and non-motion ('0') intervals:

- two periods of 30 minutes of the recording nearest the birth ($D1$);
- two periods of 30 minutes of the recording nearest the discharge ($D2$).

To illustrate video data, for two babies (010066 and 030035), an image was extracted from each annotated period and is reported in Figure 6.14. For $D1$, the camera is placed at 30 cm from the newborn using the support whereas for $D2$, the use of clamps allows a distance from the baby up to 80 cm (e.g., see Figure 6.14(a) and (b) versus Figure 6.14(c) and (d)).

In addition, along the same recording, conditions such as illumination (e.g., Figure 6.14(c) versus Figure 6.14(d)) or covering of the baby (e.g., Figure 6.14(e) versus Figure 6.14(f)) may change. Thus, to integrate this heterogeneity, two different periods for each video recording were annotated.

Moreover, a wide variety of conditions regarding the newborn equipment is integrated in the database. For examples, a pulse-oximetry sensor is placed on the newborn's hand in Figure 6.14(c) or a respiratory assistance is provided to the baby in Figure 6.14(b).

¹. An annotation tool that proposes a synchronized visualization of signals of different sources, including video. It was developed by Raphaël Weber, post-doctorant at LTSI.

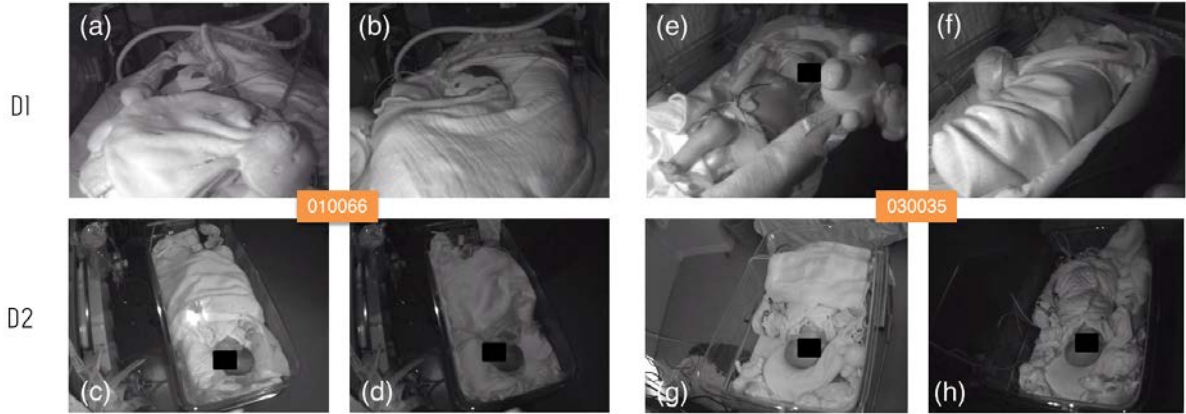


Figure 6.14 – Images illustrating the four annotated periods in both days ($D1$ and $D2$) for 010066 (a)-(d) and for 030035 (e)-(h).

2.2.2 Evaluation strategy

We pursued a leave-one-patient-out cross validation strategy to evaluate the generalization of the method for a new patient.

To compare classifiers, metrics are calculated for each cross-validation. We computed sample to sample performances metrics (Se , Sp and Acc), considering "motion" as the positive case. As in [7], we decided to minimize the impact of errors due to delay in the manual annotation by discarding one second before and after each transition in the manual scoring of the evaluation.

2.2.3 Tuning of the parameters

Parameters for flashing correction For the flashing correction method, seven parameters have been tuned in order to get the highest accuracy regarding our objective to segment motion and non-motion intervals. Table 6.3 summarizes the tests conducted for each parameter. All the combinations were tested and the final value of each parameter is marked in bold.

Table 6.3 – Parameters testing summary. Final selecting sets of parameters are marked in bold.

Method	Parameters
Butterworth bandstop filtering	$f_1 \in [0.1, 0.15, 0.2, \mathbf{0.25}, 0.3, 0.35]$ $f_2 \in [\mathbf{12.49}, 10.49, 8.49, 6.49]$ $n \in [1, 2, \mathbf{3}, 4]$
Baseline subtraction	$d \in [\mathbf{1}, 2]$ $f_c \in [\mathbf{0.006}, 0.1, 0.2, 0.3]$ $r \in [\mathbf{2}, 4, 6, 8]$ $amp \in [0.4, 0.6, \mathbf{0.8}, 1]$

An example of artifacted motion $M(t)$ and its correction $M_c(t)$ is provided by Figure 6.15. On $M_c(t)$,

periods without motion are now similar with and without flashing. Additionally, the method doesn't degrade motion and non-motion events, especially in terms of durations.

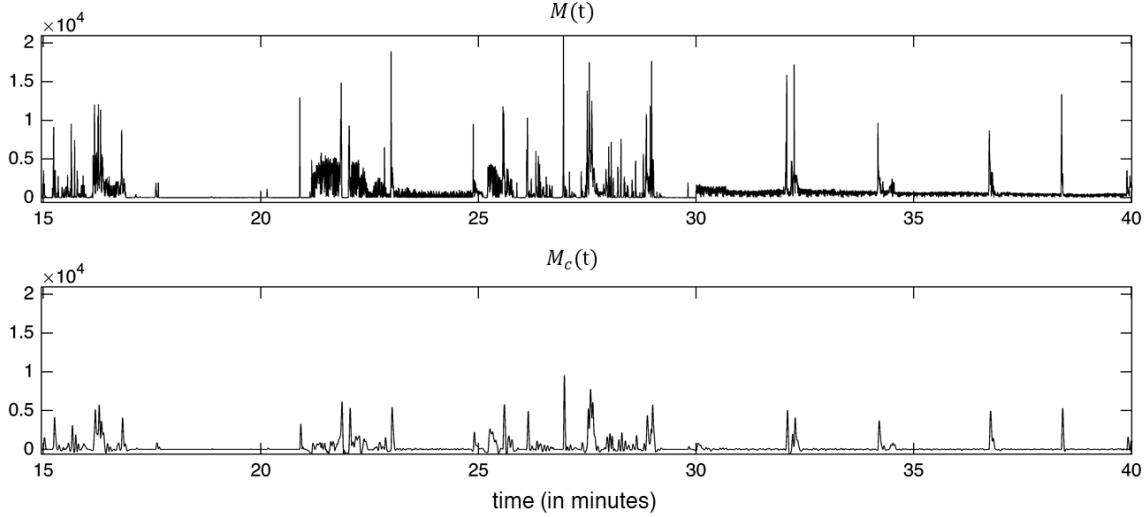


Figure 6.15 – Example of motion $M(t)$ (top) and its corrected version $M_c(t)$ (bottom).

Classifier parameters For Thresholding and LR, a threshold have to be defined. In the first case, it is used to separate motion and non-motion samples and in the second, in order to define the cut-off threshold that decide the output class regarding probability. For both methods, the threshold was defined as the one giving the smaller false positive rate ($1-Sp$) while keeping a high true positive predictive rate (Se). It is also known as the closest point to the upper left corner of the ROC (as presented in Section 2.1.2).

The number of nearest neighbors k used to predict a sample class had also to be tuned. We evaluated the performances for $k \in [1, 3, 5, 11]$ and $k = 3$ showed the best classification results.

2.2.4 Results

To evaluate the segmentation method, two approaches were investigated: with all data and working on data day by day ($D1$ or $D2$).

Global model The first approach is to construct a global model using data from both dates. Five cross-validations were performed with 8 hours for training and 2 hours of testing each time. Averaged performances on the validation sets were computed and are reported in Figure 6.16. KNN shows the best performances regarding Se with a mean value of 80.7%. High Sp and Acc are also observed with mean values of 88.4% and 87.8%, respectively. For their part, thresholding and LR, underestimated motion samples that led to lower Se and higher Sp .

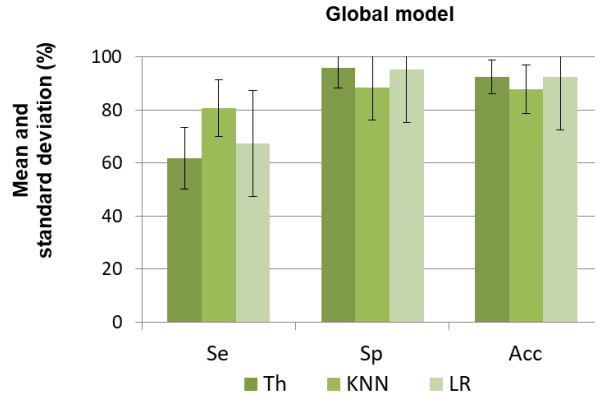


Figure 6.16 – Performances of Thresholding (Th), LR and KNN over the five cross-validations in terms of mean and standard deviation of Se , Sp and C for the global model.

Day by day models The second approach consists in constructing a model for each day. In fact, since bed configurations and newborn development differ between these two dates, we considered the hypothesis that specific models may produce more relevant predictions.

In $D1$, recordings are performed on incubators. We pursued the same leave-one-out cross-validation strategy but, this time, with 4 hours of training data and 1 hour of testing each time. Averaged performances are reported in Figure 6.17(a). For each classifier, performances are highly degraded, especially in term of Se (overfitting). In fact, more non-motion samples than motion samples were scored in $D1$ (imbalanced dataset) causing a bad generalization of the models.

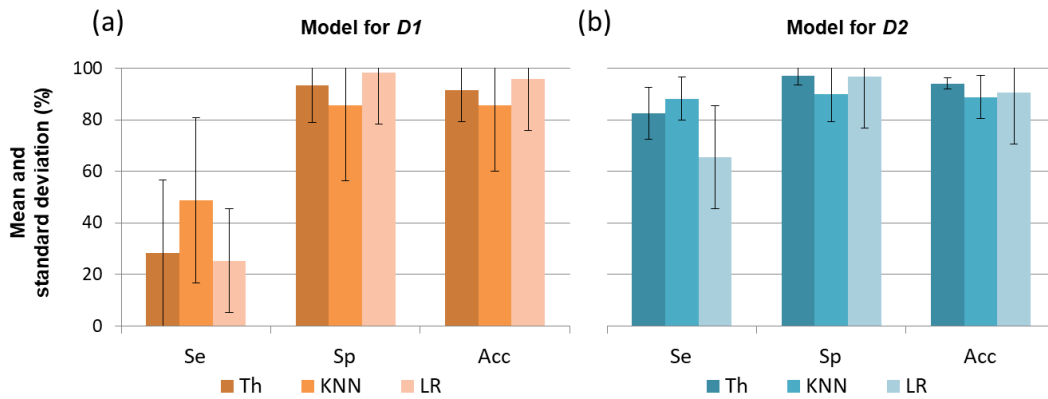


Figure 6.17 – Performances of Thresholding (Th), LR and KNN over the five cross-validations in terms of mean and standard deviation of Se , Sp and C (a) the model for $D1$ and (b) the model for $D2$.

The same strategy is applied to construct a model for $D2$ for which infants are recorded in open beds. Averaged performances are reported in Figure 6.17(b). Reversely to previous attempt ($D1$), model on

$D2$ reveals better classification results in KNN and Thresholding. In fact, a higher mean for Se (88.3%) and smaller variations along the experience in Se are observed with KNN. Mean of Sp is also increased and reached 90.1%. Results are also upgraded with Thresholding but remains slightly below than for KNN. Reversely, worst results are obtained with LR which shows a really high standard deviation over cross-validations for Se .

2.2.5 Final strategy

Finally, in the light of the previous outcomes, we chose to keep the $D2$ model to analyze motion estimated from video where baby is lying in open beds whereas the global model is applied for closed beds. In Figure 6.18, four motion segmentation results are reported.

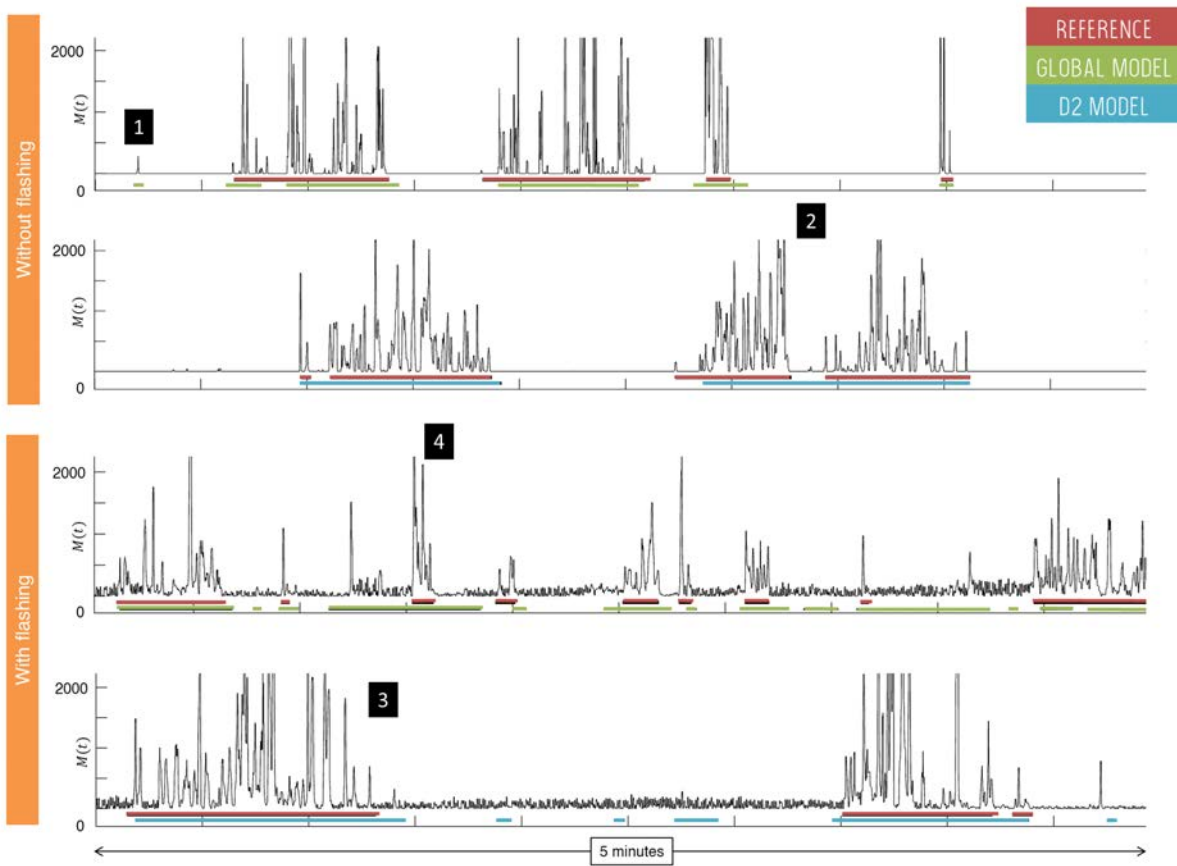


Figure 6.18 – Four segmentation results using the global model (in green) or the $D2$ model (in blue) in comparison with the reference annotation (in red). For each model, results without and with flashing are presented.

Reference epochs of motion are displayed in red while automatic segmentation are reported in green when the global model (for $D1$) was applied or in blue with $D2$ model. For each model, an example without and with flashing with the original $M(t)$ is given. In any case, the flashing correction is performed.

Four instants are numbered, from one to four, in order to discuss the algorithm limitations.

Globally, the segmentation algorithm works pretty well on each recording situations. In fact, the detected motion epochs and durations of events are generally comparable with the manual scoring.

However, false positive and false negative errors can be noticed. Errors of both types are mainly due to the subjectivity in the definition of motion events. In fact, it is sometimes difficult to evaluate, due to the blanket, if a motion event should be considered as such or if it is just a small movement induced, for example, by respiration (see the first detected epoch in Figure 6.18(1)). It is also complex to score start and stop times of motion events since small motion may happen before and after. For example, in Figure 6.18(2), the start of the motion event was scored earlier in the manual scoring than with the automatic detection. Reversely, although in the manual scoring, this period was split in two motion events, the algorithm returned only one event due to a small motion detected in between the two events.

Although segmentation on motion series with flashing works quite well (Figure 6.18(4)), the segmentation algorithm seems to work better on motion series without flashing. In fact, some motion events are overestimated on series with flashing (Figure 6.18(3)). This is also caused by small motions, including respiration movements, that are magnified by the lighting of the photo-detector.

In regard to these results, we see that a point per point evaluation is not sufficient to evaluate the process and thus, in the following section we evaluated the accuracy of the motion features that are extracted to describe the motion organization of the newborn.

2.3 Evaluation of the accuracy of motion features

Additionally, in order to discuss the relevance of the segmentation algorithm, values estimated from the automatic segmentation were compared with the ones computed from the reference segmentation. For all the recordings introduced in Section 2.2.1, the fourteen features were computed as described in Section 1.4. Three examples of feature distributions on the whole population are provided in Figure 6.19.

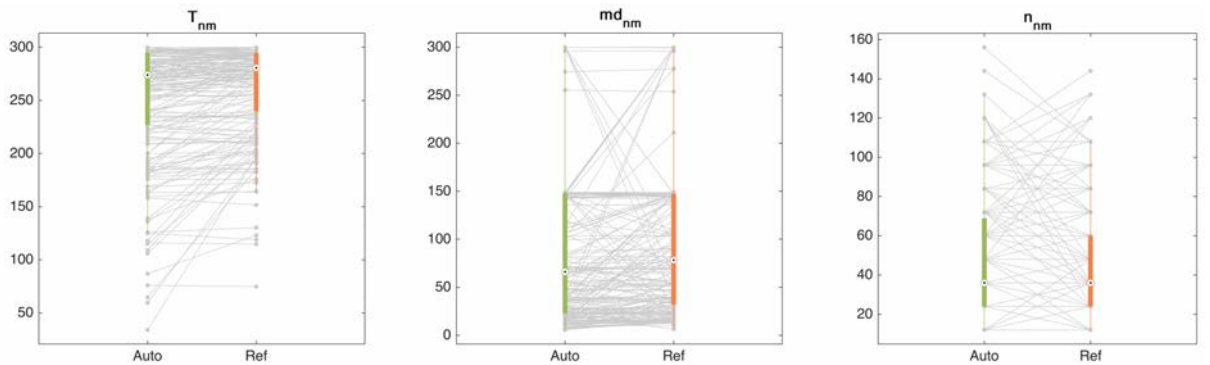


Figure 6.19 – Boxplots of the estimated values of T_{nm} , md_{nm} and n_{nm} from the automatic (green) and from the reference (orange) segmentation. Each point is linked between the two scores with a gray line.

Distributions of the three features seems to be similar between the automatic and the reference. In addition, each point was linked between both scores. For duration features T_{nm} and md_{nm} , most of the points are linked by straight lines. This directly means that most values computed from the automatic

segmentation are equal to the ones obtained from the reference segmentation. For the number of non-motion epochs n_{nm} , the differences between values seem more significant than in durations since it is a discrete feature (scaled to an hour). Nevertheless, small deviations in the values are also observed.

This was confirmed by statistical analyses (Mann Whitney U test) that were performed for each feature and for all recordings. No statistical difference ($p < 0.001$) was found for 139 of the 140 tests (number of features x number of recordings) conducted. For one test ($D1$ of 030035), the automatically computed total duration of non-motion epochs T_{nm} was revealed statistically different from those obtained with manual segmentation.

3 Evolution of the motion organization in preterm newborns

In this section, the evolution of the motion organization in preterm newborn is discussed. First, the dataset of this study, focused on maturation, is presented. Then, several approaches are pursued:

- Comparison between motion organization near birth ($D1$) and near discharge ($D2$) in very preterm infants;
- Comparison between motion organization near birth ($D1$) and near discharge ($D2$) in late preterm infants;
- Comparison between motion organization of very preterm infants, late preterm infants and full-term newborns near discharge ($D2$);
- Design of a control map for assessing motion organization development in preterm infants.

3.1 Maturation dataset

The maturation dataset, presented by Figure 6.20, is composed by 22 recordings of 14 newborns, giving a total of 182.5 hours. Recording durations are ranged from 6.70 to 8.92 hours, with a mean duration of 8.30 hours. The population have been divided in three groups, based on premature newborn categories:

- **Group 1 ($G1$):** six very preterm newborns born between 28 and 29 GA;
- **Group 2 ($G2$):** six late preterm newborns born between 33 and 37 GA;
- **Group 3 ($G3$):** two full-term newborns born between 39 and 42 GA.

Since busyness in hospital is higher by day than by night, it may influence the newborn behavior. In order to maintain consistency in the study, all recordings were studied between 22pm and 6am. For $G1$ and $G2$, two recordings ($D1$ and $D2$) were retrieved for each newborn, except for four of them (060007, 010076, 010086 and 010064) where only a recording for $D1$ or $D2$ was available. Naturally, full-term newborns were only recorded once ($D2$). PMA for $D1$ are varying between 28+4 and 37+1 weeks whereas PMA for $D2$ are ranged between 36+4 to 41+5 weeks.

The global framework of motion characterization was applied on the 22 recordings. For each recording, the fourteen features were computed on a sliding window of 5-minutes length with 50% of overlap. For each recording, the processing duration lasted about 30 minutes.

	BABY	GA (IN WEEK-DAY)	DAY	PMA (IN WEEK-DAY)	DURATION (IN HOUR)
GROUP 1	010108	27+5	D1	28+4	7.92
			D2	36+6	7.92
	010066	28+2	D1	28+5	8.92
			D2	38+1	8.92
	030033	29+0	D1	29+1	8.42
			D2	38+1	8.42
	060007	29+2	D2	38+6	6.70
GROUP 2	010081	29+5	D1	29+5	7.92
			D2	37+2	7.92
	030035	29+6	D1	30+0	8.42
			D2	38+3	8.42
	010075	33+1	D1	34+1	8.42
			D2	37+1	8.42
	010076	33+3	D1	33+4	8.42
GROUP 3	030011	33+5	D1	33+6	8.42
			D2	36+6	8.42
	010086	35+4	D1	35+5	8.42
			D2	37+1	8.42
	010084	36+2	D1	38+1	8.42
			D2	37+3	8.42
	010064	36+5	D2	37+3	8.42
	010098	39+4	D2	39+6	8.42
	010087	41+4	D2	41+5	8.42

Figure 6.20 – Presentation of the maturation dataset.

3.2 Comparison between $D1$ and $D2$ in $G1$

We firstly focus on the very preterm group and aim to compare their motion activities between $D1$ and $D2$. For that purpose, the five very preterm infants which were recorded twice were studied. Features were computed on 821 windows for $D1$ against 636 for $D2$.

Statistical analyses through Mann Whitney U test were performed to find significant differences in motion features between $D1$ and $D2$. The repetitive effect was minimized by the use of Bootstrap method. Twenty draws of 32 random windows, distributed according to a uniform distribution on the recording duration of each group, have been realized. The p-values have been computed as the median of the p-values of the 20 draws. The number of random windows which were picked was defined as 5% of the number of windows for $D2$ (smaller number of windows). Since multiple comparisons are conducted, it is wise to apply a correction regarding p-values, such as Bonferroni correction [1]. This correction reduces the risk to find significant differences by chance².

Only one feature (n_{nm}) was shown statistically significant ($p < 0.05$) to differentiate both dates. Distributions of the values between $D1$ and $D2$ in $G1$ for the number of non-motion epochs n_m are depicted by Figure 6.21. Additionally, we also chose to represent the median duration of non-motion epochs md_{nm} since its p-values remains quite low after the Bonferroni correction ($p = 0.06$).

Hence, two dynamics emerged:

1. the median duration of non-motion epochs tends to increase between $D1$ and $D2$;
2. the number of non-motion epochs appears to decrease between $D1$ and $D2$.

2. In section 2.3, we did not applied any correction although multiple comparisons were also conducted. In fact, since in that case, we wanted to find no significant difference, using a correction would have increased the risk of finding no statistical difference by chance. In fact, there is no consensus on when a correction must be applied or not [4, 11]. Thus, each time, we have chosen to put ourselves in the most restrictive conditions.

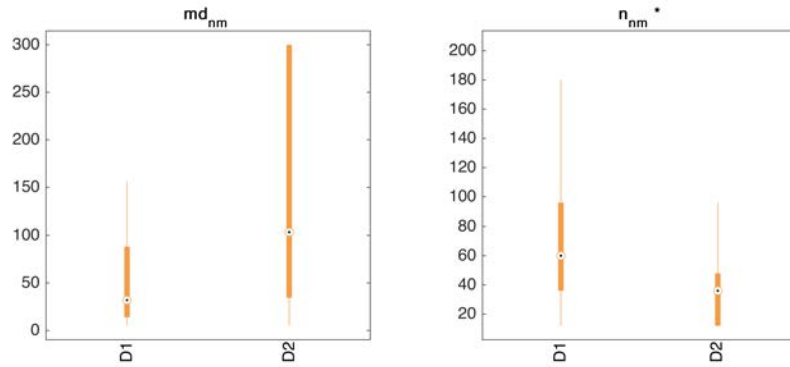


Figure 6.21 – Boxplots of the median md_{nm} durations of non-motion epochs, supplemented by the number of non-motion epochs n_{nm} between $D1$ and $D2$ in $G1$. The * indicates features with p-value < 0.05.

To take the analysis one step deeper, these features were analyzed per baby. Median values for each day are reported in Figure 6.22.

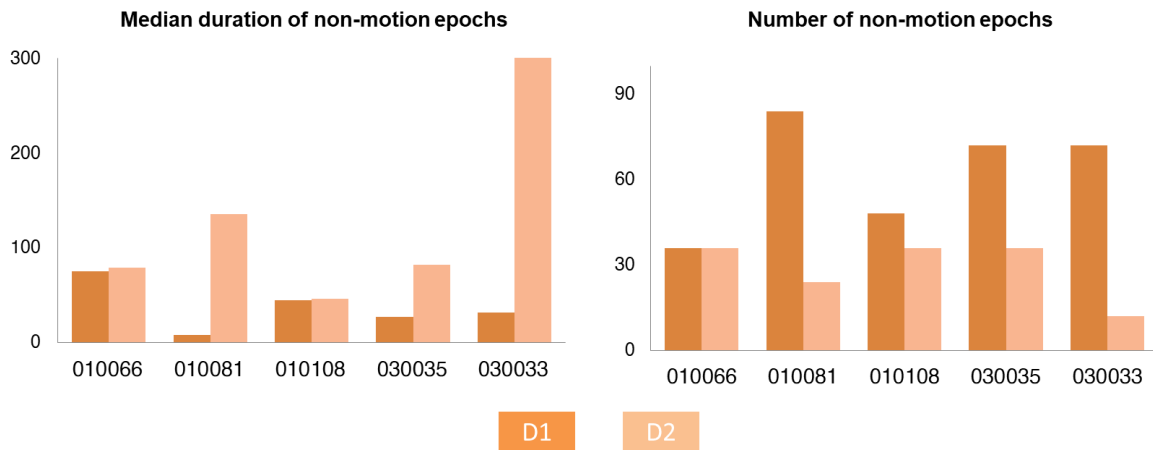


Figure 6.22 – Median values for $D1$ and $D2$ of md_{nm} and n_{nm} in $G1$.

The dynamic regarding duration is verified for three newborns (010081, 030035, 030033). In fact, the median duration of non-motion epochs increased between $D1$ and $D2$. For their part, babies 010066 and 010108 shows quite similar median duration values between $D1$ and $D2$. For all babies, the number of non-motion epochs decreased, except for 010066 which shows similar values between $D1$ and $D2$. However, after examination of his medical record, it appeared that he was treated with antibiotics during the recording in $D1$. This may give an explanation why he presented less and longer non-motion periods than others in $D1$.

The diminution of the number and the augmentation of the median duration of non-motion epochs between $D1$ and $D2$ reveals that quiet periods are less fragmented (i.e., last longer) in $D2$ than in $D1$.

in very preterm infants. These features, especially the number of non-motion periods, appear to be relevant to quantify maturation.

3.3 Comparison between $D1$ and $D2$ in $G2$

In a second time, we focus on the late preterm group and also aim to compare their motion organization between $D1$ and $D2$. For that purpose, only three infants (recorded twice) were studied. Features were computed on 513 windows for $D1$ against 440 for $D2$.

Statistical analyses through Mann Whitney U test were also performed, as in section 3.2, to find significant differences in motion features between $D1$ and $D2$. In that case, no feature was found significantly different on the whole population.

However, if we take a look closer to values studied in Section 3.2, same dynamics than in $G1$, regarding median durations of and the number of non-motion epochs, are observed for 010075 and 030011 (Figure 6.23).

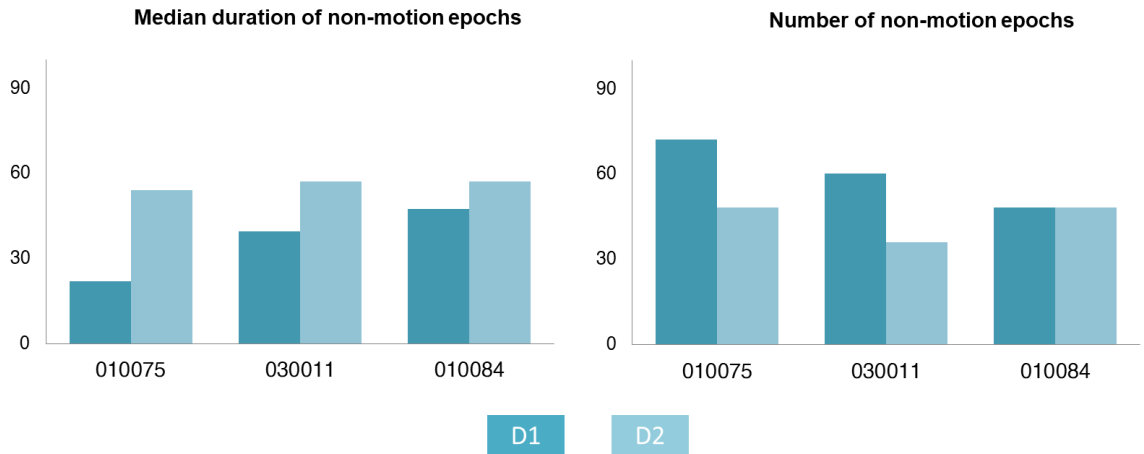


Figure 6.23 – Median values for $D1$ and $D2$ of md_{nm} and n_{nm} in $G2$.

The baby 010084 reaches the same value in $D1$ and in $D2$. It can be explained by the fact that, for this newborn, recordings of $D1$ and $D2$ took place only one week apart.

3.4 Comparison of motion activity near discharge

Another approach is to compare all motion features near discharge ($D2$) between the three groups. Twelve recordings were compared:

- Six very preterm newborns from $G1$;
- Four late preterm newborns from $G2$;
- Two full-term newborns from $G3$.

Features were computed on 1164 windows for $G1$, 996 for $G2$ and 404 for $G3$. Distribution of the values are quite similar in $G1$, $G2$ and $G3$ for all features. Distribution of the values that were revealed meaningful in previous sections are reported in Figure 6.24.

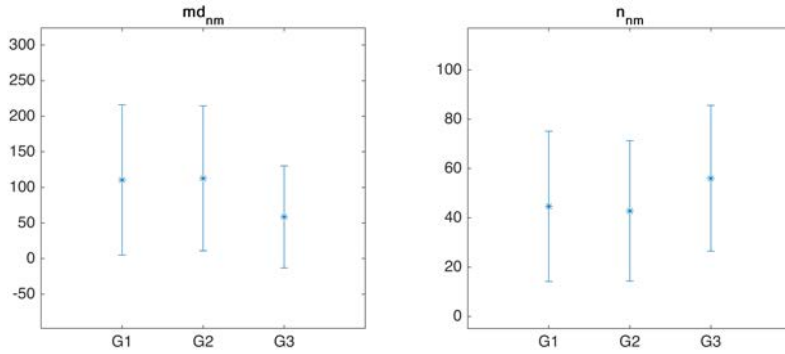


Figure 6.24 – Distribution (mean \pm std) of md_{nm} and n_{nm} between $G1$, $G2$ and $G3$ in $D2$.

This hypothesis was confirmed by Mann Whitney U tests, (see section 3.2), since no significantly difference ($p < 0.05$) has been returned. Hence, the whole group seems to reach the same order of mean values, regarding motion organization, near discharge. These results are particularly reassuring because they suggest that the gestational age may have no impact on the newborn behavior in terms of motion organization near discharge.

However, if we take a closer look, values seem quite different in $G3$ for the two features. Indeed, it appears that for md_{nm} , the range of standard deviation of values in $G3$ is smaller than the ones of $G1$ and $G2$. Increasing the population of $G3$ (only two babies here) may reveal some differences in motion organization. In that case, it would be interesting to study if this difference results from the time spent *ex-utero* or from difference in behavioral development between newborn categories itself.

3.5 Control map for assessing motion organization development in preterm infants.

This final point is both a result and a perspective. In fact, the objective of the Digi-NewB project is to monitor newborn behavioral development by studying several components, including motion. Thus, we want to construct a control group to evaluate newborn developments regarding motion.

For that purpose, we used the 21 recordings in our maturation dataset (all except $D1$ of 010066 because of the antibiotic administration). In the light of previous results, we focused on the number of non-motion epochs. For each recording, the median value on all the analyzed windows was computed. Then, in order to assess the organization of the motion in newborns, we proposed a control map, illustrated by Figure 6.25. It describes the expected development of the newborn activity in function of his/her PMA and GA.

As a confirmation of previous results, one can see a trend for the feature to be darker (= to decrease) from left to right of the graph (= with PMA). In addition, near discharge (about 38 PMA), the graph shows a quite uniformity of colors top-down (= between groups).

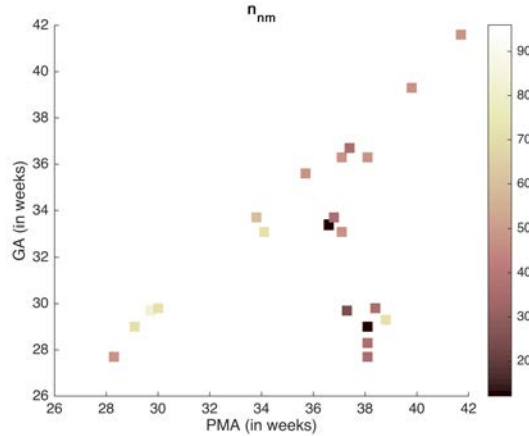


Figure 6.25 – Control map of the number of non-motion epochs for maturation evaluation.

Hence, the behavioral development of a newborn may be compared to others considering his/her GA and PMA. Diverging points in control map can alert clinicians to perform further analysis on motion development. Naturally, this assumption will have to be confirmed for a higher number of babies of different GA, recorded for diverse PMA. In addition, verified cases of abnormal behavioral development will have to be projected on this map before validating our approach.

4 Discussion and conclusion

In this chapter, a new process to characterize motion organization of the newborn from video recordings was described. It goes from the motion estimation by video processing to the extraction of an informative set of features. It also integrates the removal of irrelevant periods due to the presence of adults on the frame and a motion segmentation method to retrieve intervals of motion and non-motion.

Each method of this process was evaluated on videos acquired in a wide variety of real conditions of NICU (e.g., baby may be covered by a blanket, under respiratory assistance, lying in different beds, in dark environment, with pulse-oxymetry flashing). The adult detection algorithm reaching 96.8% of sensitivity and 95.1% of specificity has been published in IRBM [5]. Two models, based on KNN, were proposed to segment motion and non-motion periods, depending on the acquisition setup. Segmentation on closed beds ($D1$) leads to a sensitivity of 80.7% and a specificity of 88.4%. Results for the second model, designed for opened beds ($D2$), showed better results with a sensitivity of 88.3% and 90.1% of specificity. Besides, the extracted features revealed to be similar between automatic and manual segmentation. It is worthwhile to note that the annotations (adult presences and motion intervals) performed to evaluate methods of this chapter correspond to an heavy task but also conduct to a rich result since annotated database coming from real life is really relevant for further analyses and comparisons.

Additionally, let us indicate that this entire process takes only about 30 minutes to analyze eight hours of video data and requires, in some cases, only two short manual interactions at the initialization step (cropping and border definition).

The whole process was applied on video data from three different groups: very preterm, late preterm and full-term newborns. We showed that the organization of motion evolves with PMA. More specifically, duration of non-motion epochs increases and the number of non-motion epochs decreases while the newborn is growing up. These results are consistent with the fact that the time spent in motion was shown to decrease with PMA in AS and QS [8]. Our observations are also in balance with two expected behavioral developments of the newborn: the sleep cycle duration increases with PMA [8] and after 36 weeks, babies generally show their first moments of calm wakefulness [6]. In addition, we constructed a first version of control map that may be used as a reference to assess the motion organization of newborns. Nevertheless, it will be necessary to project newborn showing critical motion development to evaluate our capacity to detect disorders from video analyses.

These results have to be confirmed on a larger database regarding newborn of different GA as well as a wider diversity of PMA. This will be provided by the Digi-NewB protocol since 313 newborns of all categories were recorded, often several times. To date, only 22 recordings were analyzed for two reasons. The first one is due to the necessary time to recover data since recordings are locally stored in each hospital. The second is that the process still requires manual annotations. Indeed, for most of the recordings of *D1*, parts of the border had to be manually discarded in order to avoid false adult detections. Additionally, absences of the baby have been manually annotated and removed from the analysis. In fact, these periods, where the baby is taken out of the bed, can also impact the motion evaluation and an automatic detection of these events will be necessary to move towards a fully-automated solution for monitoring. This problem has been recently addressed for video recordings of toddlers performed at home [10]. In this study, authors proposed to study the motion on two regions of interest: in and out the bed, in order to detect instants when parents move the baby. However, the regions have to be manually defined for each new recording configurations or field of view. From our part, deep learning solutions will be investigated to classify images as relevant or not (e.g., absence of the baby, adults presence) for analyses.

We can also note that our first version of the pulse-oximetry flashing correction may be improved. In fact, in this chapter, we chose to use signal processing to resolve it. However, the correction slightly deteriorates the motion series, in terms of amplitude and shape, and the computation of a new set of features, for example, characterizing the shape of motion epochs or frequency content will not be representative of the motion reality. Another approach could have been to automatically detect and mask this bright area through video processing.

Finally, it is important to note that this work falls within the objective of combining descriptors to monitor newborn developmental behavior. Although this first set of features shows its ability to ride up informative elements about the evolution of motion organization in newborns, it could be enhanced by the addition of relevant information about newborn cries. This is the object of the next chapter.

Bibliography

- [1] ABDI, H. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics 3* (2007), 103–107.

-
- [2] ADDE, L., HELBOSTAD, J. L., JENSENIUS, A. R., TARALDSEN, G., GRUNEWALDT, K. H., AND STOEN, R. Early prediction of cerebral palsy by computer-based video analysis of general movements: A feasibility study. *Developmental Medicine & Child Neurology* 52, 8 (2010), 773–8.
- [3] ADDE, L., HELBOSTAD, J. L., JENSENIUS, A. R., TARALDSEN, G., AND STOEN, R. Using computer-based video analysis in the study of fidgety movements. *Early human development* 85, 9 (2009), 541–7.
- [4] CABIN, R. J., AND MITCHELL, R. J. To bonferroni or not to bonferroni: when and how are the questions. *Bulletin of the Ecological Society of America* 81, 3 (2000), 246–248.
- [5] CABON, S., POREE, F., SIMON, A., UGOLIN, M., ROSEC, O., CARRAULT, G., AND PLADYS, P. Motion estimation and characterization in premature newborns using long duration video recordings. *IRBM* 38, 4 (2017), 207–213.
- [6] CHALLAMEL, M.-J., FRANCO, P., AND HARDY, M. *Le sommeil de l'enfant*. Elsevier Masson, 2009.
- [7] CUPPENS, K., LAGAE, L., CEULEMANS, B., HUFFEL, S. V., AND VANRUMSTE, B. Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy. *Medical & Biological Engineering & Computing* 48, 9 (sep 2010), 923–931.
- [8] CURZI-DASCALOVA, L. Physiological correlates of sleep development in premature and full-term neonates. *Neurophysiologie Clinique/Clinical Neurophysiology* 22, 2 (1992), 151–166.
- [9] KARAYIANNIS, N. B., AND TAO, G. An improved procedure for the extraction of temporal motion strength signals from video recordings of neonatal seizures. *Image and Vision Computing* 24, 1 (2006), 27–40.
- [10] LONG, X., VAN DER SANDEN, E., PREVVOO, Y., TEN HOOR, L., DEN BOER, S., GELISSEN, J., OTTE, R., AND ZWARTKRUIS-PELGRIM, E. An efficient heuristic method for infant in/out of bed detection using video-derived motion estimates. *Biomedical Physics & Engineering Express* 4, 3 (2018), 035035.
- [11] NAKAGAWA, S. A farewell to bonferroni: the problems of low statistical power and publication bias. *Behavioral ecology* 15, 6 (2004), 1044–1045.
- [12] NAVARRO, X., PORÉE, F., KUCHENBUCH, M., CHAVEZ, M., BEUCHÉE, A., AND CARRAULT, G. Multi-feature classifiers for burst detection in single EEG channels from preterm infants. *Journal of Neural Engineering* 14, 4 (2017), 046015.
- [13] NING, X., SELESNICK, I. W., AND DUVAL, L. Chromatogram baseline estimation and denoising using sparsity (beads). *Chemometrics and Intelligent Laboratory Systems* 139 (2014), 156–167.
- [14] NTONFO, G. M. K., FERRARI, G., RAHELI, R., AND PISANI, F. Low-complexity image processing for real-time detection of neonatal clonic seizures. *IEEE Transactions on Information Technology in Biomedicine* 16, 3 (2012), 375–382.

- [15] RAHMATI, H., DRAGON, R., AAMO, O. M., ADDE, L., STAVDAHL, Ø., AND VAN GOOL, L. Weakly supervised motion segmentation with particle matching. *Computer Vision and Image Understanding* 140 (2015), 30–42.
- [16] SÄRKELÄ, M., MUSTOLA, S., SEPPÄNEN, T., KOSKINEN, M., LEPOLA, P., SUOMINEN, K., JUVONEN, T., TOLVANEN-LAAKSO, H., AND JÄNTTI, V. Automatic analysis and monitoring of burst suppression in anesthesia. *Journal of Clinical Monitoring and Computing* 17, 2 (2002), 125–134.
- [17] TEKALP, A. M. *Digital video processing*. Prentice Hall Press, 2015.

AUDIO-BASED CHARACTERIZATION OF NEWBORN CRIES FOR NEURO-DEVELOPMENTAL MONITORING

In Chapter 2, we saw that analyses of spontaneous cries are an informative tool to assess the development of premature newborns. More precisely, some authors discussed relationship between frequency content and increasing gestational age [8] and found that fundamental frequency are generally higher in preterm than in full-term newborns at term-equivalent ages [11, 16].

Nevertheless, in these studies, authors focus their analyses either after a manual extraction of cry events [16] or on short recordings [8, 11]. Nevertheless, for monitoring purpose, it is first necessary to automatically extract newborn cry events from audio recordings.

In this chapter, we present the strategy that we developed to retrieve spontaneous cry of newborns. Then, the process is evaluated and preliminary results regarding behavioral development are given.

1 Methods to extract newborn cries

Extracting newborn cries is challenging due to the fact that in NICU, several sounds from other sources (e.g., alarms, adult voices) can occur. In addition, within our purpose, the analysis has to be performed continuously over long periods. To deal with this, a two-step strategy was proposed (Figure 7.1):

1. Segmentation between silence and sound events;
2. Selection of the newborn cries.

First, we decided to use the segmentation method proposed in [13] and introduced in Section 2.2 of Chapter 4. This method leans on energy thresholding where two thresholds are automatically estimated by the mean of the Otsu method [14]. When applied to an audio recording containing only cries, it will returned only cry intervals, such it was the case in our preliminary study on sleep stage estimation. However, when applied on long recordings, irrelevant sounds segments are also extracted. Thus, we decided to apply it, as a first step, to extract segments of interest from silence periods from recordings of long duration. Only segments lasting more than 100 milliseconds were considered. It has been arbitrary chosen since no consensus on the minimal length of a cry event was revealed by the literature review. In studies dealing with premature newborns, values were ranged from 150 to 260 milliseconds [7, 11, 13].

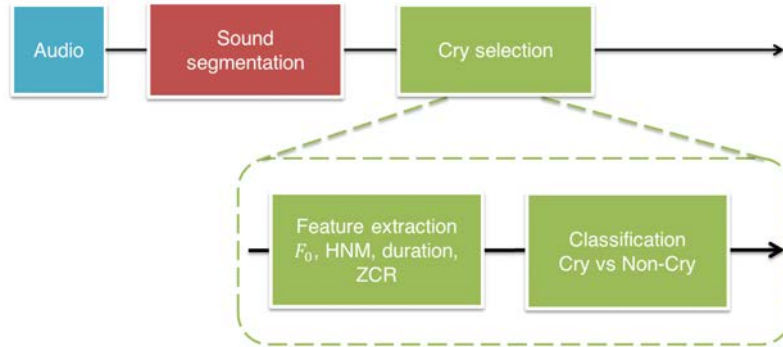


Figure 7.1 – Global workflow of cry extraction.

Then, in order to provide a relevant characterization of behavioral development, segments of cries have to be automatically recognized among them. For that purpose, machine learning techniques were investigated in order to perform a cry/non-cry classification. In this chapter, we focus on the presentation and the evaluation of this second step since the segmentation was already validated by its authors. First, the database used to train models is presented, the feature set used for classification is then described and the classifiers that have been investigated are finally introduced.

1.1 Annotated database

We selected a set of data representing a large part of the diversity encountered in the project. Hence, 27 recordings were selected from the Digi-NewB database. They involve fourteen boys and seven girls born between 27+5 and 41+4 GA and recorded between 28+5 and 41+5 PMA. Some of them were recorded two times. Recordings were performed in four hospitals: Rennes (1), Angers (2), Brest (3) and Tours (6), on both types of beds: open or close.

From each audio recording, segments were first extracted by the mean of the segmentation step. Then, based on the resulting segments and on previous works [1, 15], six classes of sounds have been defined. Figure 7.2 summarizes the annotated data.

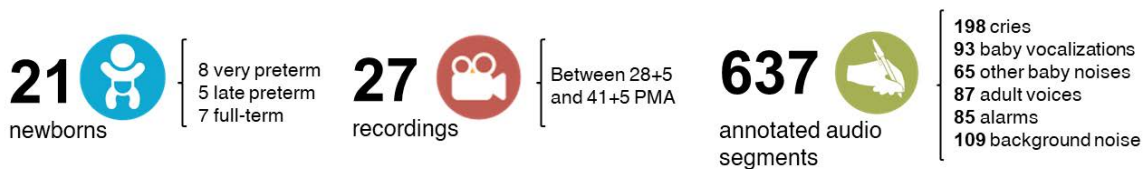


Figure 7.2 – Overview of the annotated audio data.

We chose to annotate segments containing only one type of event, meaning that segments with overlapping sounds (e.g., cry with adult voice) were not selected. In order to construct a dictionary for

the learning and evaluation phases of the classification methods, a total of 637 segments have been manually labeled.

Sounds emitted by the newborns were declined in three categories: cries, vocalizations (e.g., cooing) and other baby noises such as coughing or hiccups. The auditory differences between vocalizations and cries are subjective and, thus, only obvious cries were annotated as such. Hence, 198 cries were extracted from all recordings except one of them since no cry was found in it. The number of cries per recording ranged from 2 to 12 for the others. For the other two categories, baby vocalizations and baby noises, 93 and 65 segments, from all recordings, were respectively annotated.

Others sounds were also classified into three categories: adults voices (87 segments), alarms from devices (85 segments) and background noises (109 segments). The most diverse segments were selected to be part of each category. Thus, men and women voices, several types of alarms, many background noises coming from the adults activity (e.g., doors opening/closing, packaging friction, water flowing from the tap) as well as from devices (e.g., ventilatory support airflow, bed adjustment noises), were selected.

1.2 Feature engineering

We saw in Section 3.2.2 of Chapter 2 that three types of features can be used to characterize cry signals: time, frequency and time-varying frequency features. In recent works, authors mainly used Mel-Frequency Cepstral Coefficients (MFCCs) to describe audio signals for classification. In some cases, preprocessing steps were first applied in order to reduce the effect of noise, such as beamformer [4] or signal decomposition [2]. Coefficients were then computed frame by frame. Finally, each frame was classified by taking into account adjacent frames either using CNN [4] or using HMM [2, 10].

On our side, we chose to integrate noise by modeling audio segments using Harmonic plus Noise Model (HNM) that is commonly applied in speech synthesis [17]. HNM analysis is known to be more suitable for quasi-harmonic signals such as baby cries and vocalizations, adult voices or alarms than for background noises or other baby noises. In addition to that, in HNM, it is necessary to limit the analysis in a certain frequency band $[F_{min}, F_{max}]$ that we chose to adapt to cry analysis. The underlying hypothesis behind this choice is that an analysis focused on extracting characteristics relevant to cry analysis will give discriminating features for classification in case of other types of sounds.

Once the signal has been modeled, MFCCs and fundamental frequency are extracted. They are supplemented by time features computed from the original signal. An audio segment is then summarized by the median values of each feature over each analyzed frame.

1.2.1 Harmonic plus Noise Model features

Model description HNM is used to split speech records into small parts called phonemes that will be reorganized to construct new speeches. The principle of HNM analysis is to create a synthetic signal $s(t)$ composed of harmonic $h(t)$ and noise $n(t)$ parts that fit the original signal such as:

$$s(t) = h(t) + n(t) \tag{7.1}$$

In fact, for a voiced speech signal, the spectrum can be divided into two bands delimited by the so-called maximum voiced frequency $F_m(t)$, a time varying parameter. The lower band of the spectrum is represented by the harmonic part and the upper band by the noise part. The harmonic part $h(t)$ is modeled as a sum of harmonics such as:

$$h(t) = \sum_{k=1}^{K(t)} A_k(t) \cos(k\theta(t) + \Phi_k(t)) \quad (7.2)$$

where $A_k(t)$ and $\Phi_k(t)$ are respectively amplitude and phase at time t of the k -th harmonic. $K(t)$ represents the time-varying number of harmonics included in the harmonic part. On its part, the frequency content of the noise part $n(t)$ is described by a time-varying auto-regressive envelope where its time-domain structure is represented by a piece-wise linear energy envelope function and is defined as:

$$n(t) = e(t)[f(t, \tau)u(t)] \quad (7.3)$$

where $u(t)$ is a white Gaussian noise and $f(t, \tau)$ is a time-varying, normalized all-pole filter.

Parameters of the model First, we decided to conduct the analysis on short frames of 5 milliseconds without overlap for each audio segment. Then, in order to fit the original signal for each frame, several parameters which regulate the model have to be estimated. The analysis is performed in four steps, described hereafter.

Estimation of the fundamental and maximum voiced frequencies The first step of the analysis consists in estimating the fundamental frequency F_0 and the maximum voiced frequency $F_m(t)$. This operation is performed in four phases: initial estimation of F_0 , voiced/unvoiced decision, estimation of the maximum voiced frequency and refining the initial F_0 estimation.

In our case, we chose to estimate an initial fundamental frequency by the mean of Continuous Wavelet Transform, as presented in [12]. This choice is mainly justified by the quasi-stationary of cry signals. Continuous Wavelet Transform of a function $f(t)$ is defined as:

$$W(a, b) = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{|a|}} \Psi^* \left(\frac{t-b}{a} \right) \quad (7.4)$$

where $\Psi(t)$ is called the "mother" wavelet (here, Mexican Hat is used) and $*$ stands for the conjugate operation. For their part, a and b are, respectively, the shift parameter that locates the wavelet in time and the scale parameter which regulates the width of the mother wavelet (i.e., it compresses or dilates the wavelet) to generate daughter wavelets. The relationship between the scale and the pseudo-frequency F_{eq} of a resulting wavelet is given by:

$$F_{eq} = \frac{F_c \cdot F_s}{a} \quad (7.5)$$

where F_c is the central frequency of the Mexican Hat wavelet that maximizes the module of its Fourier Transform and F_s is the sampling frequency. Therefore, the fundamental frequency estimation is performed as follows:

1. Filtering of the signal with a band-pass ($F_{min} = 150$ Hz and $F_{max} = 750$ Hz) filter of finite impulse response (FIR);
2. Computing the CWT matrix of coefficients (size: 27×80), corresponding to an equivalent frequency varying between F_{min} and F_{max} (a is allowed to vary in the range 1-27) and b changing along the window length (80 samples);
3. Finding the best scale a , that is equivalent to find the row containing the maximum of the CWT matrix;
4. Estimating F_0 by computing the autocorrelation of the row, according to:

$$F_0 = \frac{F_s}{\delta} \quad (7.6)$$

where δ is the lag relative to the autocorrelation maximum.

Finally, to avoid spurious peaks in the F_0 estimations along the audio segment, a three points moving average smoothing function is applied.

Based on this first estimation, a voicing decision is made. A synthetic signal is constructed using the estimated pitch and its associate harmonics. Their amplitudes and phases are estimated by Discrete Fourier Transform algorithm. Then, the normalized error over the first four harmonics in the original spectrum $S(f)$ and the synthetic spectrum $\hat{S}(f)$ is compared to a given threshold, here -15 dB. The error is calculated as follow:

$$E = \frac{\int_{0.7\hat{F}_0}^{4.3\hat{F}_0} (|S(f)| - |\hat{S}(f)|)^2}{\int_{0.7\hat{F}_0}^{4.3\hat{F}_0} |S(f)|^2} \quad (7.7)$$

where \hat{F}_0 is the initial fundamental frequency. If the error E is below the threshold, the window is marked as voice, otherwise, as unvoiced.

On voiced windows, the maximum voiced frequency $F_m(t)$ is estimated. A peak picking procedure is applied to separate the two bands of the spectrum $\hat{S}(f)$: the lower band which contains the quasi-harmonic frequencies and the upper band containing the noise component.

Finally, from the initial fundamental frequency estimation and the frequencies f_i of the lower band, the fundamental frequency is refined. It is defined as the value that minimizes the following error criteria:

$$E(\hat{F}_0) = \sum_{i=1}^{L_n} |f_i - i \cdot \hat{F}_0|^2 \quad (7.8)$$

where L_n is the number of the voiced frequencies f_i .

Estimation of the parameters of the harmonic part For voiced frames, harmonic amplitudes $A_k(t)$ and phases $\Phi_k(t)$ can be estimated at the center time instant t_i of the analysis windows. Conse-

quently, from Eq.(7.2), $h(t)$ may be expressed as:

$$h(t) = \sum_{k=1}^L A_k(t_i) \cos(kt\omega_0 + \Phi_k(t_i)) \quad t_i - \frac{N}{2} \leq t \leq t_i + \frac{N}{2} \quad (7.9)$$

where N is the length of the windows in samples and $L = K(t_i) = \frac{F_m(t)}{\omega_0(t_i)}$ is the number of harmonics included in the harmonic part. Then, the estimation is made by minimizing a weighted time-domain least-squares criterion between signal $s(t)$ and $h(t)$:

$$\min_{A_k(t_i), \Phi_k(t_i)} \sum_{t=-\frac{N}{2}}^{\frac{N}{2}} \alpha(t)(s(t) - h(t))^2 \quad (7.10)$$

where $\alpha(t)$ is a Hamming window.

Estimation of the parameters of the noise part The next step of the HNM analysis is the estimation of the noise part parameters. For each analysis window (voiced or unvoiced), the spectral density function of the original signal is modeled by a 20th-order all pole filter $F(t_i, z)$, by the mean of a standard correlation-based method. The correlation function is estimated from 40 milliseconds of signal located around the center of the analysis window. Over the same duration, the variance of the original signal is also estimated, representing the gain of the filter. For that purpose, a triangular like time-domain envelope is employed.

Cepstral analysis Finally, a spectral conversion function is applied on the harmonic part to extract Mel-Frequency Cepstral Coefficients (MFCCs) that characterize the spectral envelope of the signal, such as presented in Chapter 2 (Section 3.2.3).

1.2.2 Time features

Two time features have been also defined to characterize the segments:

- **Total duration** in seconds of the segment since some events may last longer (e.g., adult speech) or take less time (e.g, beep) than typical cries;
- **Zero Crossing Rate ZCR**, already shown to be useful to distinguish alarms from cries [19, 20]:

$$ZCR = \frac{1}{T-1} \sum_{i=1}^{T-1} 1_{R<0}(s_t s_{t-1}) \quad (7.11)$$

where T is the length of the signal $s(t)$ and $1_{R<0}$ is the indicator function.

1.2.3 Synthetic resume of the set of features

HNM analysis is performed on frames of 5-milliseconds over each segment. In order to obtain only one value for each feature, the median has been computed. In total 124 features were computed. In our

case, some of these features keep for all studied segments a null value (e.g., related to high harmonics) and have been discarded from the final feature set. Table 7.1 synthesizes the remaining 73 parameters.

Table 7.1 – List of the computed features for classification

Type of feature	Estimation Method	Number of instances
Fundamental frequency	HNM	1
Number of harmonics	HNM	1
Harmonic amplitudes	HNM	18
Harmonic phases	HNM	14
Gain	HNM	1
Filter coefficients	HNM	20
Cepstral coefficients	HNM	16
Zero Crossing Rate	ZCR	1
Duration	Duration	1

This feature set appears numerous even if it is a classical approach in speech processing. Thus, in the following section, we choose to reduce its dimension.

1.2.4 Dimensionality reduction

As mentioned in Chapter 3, a common problem in classification is overfitting. This occurs when a classifier corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data. This can be due to a too high number of features describing each sample. To prevent this situation and reduce the dimension p of our feature set, we arbitrary chose to apply Principal Component Analysis (PCA) ¹.

Before PCA, a transformation need to be performed on the feature set in order to work with features on the same scale. This way, the prevalence of a feature in the dimensionality reduction process is avoided. In fact, if the differences between values within a feature are wider than for others, the variance in the whole dataset would be mostly explained by this feature although it may be incorrect.

To do so, several techniques exist such as min-max normalization or standard scaling. In our case, we chose to applied standard scaling since the maximum and the minimum values that can take each feature may not be present in our dataset. On its side, standard scaling is based on the mean and the variance of each feature. Hence, the standardized set of features Xn_{if} is computed for each sample i , and is defined as:

$$Xn_{if} = \frac{X_{if} - m_f}{s_f} \quad (7.12)$$

where X_{if} is the original value of a feature f , with $f \in [1, \dots, p]$. The mean m_f and the variance s_f of the feature are computed from all samples of the dataset.

Then, there are two ways of performing dimension reduction with PCA, either by keeping a share of the total variance or by targeting a number of principal components. In our case, we chose to keep the principal components that represent 95% of the total variance.

1. Indeed, the objective of our method being to classify cry and non-cry events, it is not necessary to recall the physiological meaning of the feature and thus, as a first step, the attention can be focused on feature extraction techniques.

1.3 Classifiers

Contrary to the methods of the literature that generally exploit the notion of temporality on not segmented signal, we chose to compare six commonly used classifiers, introduced in Chapter 3: Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbors, Random Forest, Multi-Layer Perceptron and Support Vector Machine. Indeed, temporality within the same event is not relevant in our case since we focus on already extracted audio segments that are summarized by median values of the features.

2 Evaluation of the process

2.1 Evaluation strategy

Our goal being to identify cry segments from others, a binary classification was performed. Within this objective, our annotated dataset was composed of 198 cries and 439 other sounds. Then, a training/validation set and a testing set were defined, respectively composed of 60% and 40% of the dataset.

On the training/validation set, a k-fold cross validation strategy was performed in order to tune parameters of the classifiers. The number of splits was set to three. Contrary to attempts of classification presented in previous chapters, this time, parameters were tuned within two objectives: in order to reach the highest accuracy (strategy 1) and the highest precision (strategy 2). In the first case, the idea is to performed the best classification between cry and non-cry events whereas in the second case, the lowest false positive rate is sought, i.e. recover as few non-cry segments as possible.

To finish, the generalization of each classifier was evaluated on the testing set through three metrics: the precision, the recall and the accuracy.

2.2 Dimensionality reduction

In Chapter 3, Principal Component Analysis has been presented as a popular method for feature extraction and visualization of data.

First, data has been projected into a 2-dimensional space for visualization. Resulting graph is presented in Figure 7.3.

Principal Component 1 and Principal Component 2 explain respectively 8.8% and 6.1% of the variance, for a total of 14.9%. One can see that baby cries are mostly located on the bottom of the graph with some of the baby vocalizations while other sounds tend to be in the upper band. This observation is reassuring regarding our classification purpose although we may already notice that it will be difficult to discriminate between baby cries and vocalizations on the sole basis of these two dimensions.

In a second time, we projected the feature set into principal components that represent 95% of the total variance. For that purpose, 41 components are retained. Hence, the dimension of the feature set goes from 73 to 41, reducing the number of features by 32. Figure 7.4 illustrated the coefficients applied on each feature in order to perform the projection.

Globally, each feature contributes to the projection and none of them stands out from the crowd. One can also see that the most influential features on the first three dimensions are related to the

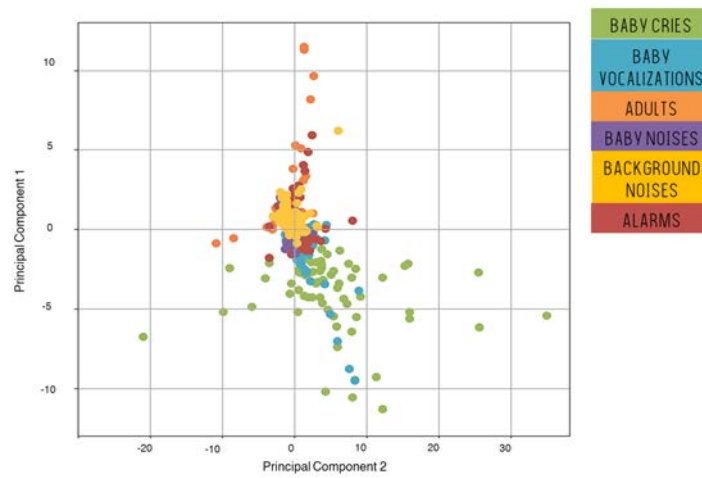


Figure 7.3 – Visualization of the dataset using the first two principal components.



Figure 7.4 – Heatmap reporting the value of the coefficients applied to project the original feature set to 41 principal components. Color represents the value of the coefficient. The deeper is the blue, the higher is the coefficient applied for a feature.

harmonics (the three first lines in Figure 7.4). These observations mean that information is contained in all features. However, it is important to remind that in our case, we used an unsupervised approach to reduce dimensionality with the sole goal to prevent overfitting. The contribution of each feature to the projection has thus no direct link with its ability by itself to be discriminant for classification.

To conclude, at this stage, we can only affirm that the projected feature set contains information from the whole original feature set that may be useful for classification. Indeed, visually, we saw that it was possible to distinguish between cries and others sounds. The relevance of the projected feature set for this purpose will be assessed along with classifiers evaluation in the following sections.

2.3 Tuning of the parameters

For each classifier, two sets of parameters, respectively, resulting to the highest accuracy (strategy 1) and the highest precision (strategy 2) during cross-validation were identified. For that purpose, several parameters and hyper-parameters have been tuned. This time, in order to compare between kernels in SVM, we chose to report the best parameters for the three SVM classifiers: linear, polynomial and Gaussian. A summary of the tests is reported in Table 7.2.

Table 7.2 – Parameters testing summary. Final selecting sets of parameters for precision are marked in bold whereas final selecting sets of parameters for accuracy are marked in gray (bold and gray means that the same parameter is obtained for both strategies).

Method	Parameters
KNN	Number of neighbors $\in [1, 3, 5, \mathbf{11}, 15]$ Distance: Manhattan or Euclidean
LDA	Solver \in [singular value decomposition, least squares solution , eigenvalue decomposition]
LR	Cut-off $\in [\mathbf{0.1}, 0.2, 0.5, 0.7]$
RF	Number of trees $\in [5, 10, 20, 50, 100, \mathbf{300}]$ Quality split criterion: gini or entropy
MLP	Number of hidden layers $\in [1, 2, 5]$ Number of perceptrons per layers $\in [1, 2, 5, 10, \mathbf{20}, 30]$ Activation function \in [identity, logistic sigmoid, hyperbolic tan, rectified linear unit]
SVM linear	No additional parameter
SVM polynomial	degree $\in [1, 2, \mathbf{3}, 4]$
SVM Gaussian	margin $\in [0.01, 0.1, \mathbf{1}, 10, 100, 10^3, 10^4]$ gamma $\in [\mathbf{0.0001}, 0.001, 0.01, 0.1, 1, \mathbf{5}, 10, 100]$

One can see that the parameters can be different regarding the classification objectives. This recalls the fact that this step is inescapable during the learning phase of machine learning and must be done according to the objective.

2.4 Classification results

In Figure 7.5, results on the test set regarding strategy 1 (accuracy) are given. The best results are obtained with MLP that presents 95.3% of accuracy, 92.9% of recall and 92.8% of precision. Generally, non-linear algorithms show weaker results than linear approaches, especially on recall (48.8% for SVM Gaussian, 52.4% for SVM polynomial and 60.7% for RF). This reveals a bad generalization of those models.

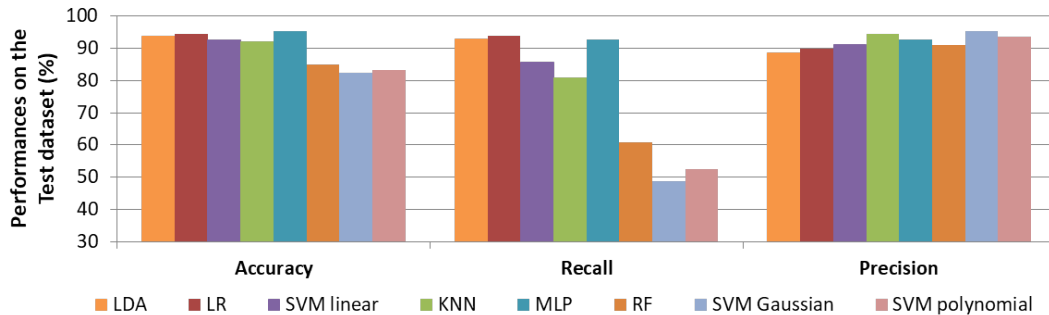


Figure 7.5 – Performances (in %) of cry selection on the test set for each machine learning approach by maximizing the accuracy in the learning phase.

The best recall value is obtained by LR with 94.0%, which is slightly higher than with MLP. KNN and SVM linear classifiers perform well but results on recall are below with values at 85.7% and 81.0%.

Results for strategy 2 are reported in Figure 7.6. A better generalization for all models is observed since values are more stable between metrics. The best precision score is obtained by KNN and reaches 92.9%. The highest recall score is once again obtained by LR, with 94.1%. Accuracies are high (above 90.2%) for all classifiers. The results with MLP are also really good since a precision of 92.7%, a recall of 90.48% and an accuracy of 94.5% are reached.

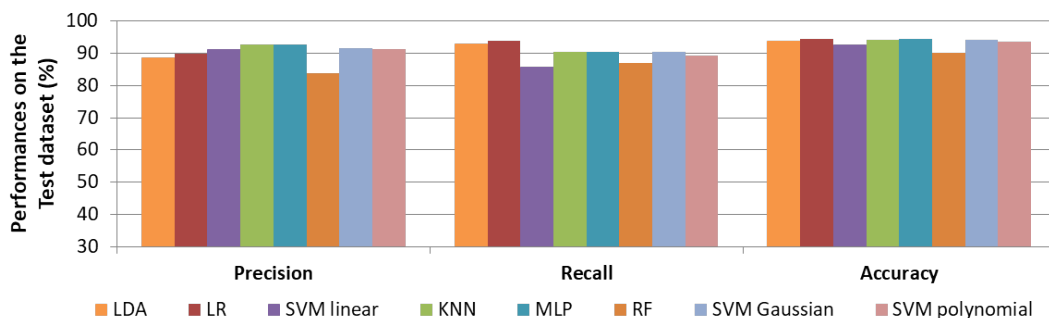


Figure 7.6 – Performances (in %) of cry selection on the test set for each machine learning approach by maximizing the precision in the learning phase.

With regard to our objective, a classifier that have learned to be precise on cry selection (strategy 2)

seems to be the most reasonable choice. Indeed, the goal is to extract cries in order to assess newborn evolution during hospitalization, in other words, over the long term. In that case, there is no need to get all the cries as long as there is a high chance that the predicted cries are actual cries. The results of the KNN of the second strategy are in line with this. In fact, it carries a high precision while keeping a high recall value. This means that there is also a low chance of missing cries with this model.

2.5 Evaluation of the model accuracy for fundamental frequency analyses

In this section, the efficiency of the model in cry selection during the deployment is assessed. Secondly, since in Chapter 2, we saw that cries were mainly investigated regarding the fundamental frequency F_0 . Thus, the accuracy of our method within this purpose is discussed regarding cry characterization. In addition, the impact of the misclassifications on the F_0 estimates over long periods is also studied.

Evaluation of the model during deployment For that purpose, the KNN model was applied on all 8 hour-recordings of the maturation dataset that was introduced in Chapter 6 (Section 3). Results are reported in Figure 7.7.

	BABY	GA (IN WEEK+ DAY)	DAY	NUMBER OF EXTRACTED CRIES	ACTUAL CRIES (%)	ACTUAL CRIES + OVERLAPPED CRIES + VOCALIZATIONS (%)
GROUP 1	010108	27+5	D1	0	-	-
			D2	302	70	100
	010066	28+2	D1	1	0	100
			D2	86	83	100
	030033	29+0	D1	295	46	63
			D2	21	57	76
	060007	29+2	D2	389	63	99
	010081	29+5	D1	50	68	86
GROUP 2			D2	28	68	100
	030035	29+6	D1	578	94	100
			D2	125	87	98
	010075	33+1	D1	35	54	85
			D2	537	76	99
	010076	33+3	D1	40	10	20
	030011	33+5	D1	89	56	70
			D2	3	0	0
GROUP 3	010086	35+4	D1	360	71	95
	010084	36+2	D1	377	68	86
			D2	393	77	99
	010064	36+5	D2	11	0	54
	010098	39+4	D2	53	87	94
	010087	41+4	D2	111	62	94

Figure 7.7 – Results of automatic cry selection on the maturation dataset. Number of extracted cries over the 8 hours are reported for each recording, as well as the percentage of actual cries among them. It is supplemented by the percentage including actual cries, overlapped cries and vocalizations.

As a first step, segments that were automatically classified as cries were verified and the percentage of actual cries, i.e. the precision, was computed for each recording. Resulting percentages are ranging

from 0% to 94% with a median value of 68%. The results are very variable since high results were obtained for four recordings (above 80%), good results for nine recordings (between 60% and 80%), moderate for four (between 40 and 60%) and weak for four of them (under 10%). Finally, a total of 3941 segments were automatically extracted and 1126 of them revealed to be false positives, giving an error rate of 25%. The distribution of errors is reported in Figure 7.8.

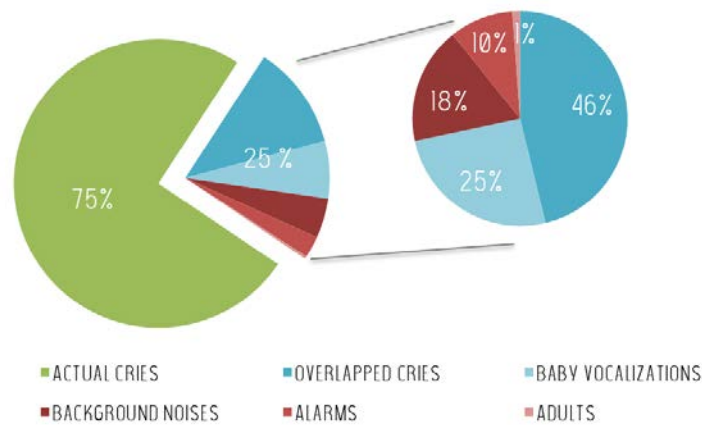


Figure 7.8 – Cry selection results during the deployment. Percentages of good and misclassifications regarding each class.

The main source of errors comes from segments on which other sound events happened in the same time than a cry (overlapped cry). It represents 46% of the errors. This was not surprising since the model was not trained for this purpose. Secondly, as expected, errors are made between cry and vocalizations (25% of the errors). Nevertheless, 18% of the errors are also made by confusing cries and background noises, 10% are due to alarms and 1% are from adults.

From these observations, percentages were recomputed by integrating vocalizations and all the cries (actual and overlapped). In that case, the median value reaches 94% and the error rate drops at 8%. For only two recordings, the percentage of actual cries and vocalizations stays under 20%. The worst results are obtained for *D2* of 030011 where three segments were extracted and none of them was a cry nor a vocalization. Indeed, two among them are segments of adult voice and one is a background noise. For *D1* of 010076, 19 segments of background and 13 alarms were misclassified.

An additional limitation was found during the deployment of the model. It occurs when several babies share a room, such as in the recording of 060007. In fact, cries of three different babies were extracted without distinction since the model is not trained for this purpose. During the verification, segments classified as cries for this situation were considered as actual cries.

To complete these observations, we can also notice that the processing duration depends on the content of each recording, notably because of the time necessary to perform HNM analysis for classification. It is directly linked to the number of segments coming from the initial segmentation. The higher the number of segments that need to be classified is, the longer will take the analysis. As an exam-

ple, for the $D2$ recording of 030033, 29069 segments had to be classified and the computation lasted 205 minutes. Reversely, 174 segments were extracted for $D1$ of 010066 and they were classified in 12 minutes. These computing performances reveals that a quasi-real time monitoring is reachable.

Estimation of the fundamental frequency To go further, the fundamental frequency estimation for cry characterization was studied. In our case, the fundamental frequency was estimated between 150 and 750 Hz in order to be integrated in the feature set for cry selection. Although this estimation was relevant for classification, further observations revealed that the band applied for the estimation of F_0 was not accurate for cry characterization. Indeed, the band of analysis had to be adapted for each cry. To illustrate this point, three examples of cries are given in Figure 7.9.

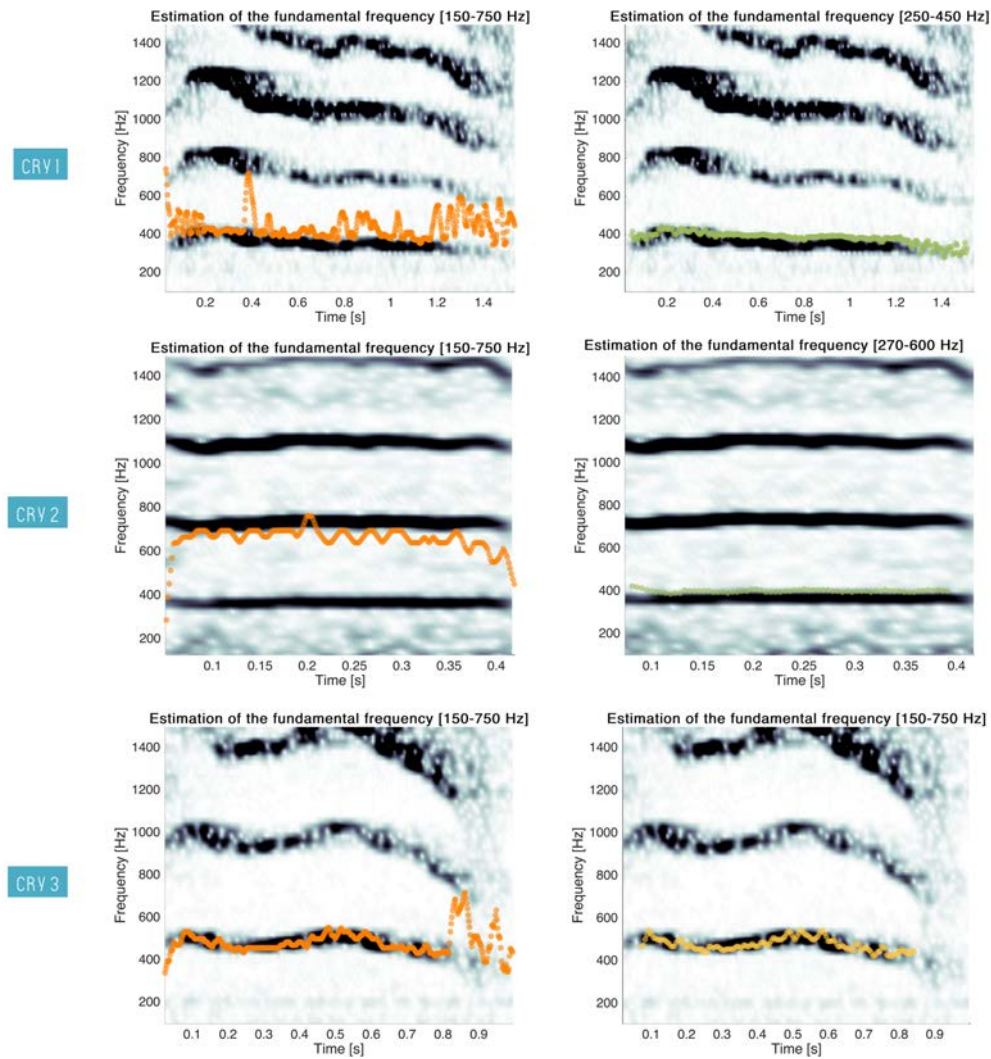


Figure 7.9 – Three examples of cry characterization using F_0 estimations. Each time, the estimation either with the fixed band 150-750 Hz (in orange) or with a manually selected band (in green) as well as with smoothing (in yellow) are superimposed on the spectrogram of the cries.

For Cry 1, several jumps between the fundamental frequency and the first formant can be observed when the band 150-750 Hz was used. A manual selection of the band 250-450 Hz allowed for a more accurate estimation of F_0 over the cry. For Cry 2, F_0 is overestimated and is found on the first formant all along the cry. The use of a band 270-600 Hz provided better estimates.

In Cry 3, noisy parts are observed at the beginning and at the end of the segment, leading to wrong estimations of F_0 . In [7], authors submitted the hypothesis that these noisy parts are due to the difficulty of the infant to initiate phonation and to inhaling and proposed to smooth the estimates. This method was composed of three steps:

1. removing the outliers by keeping all the estimates between the 25th and the 75th percentile of the distribution over the cry (Tukey Method [18]).
2. removing isolated outliers by studying previous and the next value of each F_0 estimates;
3. removing the highly irregular estimations in the first 5% and the last 5% of the cry.

The effect of the smoothing is exemplified with Cry 3. In this case, the smoothing removed the bad estimates. Thus, the estimation started after and ended before noisy parts of the segment.

We can also notice that the smoothing could have been useful to retrieve an accurate estimation of F_0 with the fixed band in Cry 1 but that it will be useless for Cry 2. Hence, a combination of an adaptive band of estimation and a smoothing seems to be the best approach to correctly characterize cries.

Impact of the sound superposition on the estimation of the fundamental frequency. To go one step deeper, the impact of overlapping sounds on the estimation of the fundamental frequency of cries has been studied. Fundamental frequencies were estimated with a manual selection of the estimation band and smoothing for six examples of cries. Results are reported in Figure 7.10.

First, we can see that an alarm either of high (e.g., Cry+Alarm1) or low (e.g., Cry+Alarm2) frequency has no impact on the estimation of the fundamental frequency. Secondly, adults have usually lower fundamental frequencies than babies, from 85 to 180 Hz for men, and from 165 to 255 Hz for women [3]. In most of the cases, the estimation of F_0 is well performed (e.g., Cry+ Adult2). Nevertheless, the first formant of adult voices can be in the same order than F_0 and thus, may induce errors in the estimation (e.g., Cry+Adult1). Finally, although background noises impact a large frequency band, the estimation of F_0 remains quite accurate in the affected periods of cries (e.g., Cry+Background1 and Cry+Background2).

If estimations of F_0 over a cry are globally not impacted by the overlapping sounds, another limitation coming from the segmentation step can be noticed in case of multiple events. Indeed, if another sound begins earlier and/or stops after the end of the cry, the algorithm will retrieve all the sound activity as a sole segment, as depicted in Figure 7.11.

In this example, three types of sounds are present: cry, alarm and vocalization. First, the segmentation started before and continued after the first cry event because of an alarm. Then, the second cry event began, immediately followed by a new cry and by an alarm. Finally, a fourth cry occurred and was also followed by a vocalization. In that case, the estimation of the fundamental frequency is performed all along the segment and noise parts of the spectrum are not avoided or corrected by the smoothing.

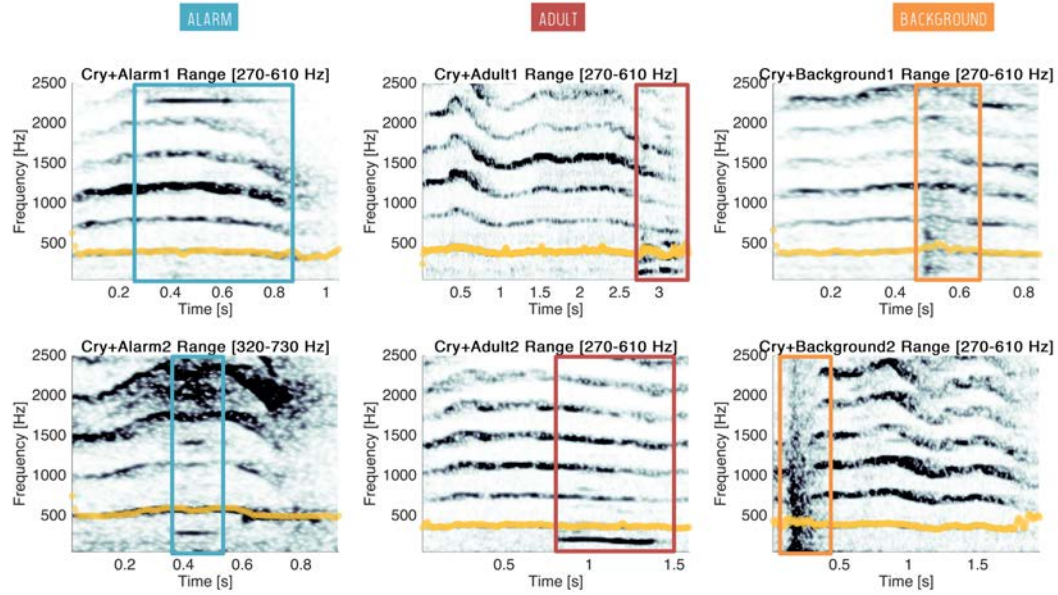


Figure 7.10 – Six examples of overlapped cries by another sound: two alarms (in blue), two adult voices (in red) and two background noises (in orange). Estimates of the fundamental frequency are superimposed on the spectrogram of the cries (in yellow).

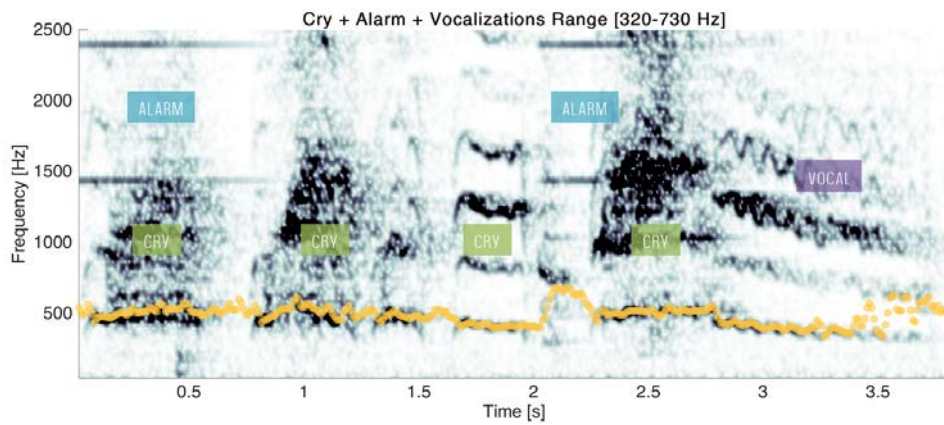


Figure 7.11 – Example of segment with several sounds and the F_0 estimates (yellow).

These errors may be identified since the resulting segments last usually longer (here, more than 3 seconds). For segments of long duration, it may be relevant to perform the extraction process another time. Results for the segment presented in Figure 7.11 is reported in Figure 7.12.

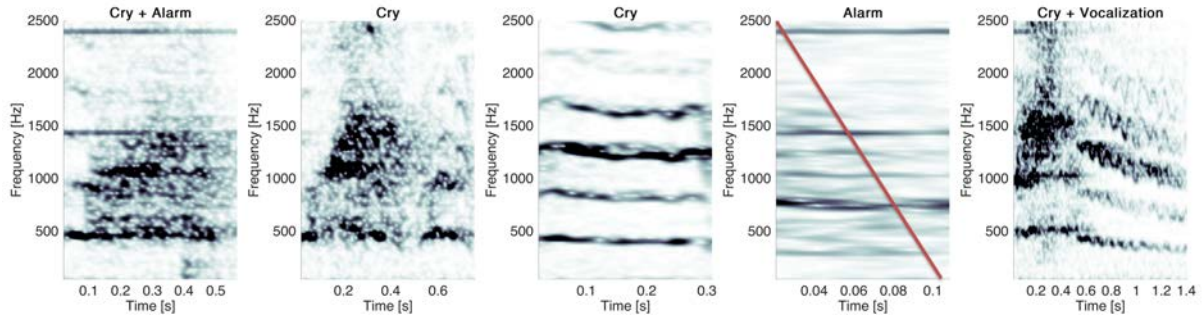


Figure 7.12 – New computation of the cry extraction method (segmentation and classification) for a segment of long duration. The red line indicates the segment that was automatically discarded.

The segmentation step resulted into five segments: one overlapped cry, two cries, one beep and a cry ending with a vocalization. After classification, the four segments containing cry were retrieved.

Impact of confounding vocalizations and cries on the estimation of the fundamental frequency.

The second type of errors made by the classifier is confounding some vocalizations with cry events. An example of misclassified vocalization is reported in Figure 7.13.

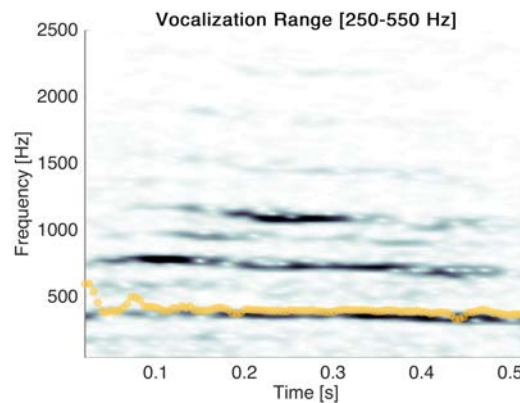


Figure 7.13 – Example of vocalization and the F_0 estimates (yellow).

Misclassified vocalizations have a spectrum similar to that of cries regarding low frequencies. Among these errors, no oscillatory pattern such as observed in Figure 7.11 was observed. In fact, these events are usually short and with values of F_0 in the same range than cries. Hence, the inclusion of these segments may have no impact on the estimation of F_0 over long periods of monitoring.

Impact of other errors on the estimation of the fundamental frequency. The last part of errors, representing 8%, are made between background noises, alarms or adults and cries. An example of

misclassified event for each class is reported in Figure 7.14.

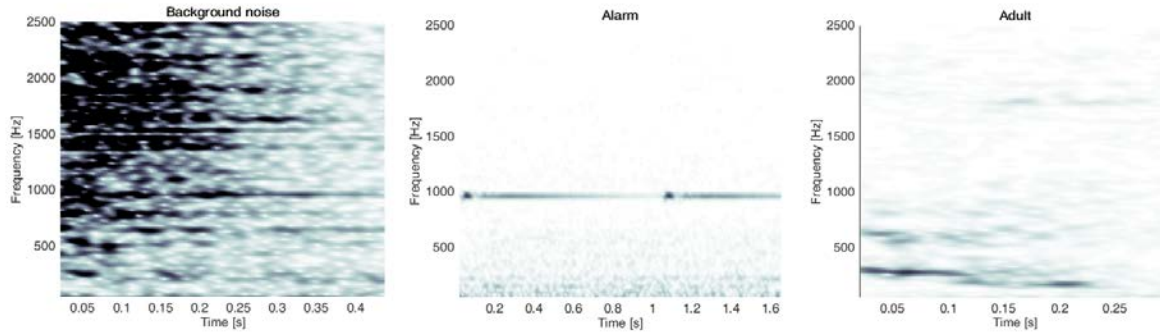


Figure 7.14 – Three examples of misclassified events.

This time, no estimation of F_0 was performed since no band of estimation would be relevant. Estimation of the fundamental frequency will be impacted by these kinds of errors if they are too many of them in regard to the number of actual cries. Nevertheless, the resulting spectrum are different from cries and may be discarded by studying upper band of frequencies (>750 Hz) of the signal. To discard misclassified segments containing adult voice, it may also be interesting to look at the decisions made on previous segments to detect if this segment has a chance to be part of period of adult speech. The adult detection algorithm based on motion and introduced in Chapter 6 may also give a piece of information for that purpose.

3 Evolution of the fundamental frequency in very preterm newborns

The analysis of cries with the objective of characterizing the maturation has only been initiated. It was performed in parallel with this work during an internship work focused on the evaluation of the evolution of the fundamental frequency in very preterm newborns.

For six newborns, eight cries were manually extracted for two dates ($D1$ and $D2$). Then, fundamental frequencies were estimated and the mean F_0 value was calculated for each cry. Resulting distributions for $D1$ and $D2$ are reported in Figure 7.15.

For all newborns, the mean fundamental frequency appears to increase between $D1$ and $D2$. This was partially confirmed by statistical analyses conducted through Mann Whitney U tests. Indeed, for five newborns significant differences (p-values < 0.05) were found.

Only one study dealt with the evolution of cry of premature newborns [9]. Authors showed that the fundamental frequency of pain-induced cries of premature newborns decreases with increasing PMA, until it becomes comparable to the ones of full-term babies. Our results show that the fundamental frequency tends to increase as premature newborns grow older. Therefore, this could be considered in contradiction. However, in our case, we are dealing with spontaneous cries. In this regard, Goberman et al. suggest that the F_0 values in premature infants compared to full-term newborns could be higher

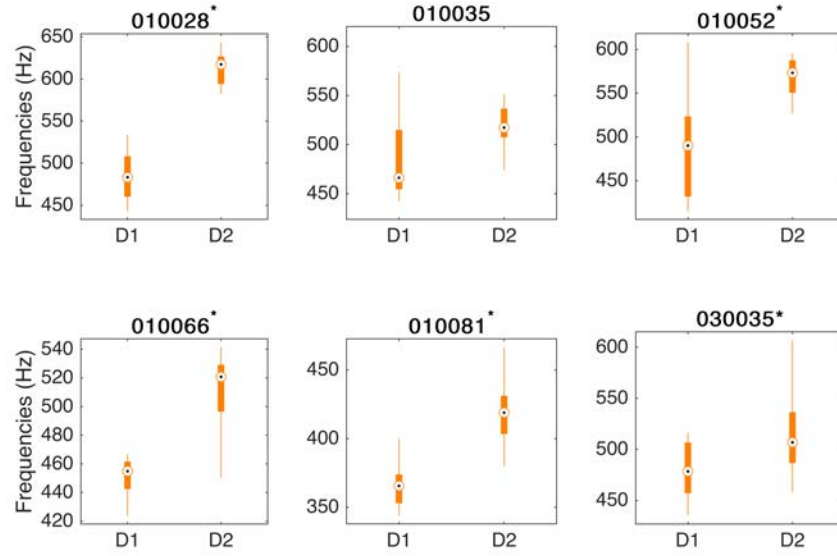


Figure 7.15 – Boxplots of the mean values of F_0 between $D1$ and $D2$ in very preterm newborns. * indicates when a p-value < 0.05 was obtained using Mann-Whitney U test.

due to the fact that premature babies are more sensitive to pain [5]. In addition, Shinya *et al.* wrote that higher frequencies observed in premature infants may be attributed to a larger post-natal age (the time spent *ex-utero*) [16]. This information was not integrated in [9] since groups of babies reaching the same PMA were composed by individuals of different GA. In the end, our results are new and not in contradiction with those of the literature. Nevertheless, given the small number of babies studied, no definitive conclusion can be drawn. Hence, more recordings will have to be processed.

4 Discussion and conclusion

In this chapter, a process was proposed to automatically extract cry events from long audio recordings of preterm newborns. First, a segmentation of the periods of interest was performed and then, a classification approach for cry selection has been proposed and evaluated. This approach was a multi-faceted problem since a relevant database had first to be manually collected and annotated, the right ensemble of features had to be defined and classifiers had to be correctly trained.

As a first step, a high number of segments (637 segments during the training and 3941 segments during the deployment) has been manually annotated. It is very important in the establishment of a database and constitutes a real contribution for further method comparisons. In fact, this dictionary contains a wide variety of sound events coming from real conditions of NICU in six hospitals.

Secondly, for the first time, Harmonic Plus Noise analysis, commonly used in speech synthesis, has been applied within the objective of cry selection. Features obtained from this analysis have proven their effectiveness since high classification performances were obtained. Indeed, performances of the KNN

classifier during the testing reached 92.9% of precision and 90.5% of recall, with an accuracy of 94.2%. This is better than what we found in the literature where accuracies of 89.2% was reported with HMM [10] and 86.6% with CNN [4].

Thirdly, an automatic classification of sounds, based on the KNN model, has been performed on real life data. For the first time, more than 180 hours of recordings were analyzed and we showed that the classifier performed well. Indeed, we saw that errors were mainly made between cries and overlapped cries or between cries and vocalizations. By integrating vocalizations and all cries, a median of 94% of good classifications was retrieved and results were high for all recordings except for two of them. Furthermore, we saw that these errors may have a limited impact for the assessment of the fundamental frequency for cry characterization.

As regard to clinical aspect, our results showed that the fundamental frequency increases with evolving age in very preterm newborns. These results were never shown before and thus, it is an open research question that will have to be investigated.

To date, it is now possible to automatically extract a large number of cries. This way, we are in position to conduct a robust maturation study on a large database.

This will help to resolve several limitations in cry selection that raised up during this work. Indeed, an 8% error rate was reported since background noises and alarms were misclassified. To reinforce the model, the annotated database could be updated by adding these segments. Then, it may be relevant to perform the HNM analysis simultaneously in a upper band (> 750 Hz) to enhance the feature set. Indeed, some of the alarm segments that were misclassified have a frequency above 750 Hz and thus, it could be characterized using HNM. It could be also relevant to detect background noises that occupy a broad band of frequencies as well as cry events with overlapping sounds. Additionally, irrelevant segments composed by multiple successive sound events (including at least a cry) can be extracted. To overcome this situation, it was shown that performing the cry extraction process another time may be relevant. Nevertheless, the segment length to trigger this new computation will have to be studied.

Another limitation was observed during the deployment of the model. In fact, our approach is not designed for crying of multiple babies. This is a tough question to deal with but the use of stereo may be relevant to define the positioning of the crying baby and thus, considering only the baby in between the two microphones. However, since the distance between the microphones and the baby may change due to room configurations, it could be interesting to propose, as for motion, an adaptive model.

To finish, during the annotation of the database, an important question was raised up regarding the choice to consider a cry as an informative one for behavioral development assessment. Studies, reported in Chapter 2, have not clarified this point and thus, it will be necessary to clearly define the scope of what we consider as a relevant cry for further analyses. For that purpose, we may rely on the work of Golub *et al.* which proposed a physioacoustic model of the infant cry [6].

Bibliography

- [1] ABOU-ABBAS, L., ALAIE, H. F., AND TADJ, C. Automatic detection of the expiratory and inspiratory phases in newborn cry signals. *Biomedical Signal Processing and Control* 19 (2015), 35–43.

-
- [2] ABOU-ABBAS, L., TADJ, C., GARGOUR, C., AND MONTAZERI, L. Expiratory and inspiratory cries detection using different signals' decomposition techniques. *Journal of Voice* 31, 2 (2017), 259.e13 – 259.e28.
- [3] BAKEN, R. J., AND ORLIKOFF, R. F. *Clinical measurement of speech and voice*. Cengage Learning, 2000.
- [4] FERRETTI, D., SEVERINI, M., PRINCIPI, E., CENCI, A., AND SQUARTINI, S. Infant cry detection in adverse acoustic environments by using deep neural networks. In *2018 26th European Signal Processing Conference (EUSIPCO)* (2018), pp. 992–996.
- [5] GOBERMAN, A. M., AND ROBB, M. P. Acoustic examination of preterm and full-term infant cries: The long-time average spectrum. *Journal of Speech, Language, and Hearing Research* 42, 4 (1999), 850–61.
- [6] GOLUB, H. L., AND CORWIN, M. J. A physioacoustic model of the infant cry. In *Infant crying*. Springer, 1985, pp. 59–82.
- [7] MANFREDI, C., BANDINI, A., MELINO, D., VIELLEVOYE, R., KALENGA, M., AND ORLANDI, S. Automated detection and classification of basic shapes of newborn cry melody. *Biomedical Signal Processing and Control* 45 (2018), 174–181.
- [8] MANFREDI, C., BOCCHI, L., ORLANDI, S., SPACCATERRA, L., AND DONZELLI, G. P. High-resolution cry analysis in preterm newborn infants. *Medical Engineering & Physics* 31, 5 (2009), 528–32.
- [9] MICHELSSON, K., JÄRVENPÄÄ, A., AND RINNE, A. Sound spectrographic analysis of pain cry in preterm infants. *Early Human Development* 8, 2 (1983), 141–149.
- [10] NAITHANI, G., KIVINUMMI, J., VIRTANEN, T., TAMMELA, O., PELTOLA, M. J., AND LEPPÄNEN, J. M. Automatic segmentation of infant cry signals using hidden Markov models. *EURASIP Journal on Audio, Speech, and Music Processing* 2018, 1 (2018), 1–14.
- [11] ORLANDI, S., GARCIA, C. A. R., BANDINI, A., DONZELLI, G., AND MANFREDI, C. Application of pattern recognition techniques to the classification of full-term and preterm infant cry. *Journal of Voice* 30, 6 (2016), 656–663.
- [12] ORLANDI, S., GUZZETTA, A., BANDINI, A., BELMONTI, V., BARBAGALLO, S. D., TEALDI, G., MAZZOTTI, S., SCATTONI, M. L., AND MANFREDI, C. AVIM—a contactless system for infant data acquisition and analysis: Software architecture and first results. *Biomedical Signal Processing and Control* 20 (2015), 85–99.
- [13] ORLANDI, S., MANFREDI, C., BOCCHI, L., AND SCATTONI, M. Automatic newborn cry analysis: A non-invasive tool to help autism early diagnosis. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (2012), IEEE, pp. 2953–2956.

- [14] OTSU, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9, 1 (1979), 62–66.
- [15] RABOSHCHUK, G., NADEU, C., JANČOVIČ, P., LILJA, A. P., KÖKÜER, M., MAHAMUD, B. M., AND DE VECIANA, A. R. A knowledge-based approach to automatic detection of equipment alarm sounds in a neonatal intensive care unit environment. *IEEE journal of Translational Engineering in Health and Medicine* 6 (2018), 1–10.
- [16] SHINYA, Y., KAWAI, M., NIWA, F., AND MYOWA-YAMAKOSHI, M. Preterm birth is associated with an increased fundamental frequency of spontaneous crying in human infants at term-equivalent age. *Biology Letters* 10, 8 (2014).
- [17] STYLIANOU, Y. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. *Ph. D thesis, Ecole Nationale Supérieure des Telecommunications* (1996).
- [18] TUKEY, J. Exploratory data analysis. *Reading, Addison-Wesley Publishing* (1977).
- [19] VÁRALLYAY, G. Future prospects of the application of the infant cry in the medicine. *Periodica Polytechnica Electrical Engineering* 50, 1-2 (2006), 47–62.
- [20] VÁRALLYAY, G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology* 71, 11 (2007), 1699–1708.

CONCLUSION AND PERSPECTIVES

The objective of this thesis was to propose a new non-invasive solution for the monitoring of the behavioral development of premature newborns. Within this purpose, the attention of this thesis has been focused on video and audio processing.

First, the relevance of such analyses in the context of pediatric health was studied. Hence, more than 150 papers were reviewed and discussed regarding their capacity to be applied for continuous monitoring. This work, which has been published [5], has shown the clinical interest of the proposed approach and the need to develop an acquisition and processing workflow adapted to the clinical routine.

Simultaneously, a preliminary study about sleep states estimation in preterm newborn was conducted. For the first time, a semi-automatic approach combining video and audio processing was developed and published [4]. From there, several decisions were taken regarding the orientation and the priority developments to move towards a fully automatic monitoring solution.

As a first step, a new audio-video acquisition system was proposed. Integration in Neonatal Intensive Care Units was studied in order to fit a wide variety of room configurations, keeping in mind the non-invasive objective regarding both newborns and healthcare personnel. In parallel with this thesis, a focus group studying the ethical aspect of this new monitoring approach was conducted [6]. In brief, caregivers and parents perceived audio and video in care as useful and acceptable provided that measures are taken to ensure informed consent, data protection and to limit the negative impact for caregivers. These points are important and should be considered for the acceptance of a new monitoring system integrating audio and video analyses in NICU.

Secondly, the video data analysis was focused on motion analyses. A process going from video acquisition to the extraction of relevant features for neuro-behavioral characterization has been proposed. It is based on the extraction of motion series by the mean of classical video techniques. The strength of this work relies on the fact that several difficulties, inherent to long time monitoring, have been tackled. The most significant one is the automatic detection of irrelevant periods of analyses due to the presence of adult in the field of view. This specific part of the process was presented [2] and published [3]. Additionally, the developed process is robust to a wide variety of recording conditions whether for very covered or equipped newborns, different beds and changing light conditions.

Additionally, a new set of features characterizing motion organization in terms of duration and number of motion and non-motion epochs was retrieved by the mean of classification techniques. Results reveal that the proposed set of features (especially the number of non-motion epochs) is relevant to evaluate the evolution of the motion organization of preterm newborns.

Finally, a process going from audio recording to the automatic extraction of cry events was proposed. As for motion, the strength of this works resides in the fact that it was designed in order to fit the real conditions of NICU where diverse irrelevant sounds can occur such as adult voices, device alarms or background noises. Our approach for cry extraction is based on a set of features computed from

harmonic plus noise analysis. This technique, usually applied in speech synthesis, was never tackled within this purpose and results reveal to be better than the ones observed in the literature.

All these results augur well for the development of non-invasive systems for the monitoring of behavioral development in preterm newborns. However, several difficulties will have to be overcome to move towards a fully automatic solution and constitute some relevant technical perspectives.

First, it has been shown that the processing of long time recordings is achievable either for motion analysis or for cry assessment as well as by combining audio and video analyses to characterize sleep. Nevertheless, solutions will have to be developed to deal with the remaining limitations of continuous monitoring. The most important point, regarding the motion analysis, is the automatic detection of the presence of the baby. Deep learning based approaches are promising for this purpose. They have not been considered in this work since the database was in construction. Other approaches, like the exploitation of the temperature, although complicated by the temperature regulation, or the use of intelligent mattress may also be investigated. At the beginning of the project, this equipment was discarded since it was expected that cameras could fully stick to observational techniques and thus, captured all the information collected by nurses such as facial activities, movements and temperature. However, we saw that monitoring of facial activities in real conditions of NICU is a complex task since the face of the baby is often partially or completely hidden. About audio analyses, an automatic search of the relevant band of frequencies to perform the fundamental frequency estimation will have to be proposed. This will enhance the process efficiency in terms of cry extraction and characterization. For that purpose, the spectrogram can be studied either by classical signal techniques (e.g., Fourier Transform and peak-picking procedure) or by the mean of image processing (e.g., spectrogram and contour detections). In addition, other sounds produced by the newborn (e.g., coughing, vowel sounds) may be studied and be markers of neuro-behavioral development [7]. This can be performed using the same classification approach based on signal processing provided that annotations and feature extraction had been adapted to this purpose. Otherwise, classification of spectrogram images may also be considered. To date, the audio part of this work has been taken over in the context of a new PhD thesis.

As regard to clinical perspectives, for the first time, the integration of an audio and video system in NICU was investigated. Although it was globally successful, some drawbacks have been raised up. Indeed, the system, in its current design, is cumbersome for acquisition in closed beds. In addition, for now, the clinicians were unable to assess the interest of such system in their daily care since no processing was embedded. To answer these points, a new project, called NEOVIDEO, has been proposed and accepted (Programme Hospitalier de Recherche Infirmière et Paramédicale - PHRIP). Its objective is to measure the impact of monitoring of the video monitoring on the sleep of preterm newborns. For this purpose, a system integrating motion analysis will be studied and deployed in NICU. Video streams and motion series will be sent back constantly to the central monitoring room so that nurses have access to it. From there, individualized care routine can be provided so as not to disturb the newborn sleep.

Simultaneously to these works, the huge amount of data that is collected during Digi-NewB will have to be processed in order to validate the relevance of the features that were extracted in this thesis. Indeed, until there, only a small part of the recordings has been processed, notably because this first part of the project was dedicated to the development of the methods. In particular, only two dates of the care were investigated (near birth and near discharge), although the analysis of the recordings

performed in between will gives us a more precise evaluation of preterm newborn development. Once the best descriptors will be selected, it will be possible to improve our control group in order to follow the evolution over the time of the premature newborns. Then, this set of indicators will have to be integrated in a monitoring system in order to help clinicians in their decision-making.

In addition, the sepsis side of the Digi-NewB project was not addressed in this thesis. We particularly think that motion analyses can also give relevant information about this episode since it is associated with an atonic behavior of preterm newborns [1]. The obstacle on this study will be lifted once the processes would have been fully automated. However, new algorithms may have to be developed. For example, periods of photo-therapy will have to be excluded from the analyses. It may be automatically discarded either by studying the sudden variations in pixel intensities in the entire frame or by the use of the color camera to detect the blue lightening induced by this technique.

As a conclusion, one thing is for sure, there is always room for improvements and this work will serve as a basis for future developments to move towards fully automatic solutions for a new generation of monitoring systems in preterm newborns.

Bibliography

- [1] BEKHOF, J., REITSMA, J. B., KOK, J. H., AND VAN STRAATEN, I. H. Clinical signs to identify late-onset sepsis in preterm infants. *European journal of pediatrics* 172, 4 (2013), 501–508.
- [2] CABON, S., PORÉE, F., SIMON, A., , UGOLIN, M., ROSEC, O., CARRAULT, G., AND PLADYS, P. Caractérisation du mouvement chez les nouveau-nés prématurés par analyse automatique de vidéos. In *Recherche en Imagerie et Technologies pour la Santé (RITS)* (2017).
- [3] CABON, S., PORÉE, F., SIMON, A., , UGOLIN, M., ROSEC, O., CARRAULT, G., AND PLADYS, P. Motion estimation and characterization in premature newborns using long duration video recordings. *IRBM* 38, 4 (2017), 207–213.
- [4] CABON, S., PORÉE, F., SIMON, A., MET-MONTOT, B., PLADYS, P., ROSEC, O., NARDI, N., AND CARRAULT, G. Audio- and video-based estimation of the sleep stages of newborns in neonatal intensive care unit. *Biomedical Signal Processing and Control* (2019).
- [5] CABON, S., PORÉE, F., SIMON, A., ROSEC, O., PLADYS, P., AND CARRAULT, G. Video and audio processing in paediatrics: a review. *Physiological Measurement* (2019).
- [6] MAZILLE, N., LEBRIS, A., SIMONOT, P., LUHERNE, M., GASCOIN, G., HARTE, R., AND PLADYS, P. Parents' and caregivers' perceptions of the use of live video recording in neonatal units, a qualitative study. In *American Paediatrics Society* (2019).
- [7] THACH, B. T. Maturation of cough and other reflexes that protect the fetal and neonatal airway. *Pulmonary Pharmacology & Therapeutics* 20, 4 (2007), 365–370.

LIST OF PUBLICATIONS

International journal with peer-review processes

[A1] Cabon, S., Porée, F., Simon, A., Met-Montot B., Pladys, P., Rosec, O., Nardi, N. and Carrault, G. Audio- and Video-based estimation of the sleep stages of newborns in Neonatal Intensive Care Unit. *in Biomedical Signal Processing and Control*, 52, 362-370. (2019).

[A2] Cabon, S., Porée, F., Simon, A., Rosec, O., Pladys, P., Carrault, G. Video and audio processing in paediatrics: a review. *Physiological Measurement*, 40(2), 1-20. (2019)

[A3] Cabon, S., Porée, F., Simon, A., Ugolin, M., Rosec, O., Carrault, G. and Pladys, P. Motion Estimation and Characterization in Premature Newborns Using Long Duration Video Recordings. *Innovation and Research in BioMedical engineering*, 38(4), 207-213. (2017).

National conferences

[C1] Cabon S., Porée F., Simon A., Ugolin M., Rosec O., Carrault G. et Pladys P. Caractérisation du mouvement chez les nouveau-nés prématurés par analyse automatique de vidéos. *Journées RITS*. Lyon, France, résumé. (2017).

[C2] Porée F., Simon A., Cabon S., Corolleur A., Nardi N., Pladys P., Carrault G. Traitement de vidéos de polysomnographie pour l'estimation de l'état des yeux chez le nouveau-né prématuré. *In XXVe Colloque GRETSI*, pp. 1–4. (2015).

Software

[D1] Weber R., Cabon S., Porée F., Simon A. et Carrault G. ViSiAnnoT: Video and Signal Annotation Tool. Dépôt APP, déposé le 2 avril 2019.

[D2] Porée F., Simon A., Carrault G., Pladys P., Cabon S., Péron F., Corolleur A., Zhang K. CAPTIV: Caractérisation Automatique du comportement en Pédiatrie par Traitements Informatisés de Vidéos. Dépôt APP no IDDN.FR.001.050018.000.S.P.2016.000.31230. (2016).

Appendices

USER MANUAL



MANUEL D'INSTALLATION ET D'UTILISATION DU SYSTEME D'ACQUISITION DES DONNEES DIGI-NEWB



*Ce projet est financé par le programme de Recherche et
d'Innovation Horizon 2020 (GA n°689260)*

Lead Beneficiary	2-Voxygen Health		
Responsible author(s)	Sandie Cabon, Guillaume Cuffel	Email	sandie.cabon@voxygen.fr guillaume.cuffel@voxygen.fr
	Beneficiary	Voxygen Health	Phone +(33)296141281
Contributing authors	<p>Voxygen Health : Olivier Rosec, Sebastien Vermandel</p> <p>UR1-INSERM1099: Guy Carrault, Antoine Simon, Fabienne Porée, Fabrice Tudoret, Thomas Janvier</p> <p>CHU Rennes/GCS HUGO: Maude Luherne, Patrick Pladys, Florence Geslin, Philippe Cozic, Pierre-Yves Donnio, Frédérique Rocaboy</p>		

Table des matières

1. LISTE DU MATERIEL	3
2. PREPARATION DE L'ENREGISTREMENT	4
1.1. Inclusion du patient.....	4
1.2. Nettoyage du matériel.....	4
2. INSTALLATION DU DISPOSITIF D'ENREGISTREMENT	5
2.1. Système d'enregistrement des signaux électrophysiologiques	5
2.2. Système d'enregistrement vidéo.....	6
2.2.1. Configuration « Couveuse fermée ».....	6
2.2.2. Configuration « Lit ouvert avec barres ».....	8
2.2.3. Configuration « Lit ouvert avec pied à perfusion »	9
2.3. Mise en route.....	10
3. CONFIGURATION DE L'ENREGISTREMENT.....	11
3.1. Branchement du PC de configuration	11
3.2. Accès à l'interface utilisateur	11
3.1. Vérification de l'installation.....	12
3.1.1. Vérification du dispositif d'acquisition.....	12
3.1.2. Vérification de l'orientation des caméras.....	13
3.2. Gestion de l'enregistrement.....	15
3.2.1. Démarrage de l'enregistrement.....	15
3.2.2. Vérification de l'enregistrement.....	16
4.1.1. Arrêt de l'enregistrement.....	16
5. STOCKAGE DES DONNEES.....	17
4.2. Transfert des enregistrements	17
4.2.1. Mise en route du NAS.....	17
4.2.2. Procédure de stockage des données.....	17
4.2.3. Procédure de collecte des données	18
5. MISE A JOUR DU LOGICIEL.....	19
6. EN CAS DE DIFFICULTES.....	19

1. Liste du matériel

Capteur multimodal :



1. Boîtier principal (+rotule)
2. Boîtier secondaire (+rotule)

3. 2x câbles de connexion inter-boîtier gris
4. Câble d'alimentation
5. Câble de connexion

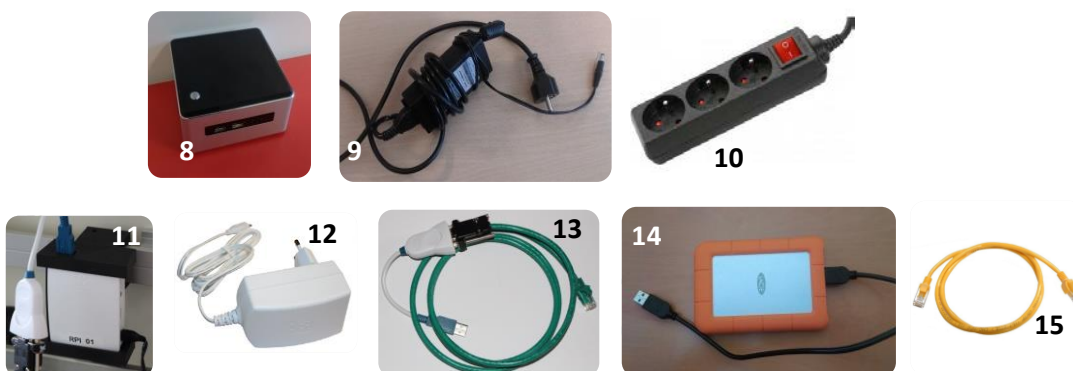
Système de fixation :



6. Support pour couveuse

7. 2x pinces pour lit ouvert

Système d'enregistrement :



8. Mini PC « NUC »
9. Câble d'alimentation du NUC
10. Multiprise électrique
11. Raspberry Pi ou « RPi »

12. Câble d'alimentation du RPi
13. Câble de connexion au moniteur Philips (Vert pour les MP5, bleu pour les autres)
14. Disque dur externe
15. Câble de synchronisation

Unité de configuration (*) :



16. Câbles de connexion noirs
17. PC de configuration (+alimentation)

18. Adaptateur USB/RJ45

- Support de stockage (*) :



19. NAS 5 disques (+alimentation)
20. Câble de connexion noir

(*) Une seule unité par centre

2. Préparation de l'enregistrement

1.1. Inclusion du patient

Chaque nouveau patient doit être inclus dans l'eCRF dont voici la procédure d'inclusion :

1. Se connecter à <https://chu-rennes.hugo-online.fr/CSOnline>
2. Rentrer les codes de connexion fournis par le support clinique
3. Cliquer sur « Tableau de suivi des patients »
4. Ajouter un patient



Le système propose un code patient automatique, cliquer sur « suivant »

OU

Cliquer sur l'icône de l'eCRF de la ligne d'un patient de la liste pour accéder aux pages

Investigateur	Statut	Patient ▲ 2	CRF ▲ 3	eCRF
DUPONT JEAN		990001	1-Digi-NewB	
DUPONT JEAN		990002	1-Digi-NewB	

© 2016 Ennov Clinical

1.2. Nettoyage du matériel

Produit à utiliser :

Lingettes Bactynéa® (Laboratoire Garcin-Bactinyl)

Procédure d'entretien :

Le capteur multimodal complet et son système de fixation (cf. « Liste du matériel ») devra être nettoyé avant chaque session d'enregistrement. Il devra également être retiré du lit de l'enfant afin de permettre un entretien complet à chaque nettoyage de la couveuse.

⚠ Penser à éteindre le capteur en appuyant sur le bouton du boîtier principal.

Le dispositif vidéo devra être repositionné après l'entretien quotidien dans le lit de l'enfant au même emplacement que lors de la mise en place du système.

⚠ Ne pas oublier de rallumer le boîtier si vous l'avez éteint.

Le reste du matériel d'enregistrement vidéo et signal seront intégrés au bionettoyage de la chambre.

Protocole d'utilisation des lingettes Bactynéa® :

1. Essuyage humide des surfaces avec la lingette en portant une attention particulière au niveau des aspérités du système (pas de vis, rainures...).
 2. Séchage passif
 3. Rinçage avec une chiffonnette à usage unique, propre, imprégnée (bien essorée) d'eau du robinet
- ⚠ Bien refermer le couvercle des lingettes après utilisation pour que le taux d'imprégnation soit optimum jusqu'à la dernière lingette.

2. Installation du dispositif d'enregistrement

2.1. Système d'enregistrement des signaux électrophysiologiques

L'installation du matériel d'enregistrement des signaux électrophysiologiques est la même quel que soit le type de chambre. Cette partie est indépendante des choix d'installation de la partie vidéo.

Matériel :

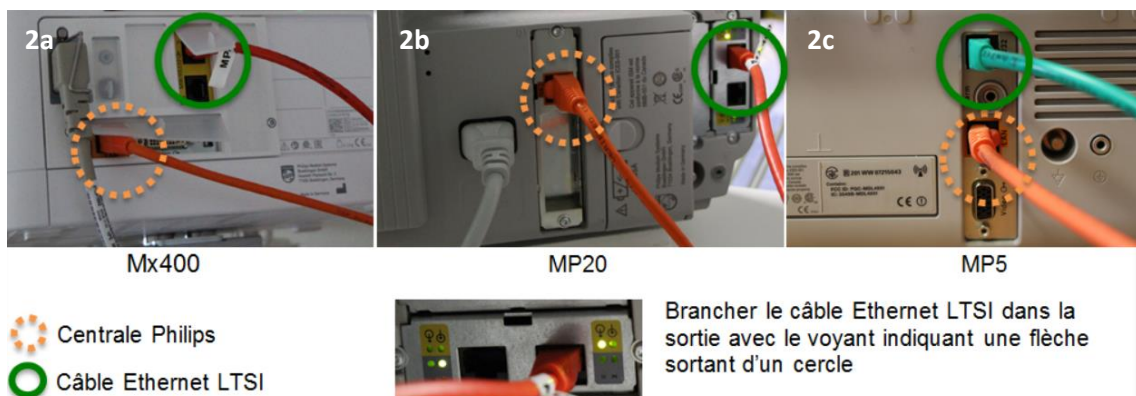
- La RPi (#11) et son alimentation (#12)
- Le câble de connexion au moniteur Philips (#13)

Etapes d'installation :

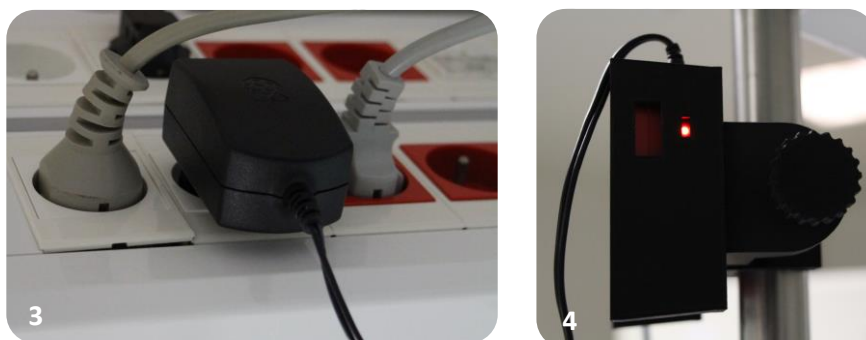
1. Fixer en vissant délicatement la RPi à l'aide de son système d'attache.



2. Brancher le câble Ethernet au moniteur Philips en fonction du modèle.

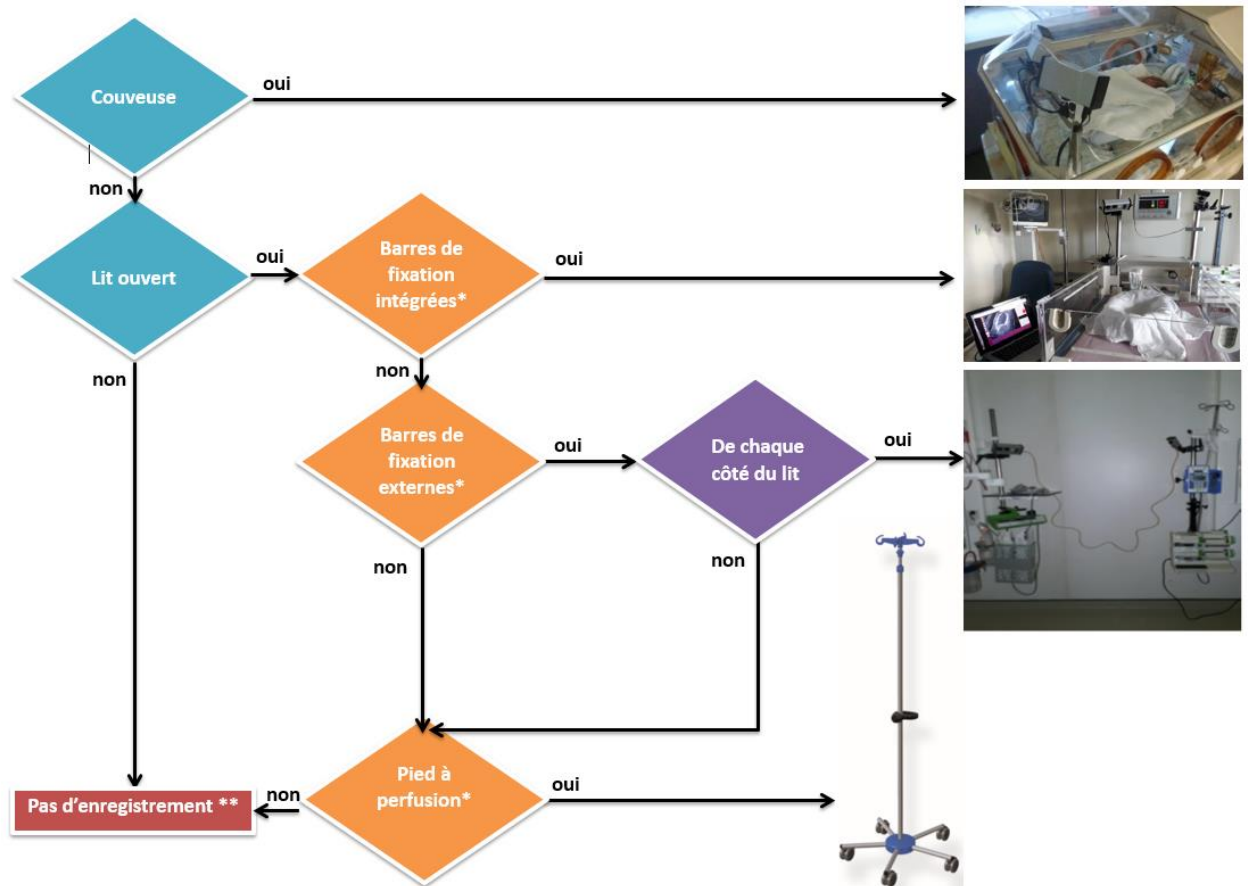


3. Brancher l'alimentation du RPi au secteur (220V), de préférence sur une prise rouge.
4. Le voyant rouge/vert indique que la RPi est correctement démarrée.



2.2. Système d'enregistrement vidéo

L'installation du système d'enregistrement vidéo doit s'adapter à l'environnement de la chambre en fonction des différents types de lit et systèmes de fixation disponibles. On distingue 4 types de configuration ; « couveuse fermée », « lit ouvert avec barres » et « lit ouvert avec pied à perfusion ». Le choix de la configuration doit être effectué en suivant l'arbre décisionnel ci-dessous :



*Matériel obligatoirement validé par le support technique

**Contacter le support technique

2.2.1. Configuration « Couveuse fermée »

Définition :

Couveuse munie d'un habitacle clos

Matériel :

- 1 capteur multimodal complet avec le câble de connexion (#3) le plus court
- 1 support pour couveuse fermée (#6)

Etapes d'installation :

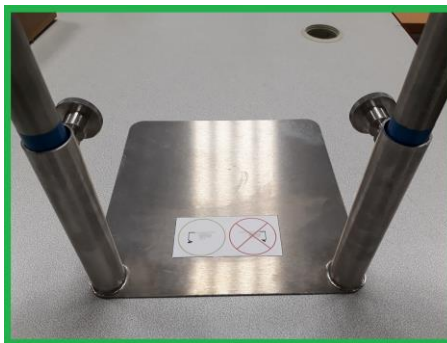
1. Visser les 2 boîtiers sur le support sans connectique pour le moment. Si l'on se place aux pieds du bébé face à la couveuse, le boîtier principal (#1) doit être situé à droite et le boîtier secondaire (#2) à gauche.
2. Préparer le système pour l'installer dans la couveuse :
 - a. Pencher les caméras

- b. Régler au plus bas la hauteur du support
- c. Brancher uniquement le câble de connexion (#3)
- d. Ouvrir la couveuse et faire glisser le support sous le matelas



3. Ajuster le positionnement du capteur:
- a. Redresser les deux boîtiers en les orientant vers le bébé
 - b. Régler au plus haut la hauteur du support sans que les boîtiers ne touchent la paroi

⚠ Bien faire attention de ne pas dépasser la limite rouge



4. Brancher le reste de la connectique comme décrit dans le paragraphe suivant en prenant soin de ne pas faire traîner de fils dans le passage.



2.2.2. Configuration « Lit ouvert avec barres »

Définition :

Lit à ciel ouvert qui permet l'installation de « pinces de fixation » grâce à la présence de l'un des équipements conformes* suivants :

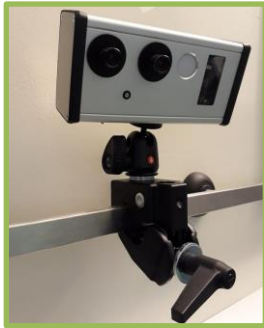
- 2 mâts intégrés au lit
- 2 mâts muraux de chaque côté du lit
- 1 rail accessible de chaque côté du lit

Matériel :

- 1 capteur multimodal complet avec le câble de connexion (#3) le plus long
- 2 « pinces de fixation » (#7) pour lit ouvert

Etapes d'installation :

1. Visser chaque boîtier sur une « pince de fixation »
2. Fixer chaque pince sur une barre de chaque côté du lit



⚠ Cas particulier : Si un moniteur ou autre élément de la table gêne la mise en place du boîtier principal sur la gauche inverser les 2 boîtiers.

3. Accrocher l'élingue de sécurité autour de la barre



4. Relier les 2 boîtiers avec le câble de connexion (#3)
5. Brancher le reste de la connectique comme décrit dans le paragraphe suivant en prenant soin de gêner le moins possible le travail du personnel médical



⚠ Ne pas laisser la vis dans l'axe de la caméra, placer-la de préférence sur le côté

2.2.3. Configuration « Lit ouvert avec pied à perfusion »

Définition :

Lit à ciel ouvert qui **NE** permet **PAS** l'installation des « pinces de fixation ».

Matériel :

- 1 capteur multimodal complet avec le câble de connexion (#3) le plus long
- 2 « pinces de fixation » (#7) pour lit ouvert
- 1 pied à perfusion estampillé « Digi-NewB »

Etapes d'installation :

1. Visser chaque boîtier sur une « pince de fixation »



2. Fixer chaque pince sur le pied à perfusion
3. Accrocher l'élingue de sécurité autour du mât
4. Relier les 2 boîtiers avec le câble de connexion (#3) le plus court
5. Positionner le pied à perfusion aux pieds du bébé, face à la couveuse. La tablette permet de positionner le pied à la perfusion à la bonne distance du lit.
6. Enclencher les freins des roulettes
7. Brancher le reste de la connectique comme décrit dans le paragraphe suivant en prenant soin de positionner le reste du matériel sur la tablette



2.3. Mise en route

Suivre ces étapes pour la mise en route du dispositif d'enregistrement :

1. Brancher la multiprise (#10)
2. Placer le NUC (#8) sur un emplacement ouvert et disponible autour du lit. Faire en sorte que son positionnement soit le moins gênant possible pour le personnel soignant et la famille.
3. Connecter l'alimentation du NUC (#9) à la multiprise (*marquage blanc*)
4. Brancher le disque dur externe (#14) sur l'une des prises USB à l'arrière du NUC (*marquage vert et jaune*)



5. Connecter le câble USB (#5) entre le boîtier principal et le NUC (*marquage bleu*)



6. Connecter l'alimentation du boîtier (#4) à la multiprise (*marquage rouge*)
7. Connecter le câble de connexion **gris** (#3) d'un boîtier à l'autre
8. Connecter le câble de synchronisation **jaune** (#15) entre le NUC et la RPi



9. Allumer la multiprise, le boîtier principal et le NUC.

3. Configuration de l'enregistrement

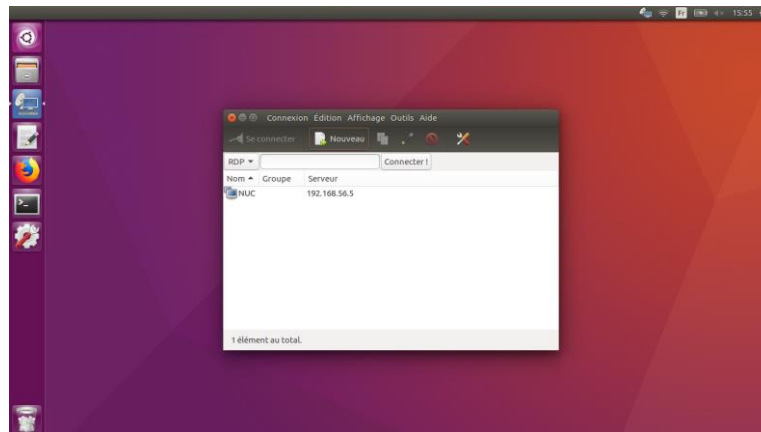
3.1. Branchement du PC de configuration

1. Ajouter l'adaptateur (#18) au câble de connexion **noir** (#16)
2. Brancher le tout entre le NUC sur un des ports USB à l'avant et le PC sur le port Ethernet
3. Brancher et allumer le PC de configuration (#17).

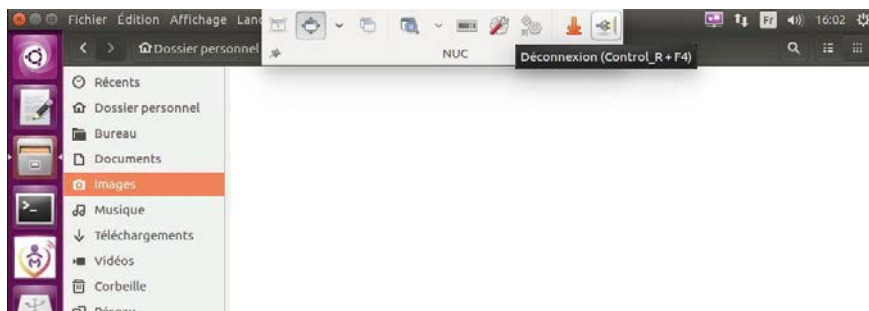


3.2. Accès à l'interface utilisateur

Lors de l'allumage du PC de configuration, une fenêtre intitulée « Visionneur de bureaux distants Remmina » s'ouvre automatiquement. Pour accéder à l'interface utilisateur, il suffit de double-cliquer sur « NUC ».



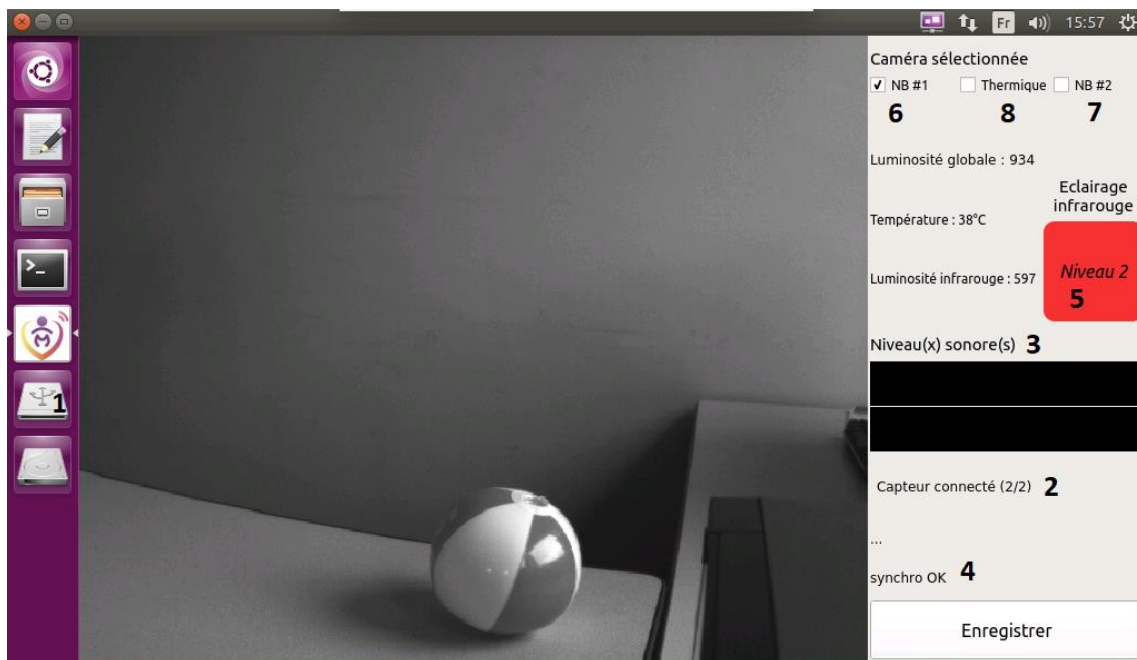
Pour se déconnecter de l'interface, cliquer sur le bouton « Déconnexion » tout en haut de l'écran.



3.1. Vérification de l'installation

3.1.1. Vérification du dispositif d'acquisition

Lorsque le capteur multimodal est correctement détecté par le NUC, l'interface graphique ci-dessous s'affiche. Il faut alors s'assurer que les différents points listés ci-dessous soit respecté pour que l'enregistrement fonctionne correctement.



1. Vérifier que le disque externe est correctement connecté : le logo doit être présent dans le lanceur. Sinon, le débrancher puis le rebrancher.
2. Vérifier que les deux boîtiers sont bien détectés : le message « Capteurs connectés 2/2 » doit s'afficher. Sinon :
 - a. Vérifier que le boîtier principal est bien allumé
 - b. Vérifier que le câble de connexion (#5) est bien branché entre les 2 boîtiers
 - c. Eteindre puis relancer l'application en cliquant sur le logo « Digi-NewB » situé dans la barre de gauche
3. Vérifier le son : une barre rouge doit être visible sur le fond noir pour chaque microphone lorsqu'il y a du bruit. Pour rester silencieux à côté du bébé, tapoter simplement chacun des boîtiers. Si aucune barre rouge n'apparaît, éteindre puis rallumer l'application et le capteur multimodal.
4. Vérifier que la synchronisation est bien faite. Le message « Synchro OK » doit s'afficher. Sinon, attendre et redémarrer l'application ainsi que la RPi.
5. Vérifier que l'éclairage infra-rouge est bien au « Niveau 2 ».
6. Vérifier l'affichage et l'orientation de la caméra noir et blanc du boîtier principal en cliquant sur le bouton « NB#1 » en suivant les recommandations décrites dans le paragraphe 0.
7. Vérifier l'affichage et l'orientation de la caméra noir et blanc du boîtier secondaire en cliquant sur le bouton « NB#2 » en suivant les recommandations décrites dans le paragraphe 0.
8. Vérifier l'affichage de la caméra thermique en cliquant sur le bouton « Thermique »

3.1.2. Vérification de l'orientation des caméras

Recommandations générales :

1. Voir le bébé en entier
2. Centrer l'image sur le bébé
3. Laisser un cadre autour du bébé : le bébé ne doit pas pouvoir bouger dans ce cadre. Ce cadre permettra d'enlever automatiquement les passages de présence d'adultes.

Capteur installé sur barres :

Caméra trop basse sur la barre



Caméra trop orientée à gauche



Caméra trop orientée à droite



Caméra trop orientée vers le bas

 Image acceptée

 Image rejetée

 Cadre à respecter

Capteur installé au pied du lit :

Caméra trop orientée vers le haut



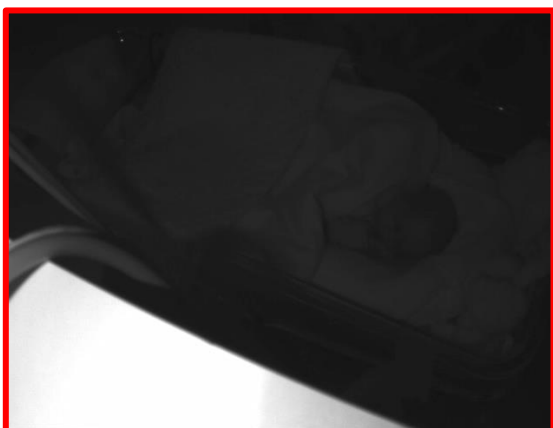
Caméra trop orientée à gauche



Caméra trop orientée à droite



⚠ Ne pas orienter la caméra trop vers le bas pour éviter la surexposition de la couverture



⚠ Faire attention de ne pas introduire un autre objet au premier plan tel qu'une tablette, une couverture ou une peluche. Cela provoquerait une surexposition du premier plan au détriment de l'image à exploiter.

3.2. Gestion de l'enregistrement

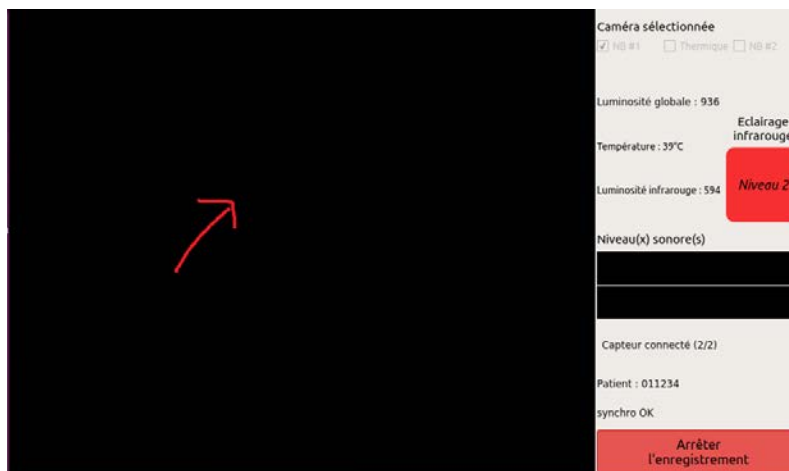
3.2.1. Démarrage de l'enregistrement

Pour démarrer une nouvelle session d'enregistrement, suivre les étapes suivantes :

1. Cliquer sur « Enregistrement »
2. Entrer le numéro d'inclusion obtenu par l'eCRF. Ce numéro doit obligatoirement être composé de 6 chiffres.



3. Après avoir entré un numéro valide, l'enregistrement est lancé. Ceci a pour conséquence de désactiver l'affichage des caméras. Il est cependant possible de le réactiver en cliquant sur l'image. Elle se désactivera automatiquement au bout d'une dizaine de minutes.

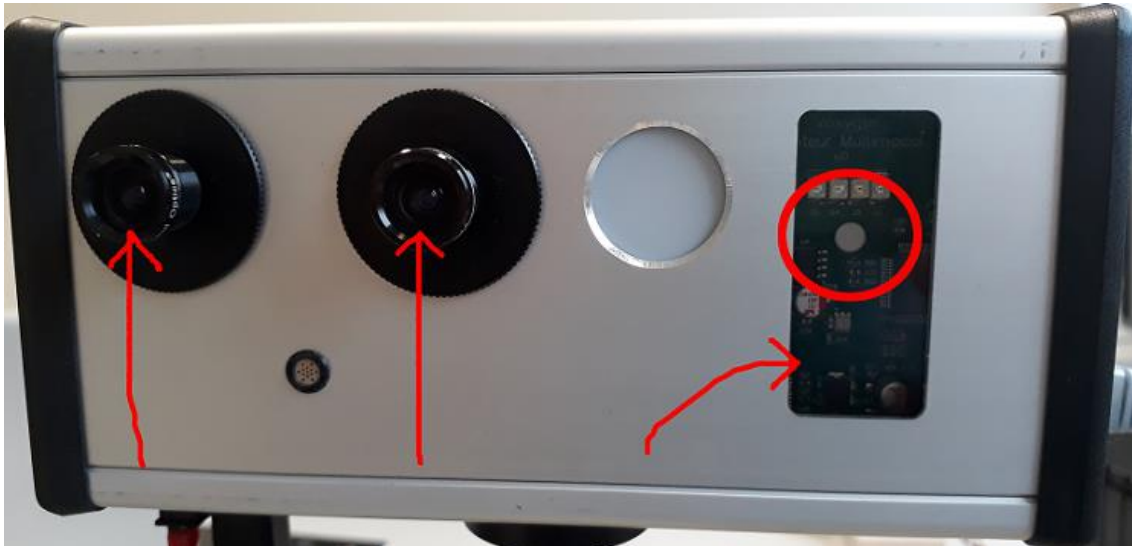


4. Laisser le matériel en place pendant toute la durée de l'enregistrement sauf le PC de configuration avec son câble de connexion et l'adaptateur. Ces 2 éléments peuvent être instantanément retirés sans risques.

3.2.2. Vérification de l'enregistrement

Lorsqu'une session d'enregistrement est en cours, il est conseillé de venir vérifier son bon fonctionnement à minima tous les jours. Pour cela, il suffit de :

1. Accéder à l'interface utilisateur en reprenant les étapes 3.1 et 3.2 décrites au début de ce paragraphe.
2. Vérifier que l'enregistrement fonctionne en cliquant sur l'écran noir
3. Vérifier que les 2 caméras NB#1 et NB#2 sont correctement orientées (cf. partie 3.1.2)
4. Dans le cas d'une couveuse avec taux d'hygrométrie augmentée, vérifier que :
 - Il n'y a pas de condensation sur les boîtiers, particulièrement sur leur vitre extérieure et leurs objectifs (flèches rouges).
 - △ Cela peut néanmoins arriver dans les 30 minutes après l'installation du matériel. Si tel est le cas, essuyez simplement la partie concernée. Si le phénomène persiste, contacter le support technique.
 - Le témoin d'humidité est de couleur blanche (cercle rouge).
 - △ S'il est de couleur rouge, arrêter IMMEDIATEMENT l'enregistrement en cours et retirer le matériel d'enregistrement de la couveuse, puis contacter le support technique.



4.1.1. Arrêt de l'enregistrement

Pour arrêter la session d'enregistrement en cours, suivre les étapes suivantes :

1. Appuyer fortement mais brièvement (environ 1s) sur le bouton d'allumage du NUC. Attendre une vingtaine de seconde avant que le NUC ne s'éteigne.
 2. Eteindre la multiprise
 3. Débrancher et récupérer tout le matériel installé pour l'enregistrement.
- △ Débrancher le NUC ou la multiprise **uniquement** lorsque la lumière bleue sur sa face avant est éteinte. En fonction de la quantité de données à copier, la durée d'extinction peut varier de 5 à 30 secondes.

5. Stockage des données

Les données enregistrées doivent être transférées de façon hebdomadaire. Dans le cas d'un enregistrement long, l'enregistrement doit être arrêté et relancé à la suite de ce transfert.

4.2. Transfert des enregistrements

4.2.1. Mise en route du NAS

Pour allumer le NAS (#19), faire un appui long sur le bouton « POWER ». Suite à cela, un signal sonore annonce le démarrage, puis une dizaine de minutes plus tard, un second signal sonore signale que le NAS est prêt à l'emploi. Il est prévu pour tourner en continu, mais il est préférable de l'éteindre s'il reste inutilisé plusieurs jours.

⚠ Ne surtout pas déplacer le NAS lorsqu'il est allumé !



Pour l'éteindre, faire un appui long sur le bouton « POWER ». Suite à cela, un signal sonore annonce l'extinction, puis quelques minutes plus tard, le NAS et les led s'éteignent.

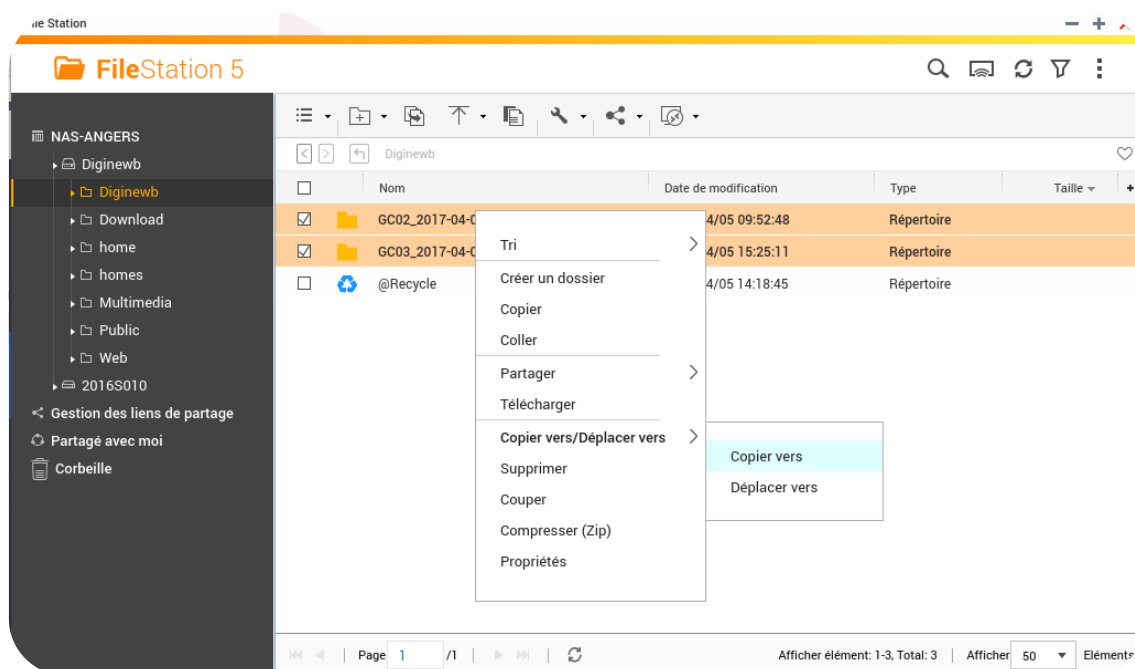
4.2.2. Procédure de stockage des données

Le disque dur externe (#14) est un outil de transfert, il permet de copier les données enregistrées vers l'espace de stockage final, à savoir le NAS. Ce support mobile est dimensionné pour contenir jusqu'à 5 semaines d'enregistrement continu.

⚠ Il est fortement recommandé de procéder au stockage des données de manière hebdomadaire !

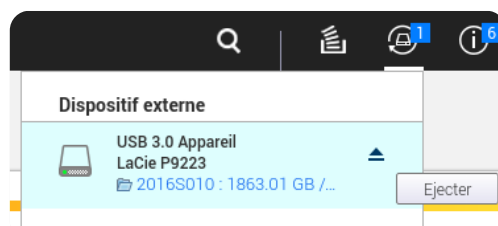


1. Brancher le disque dur externe sur l'avant du NAS
2. Connecter le PC de configuration (#17) au NAS grâce à l'un des câbles Ethernet noirs (#16)
3. Double cliquer sur l'icône 
4. Se connecter en utilisant l'identifiant « voxyvi » et le mot de passe communiqué par le support technique
5. Cliquer sur l'icône File Station 
6. Déplacer les données du disque dans le dossier Digi-NewB en faisant un clic droit sur le ou les répertoire(s) à stocker (ex : 010012_2017-09-21T09-57-52) > « Copier vers/Déplacer vers » > Déplacer vers > Diginewb



⚠ Contacter le support technique si vous vous rendez compte d'une incohérence entre le temps de l'enregistrement et la durée des fichiers audio ou vidéo enregistrés. L'enregistrement peut avoir été découpé en plusieurs sessions, vérifiez donc aussi les autres dossiers.

7. Une fois la copie terminée, vérifier qu'il n'y a plus de fichiers sur le disque externe. Sinon, vérifier si la copie s'est bien effectuée. Si c'est le cas supprimer tout le contenu du disque externe.
8. Ejecter le disque en cliquant sur l'icône du disque dur en haut à droite de la fenêtre, puis sur le bouton « Ejecter »
9. Débrancher le disque externe



4.2.3. Procédure de collecte des données

Les données brutes stockées dans le NAS doivent être remontées vers le laboratoire de recherche de manière périodique, sur une base semestrielle. Il faudra donc récupérer les données sur un disque de collecte prévu à cet effet.

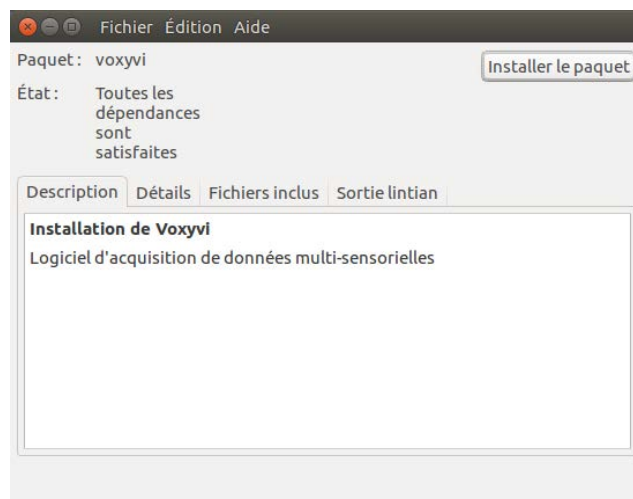
1. Brancher le câble Ethernet noir (#20) sur le PC de configuration et sur la prise Ethernet 3 du NAS.
2. Brancher le disque de collecte sur l'une des prises USB du NAS.
3. Faire un double-clic sur l'icône du NAS sur le bureau du PC de configuration.
4. Un navigateur s'ouvre et demande une identification ; utiliser les identifiants communiqué par le support technique.
5. L'interface de pilotage du NAS s'affiche. Cliquer sur l'icône « File Station » :
6. Copier les données à collecter (cf. fichier de suivi des collectes) vers le disque de collecte.
7. Renseigner le fichier de suivi des collectes
8. Envoyer le disque dur au LTSI à l'adresse :

Laboratoire Traitement du Signal et de l'Image (LTSI) Université de Rennes 1. Campus de Beaulieu. Bât 22. 35042 Cedex - Rennes - FRANCE

5. Mise à jour du logiciel

Lorsqu'une nouvelle version du logiciel d'acquisition de Voxyvi est disponible, vous serez contacté par l'équipe du support technique par mail. Pour installer la version sur chaque NUC, il faut:

1. Récupérer la dernière version du logiciel « **voxyvi-x.y-z_amd64.deb** »* auprès du support technique.
2. Copier le fichier sur un support amovible (clé USB ou disque dur externe) et le brancher au NUC.
3. Accéder au NUC avec le PC de configuration via la procédure habituelle.
4. Cliquer sur l'icône du support amovible externe pour accéder à son contenu.
5. Double-cliquer sur le fichier « **voxyvi-x.y-z_amd64.deb** »* pour faire apparaître la fenêtre ci-dessous. Si ce n'est pas cette fenêtre qui s'affiche, mais une autre intitulée « Logiciels Ubuntu », suivre ces étapes:
 - a. Fermer la fenêtre « Logiciels Ubuntu »
 - b. Faire un clic droit sur le fichier « **voxyvi-x.y-z_amd64.deb** »*
 - c. Cliquer sur « Ouvrir avec » puis sur « Installateur de paquet GDebi »



6. Cliquer sur le bouton « Installer le paquet » puis entrer le mot de passe qui vous a été communiqué.
7. Si aucune erreur n'est apparue, cela signifie que la nouvelle version du logiciel d'acquisition de Voxyvi est installée! Sinon, contacter le support technique.
8. Redémarrer la machine avant de lancer un nouvel enregistrement.

*Un fichier « .deb » est l'équivalent Linux d'un fichier « .exe » sous Windows.

6. En cas de difficultés

Contacter l'équipe du support technique :

Mail : support-technique@lists.digi-newb.eu

Téléphone : 09 72 54 06 01

Titre : Système de surveillance des nouveau-nés prématurés par analyses vidéo et audio.

Mots clés : Nouveau-nés prématurés, développement neurocomportemental, surveillance, unités de soins néonatales, traitements audio, traitements vidéo.

Résumé : L'objectif de ces travaux, conduits dans le cadre du projet européen Digi-NewB et d'une thèse CIFRE, était de proposer une nouvelle approche non-invasive de monitoring en unités de soins intensifs néonatales (NICU). Ce nouveau monitoring doit permettre d'évaluer de façon continue l'évolution neurocomportementale des nouveau-nés prématurés à partir de modalités non-invasives telles que la vidéo et l'audio.

Après une étude bibliographique de plus de 150 documents, une première étude portant sur une estimation semi-automatique des stades de sommeil a été effectuée. L'approche proposée combinait pour la première fois des analyses vidéo et audio.

Les limites identifiées lors de cette étude ont permis de proposer un nouveau système audio-vidéo et d'étudier son intégration en NICU.

Des méthodes d'analyse vidéo, du son et de

classification (Random Forest, KNN, Réseaux de Neurones...) ont été proposées.

Elles permettent une caractérisation continue du comportement des nouveau-nés en termes de quantification des mouvements et d'analyse des pleurs.

Les difficultés liées aux contraintes des conditions réelles de NICU ont été étudiées et des solutions pour écarter les périodes non analysables (e.g., parents ou personnel médical dans le champ de la caméra, alarmes provenant des appareils médicaux) ont été développées.

Les résultats sont encourageants et montrent qu'il est aujourd'hui possible d'imaginer une nouvelle génération de monitoring basée sur des analyses non-invasives pour caractériser le développement neurocomportemental du nouveau-né.

Title: Monitoring of premature newborns by video and audio analyses.

Keywords: Neuro-behavioral development, monitoring, neonatal intensive care units, video processing, audio processing, preterm newborns.

Abstract: The objective of this work, conducted as part of the European project Digi-NewB and a CIFRE thesis, was to propose a new non-invasive approach to monitoring in neonatal intensive care units (NICUs). This new monitoring should make possible a continuous evaluation of the neuro-behavioural evolution of premature newborns using non-invasive modalities such as video and audio.

After a bibliographical study of more than 150 papers, a first study was carried out on a semi-automatic estimation of sleep stages. The proposed approach combined for the first time video and audio analyses.

The limitations identified during this study led to the proposition of a new audio-video system.

Its integration into NICU was studied and evaluated.

Then, methods, based on video and audio processing techniques and classification (Random Forest, KNN, Multi-layer Perceptron ...), were proposed. They allow a continuous characterization of the newborn behaviour in terms of movement quantification and cry analysis.

The difficulties related to the constraints of the real NICU conditions were studied and solutions to avoid irrelevant periods (e.g., parents or medical staff in the camera field of view, alarms coming from medical devices) were developed. The results are encouraging and show that it is now possible to imagine a new generation of monitoring based on non-invasive analyses to characterize the neuro-behavioural development of the newborn.