



Detection and characterization of vocalizations in preterm newborns

Bertille Met-Montot

► To cite this version:

Bertille Met-Montot. Detection and characterization of vocalizations in preterm newborns. Other. Université de Rennes, 2022. English. NNT : 2022REN1S043 . tel-03906092

HAL Id: tel-03906092

<https://theses.hal.science/tel-03906092>

Submitted on 19 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601

*Mathématiques et Sciences et Technologies
de l'Information et de la Communication*

Spécialité : Signal, Image, Vision

Par

Bertille MET--MONTOT

Detection and characterization of vocalizations in preterm newborns

Thèse présentée et soutenue à Rennes, le 28 Juin 2022

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

Rapporteurs avant soutenance :

VESIN Jean-Marc Maître d'Enseignement et de Recherche à l'EPFL, Lausanne, Suisse
HUEBER Thomas Chargé de Recherche CNRS (HdR), GIPSA-lab, Grenoble

Composition du Jury :

Président :	BEUCHÉE Alain	Professeur des Universités à l'Université de Rennes 1 et Praticien Hospitalier au CHU de Rennes
Examineurs :	VESIN Jean-Marc	Maître d'Enseignement et de Recherche à l'École Polytechnique Fédérale de Lausanne, Suisse
	HUEBER Thomas	Chargé de Recherche CNRS (HdR) au laboratoire Grenoble Images Parole Signal Automatique de Grenoble
	PATURAL Hugues	Professeur des Universités à l'Université Jean Monnet et Praticien Hospitalier au CHU de Saint-Etienne
	BEUCHÉE Alain	Professeur des Universités à l'Université de Rennes 1 et Praticien Hospitalier au CHU de Rennes
Dir. de thèse :	Fabienne PORÉE	Maitre de conférence (HdR) au LTSI-INSERM à l'Université de Rennes 1
Examineur :	Guy CARRAULT	Professeur au LTSI-INSERM à l'Université de Rennes 1

Toutes les grandes personnes ont d'abord été des enfants,
mais peu d'entre elles s'en souviennent.
Antoine de Saint-Exupéry - *Le Petit Prince*

... 47, 48, 49, ... 50 !!

J'espère que tu es bien caché Renard

J'arrive !!!



Résumé en français

Les naissances prématurées sont définies par l'Organisation mondiale de la santé comme survenant avant 37 semaines d'âge gestationnel [1]. Chaque année dans le monde, environ 15 millions de bébés naissent prématurément, soit plus d'un bébé sur dix. En France, cela représente un total de 60 000 naissances soit 8% chaque année et en raison de l'augmentation de l'âge moyen des femmes enceintes, de l'évolution des modes de vie ou du recours à la procréation médicalement assistée, ce nombre est en augmentation [2].

Un enfant né prématurément n'a pas fini de se développer, ses organes ne sont pas encore matures, fonctionnels, ou autonomes. L'immaturité des fonctions vitales telles que les fonctions digestives, cardio-respiratoires, immunologiques ou neurologiques entraînent une prise en charge particulière du nourrisson en Unités de Soins Intensifs Néonataux (USIN). Dans celles-ci le personnel assure une surveillance médicale élevée pour garantir le développement optimal du nourrisson.

Au début de sa vie extra-utérine, un grand prématuré est accueilli en couveuse, un espace qui tente de reproduire celui dont il bénéficiait dans le ventre de sa mère et où la température et le taux d'humidité y sont régulés (Figure 1). En fonction de ses besoins, il peut obtenir de l'aide i) respiratoire par intubation, ii) alimentaire ou médicamenteuse par perfusion intraveineuse centrale et/ou iii) alimentaire par sonde naso-gastrique. De plus, ses constantes vitales, c'est-à-dire sa respiration, son rythme cardiaque et son taux d'oxygène, sont constamment surveillées grâce à des électrodes placées sur son torse et à une sonde placée sur son pied. Ces dispositifs, illustrés sur la Figure 2, sont progressivement retirés au fur et à mesure que le nouveau-né se développe.

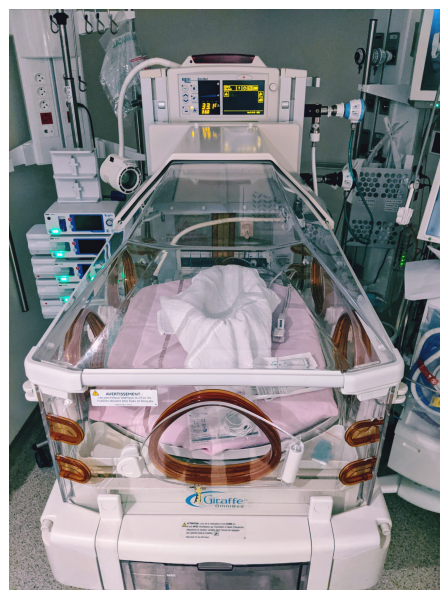


Figure 1: Couvercle au CHU de Rennes.

À la lumière de ces informations, il semble que de nouvelles solutions pour assurer une surveillance neuro-comportementale continue pourraient améliorer la prise en charge des nouveau-nés. C'est dans ce contexte que le projet européen Digi-NewB a vu le jour. Son objectif était de proposer un nouveau système de surveillance pour les soins des prématurés en s'appuyant à la fois sur des signaux traditionnels (signaux électrophysiologiques, signes cliniques) et à la fois sur de nouveaux capteurs non-invasifs encore jamais déployés en unités de soins intensifs tels

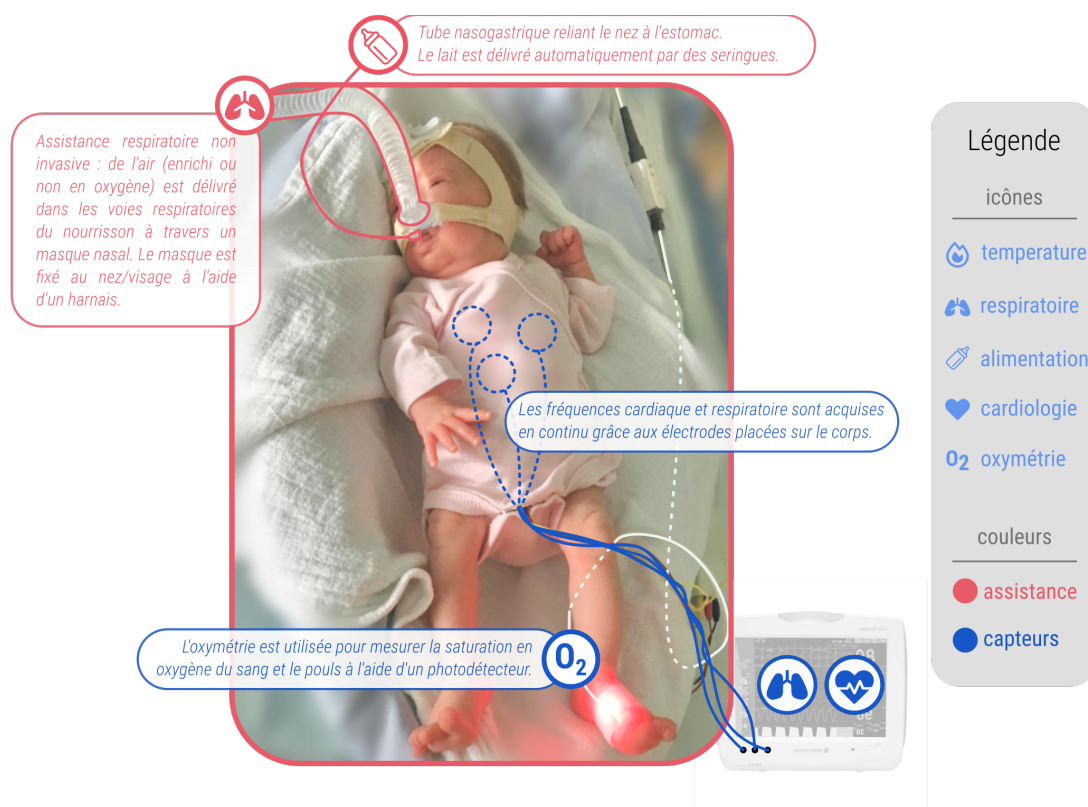


Figure 2: Illustration d'équipements médicaux nécessaires à la survie d'un nourrisson en USIN.

que des caméras et des microphones. Grâce à ces dispositifs, le projet Digi-NewB avait deux cibles cliniques : la détection précoce de l'infection nosocomiale et l'évaluation de la maturation cardio-respiratoire et neuro-comportementale des prématurés pendant leur hospitalisation. Ce projet a été réalisé grâce à la collaboration de sept partenaires européens publics et privés situés en Finlande, en France, en Irlande et au Portugal. Il a permis de recueillir des données sur plus de 600 nouveau-nés prématurés dans six centres hospitaliers de la région Grand Ouest en France.

Le travail décrit dans ce manuscrit se concentre sur l'un des objectifs de Digi-NewB : la détection et l'analyse de vocalisations de nourrissons prématurés. En effet, pleurer implique l'activation du système nerveux central et requiert des efforts coordonnés entre plusieurs régions du cerveau. En pleurant ou en ne pleurant pas, un nouveau-né alerte et informe sur son état physique et psychologique. C'est à partir de ce principe que de nombreuses études ont vu le jour sur le sujet. D'abord concentrés sur l'analyse des pleurs induits par la douleur [3–9], les scientifiques et cliniciens se sont ensuite tournés vers l'analyse des pleurs spontanés [10–12], notamment en tentant d'expliquer les différences observées entre les pleurs de prématurés et ceux d'enfants nés à terme [13–15]. Grâce au dispositif audio du projet Digi-NewB, c'est la première fois qu'autant de données ont été enregistrées. Aussi, l'objectif de ce travail de thèse était de développer une chaîne de traitement automatique pour la détection et la caractérisation des vocalisations des nouveau-nés prématurés dans un contexte hospitalier.

Comme nous l'avons mentionné précédemment, l'environnement des USIN est très chargé en matériel de soins. Les différentes machines utiles à la survie des nourrissons peuvent être très bruyantes et parfois émettre de nombreuses alarmes lorsqu'elles nécessitent une intervention de la part des infirmières. La difficulté de ce travail était donc l'automatisation du processus d'analyse des signaux de pleurs enregistrés dans cet environnement sonore parfois très bruyé. Pour exemple, les différentes sources sonores apparaissant dans les bandes-son sont présentées en **Figure 3**.

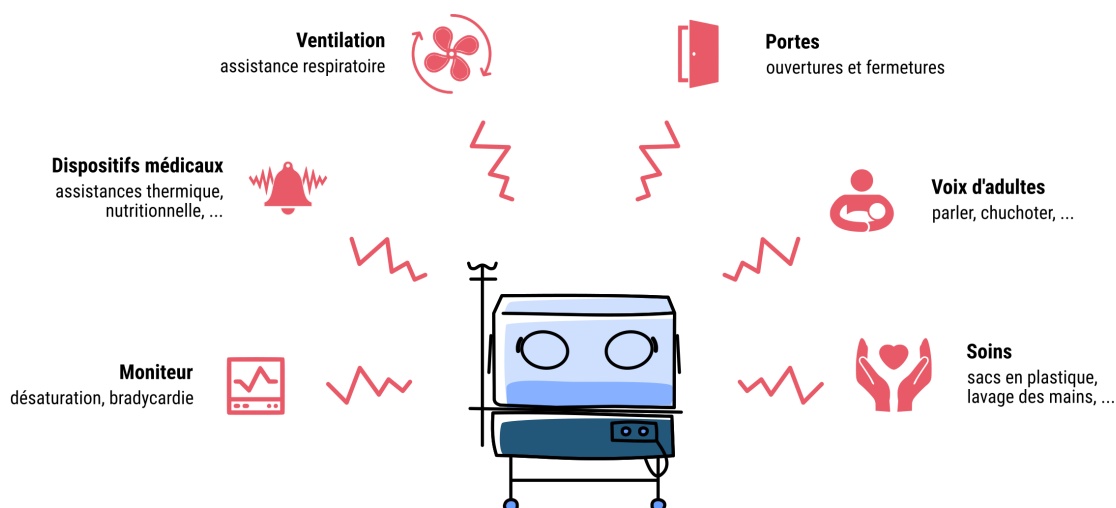


Figure 3: Sources sonores susceptibles de produire du son dans les chambres des nouveau-nés.

Pour répondre à cette problématique, nous avons choisi de développer une chaîne de traitement composée de trois étapes. D'abord à l'aide d'une segmentation des enregistrements, on extrait les portions d'audio qui contiennent les événements sonores, puis grâce à un modèle d'apprentissage profond, on détecte parmi les segments isolés ceux qui contiennent des pleurs. Enfin, on estime la fréquence fondamentale (F_0) de ces derniers à l'aide d'une méthode de détection de contours dans le spectrogramme. Ce processus est illustré sur la **Figure 4**.

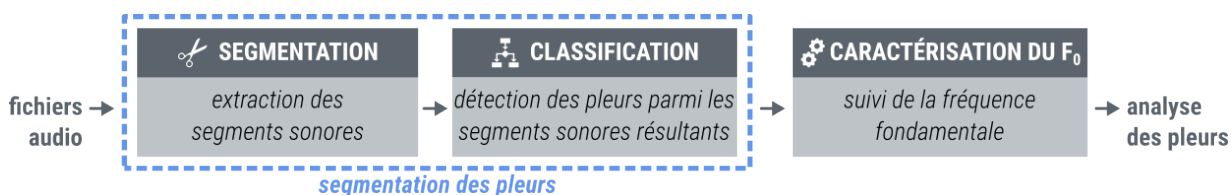


Figure 4: Chaîne de traitement automatique proposée pour l'analyse automatique des pleurs.

SEGMENTATION - La méthode de segmentation développée s'est inspirée de celle proposée par Orlandi et al. qui repose sur un calcul de l'énergie à court terme suivi d'un seuillage par la méthode d'Otsu [16]. Après avoir éliminé les fichiers audio de 30 minutes ne contenant pas de son, deux étapes sont ajoutées à la méthode pour l'améliorer. La première étape est un double filtre fréquentiel, la seconde est une re-segmentation.

L'évaluation de la méthode de segmentation en comparaison à des annotations manuelles, réalisées sur trois fichiers, a donné de bons résultats. En effet, nous avons montré qu'elle permet une bonne extraction des événements contenant des pleurs tout en réduisant le nombre de segments audio extraits. Pour aller plus loin, nous avons également proposé d'utiliser les informations de mouvements des nouveau-nés calculés par une autre équipe du LTSI au cours du projet Digi-NewB [17, 18]. En nous limitant aux sons apparaissant dans les intervalles détectés comme du mouvement, nous avons montré qu'il était possible de réduire considérablement la quantité de données à traiter tout en conservant des périodes riches en vocalisations. Ces dernières sont également très présentes dans les périodes de présence d'adultes. Cependant, nous avons choisi de les ignorer par souci de quantité et de complexité des données (superposition des voix et de pleurs, beaucoup de bruits liés aux soins, etc). L'évaluation de cette stratégie sur 303 heures d'enregistrements audio réalisés auprès de 22 nourrissons a montré que les nourrissons sont très peu en mouvement (12% du temps) et qu'en ne retenant que les sons issus de ces seules périodes cela permettait de supprimer jusqu'à 87% des segments initialement extraits.

CLASSIFICATION - La classification, après l'étape de segmentation, est nécessaire pour identifier les pleurs parmi les segments sonores extraits. Nous avons choisi d'utiliser une représentation temps-fréquence des pleurs (spectrogrammes) en entrée d'un algorithme de réseau de neurones convolutifs Resnet. La classification est ainsi réalisée en quatre étapes : i) calcul du spectrogramme par transformée de Fourier à l'aide de fenêtres de Hamming successives de 0.04 ms et d'un recouvrement de 95%, ii) découpage des spectrogrammes en images de même durée avec un recouvrement de 50%, iii) utilisation du réseau de neurones convolutifs pour la prédiction de la présence de pleurs dans les images, et iv) reconstitution des prédictions pour les sons en retenant la prédiction majoritaire sur l'ensemble des images. Grâce à un apprentissage réalisé par transfert, les poids initiaux de modèle ResNet ont été pré-entraînés avec ImageNet puis optimisés à notre tâche, c.-à-d., la classification pleurs vs non-pleurs, en réalisant un nouvel apprentissage. Pour adapter le modèle à nos données, les paramètres de durée des images d'entrées, de profondeur du réseau de neurones ainsi que le taux d'apprentissage initial ont été optimisés. Après une stratégie en deux étapes permettant d'abord de fixer le taux d'apprentissage, une évaluation de plusieurs combinaisons à l'aide d'une validation croisée a permis d'identifier le modèle avec la meilleure précision. Celui-ci correspond à des images d'entrée d'une durée de 0.25 s, une architecture ResNet34 et un taux d'apprentissage initial de 10^{-4} . Après avoir été à nouveau entraîné sur 30 bébés (17 042 sons), le modèle a obtenu de bonnes performances de classification sur un ensemble de trois nouveaux bébés (2 765 sons). Les résultats montrent que 86% des pleurs initialement annotés ont été détectés (sensibilité) et que 93% des sons classés comme pleurs sont effectivement des pleurs (précision).

CARACTÉRISATION DU F0 - Pour l'estimation du suivi de la fréquence fondamentale F_0 , nous avons proposé une nouvelle méthode de suivi de la fréquence fondamentale des pleurs de nourrissons dans le contexte d'un suivi en temps réel dans les unités de soins intensifs néonatales. Si les méthodes de la littérature fixent généralement la bande de fréquence dans laquelle effectuer le

suivi du F_0 [11, 16, 19, 20], nous avons proposé une étape initiale pour identifier automatiquement cette bande. Une fois calculé, le suivi de la fréquence fondamentale est effectué en utilisant une détection de contour dans le spectrogramme.

Pour valider notre méthode, nous avons comparé nos résultats d'estimation F_0 à ceux calculés par le logiciel BioVoice que nous avons identifié comme le programme de référence pour l'analyse de pleurs de nouveau-nés. En effet, la méthode développée par Manfredi et al. a obtenu de bonnes performances sur des formes mélodiques synthétiques de cris de nouveau-nés : [20, 21]. La comparaison qualitative des suivis de la fréquence fondamentale obtenus sur 806 pleurs a montré des estimations correctes dans 87% des cas avec BioVoice et 97% des cas avec notre méthode.

Finalement, la chaîne automatique de traitement a été déployée sur une base de données de 57 bébés nés prématurément et à terme et correspondant à 232 jours d'enregistrement. Grâce aux traitements successifs des trois méthodes proposées, nous avons été capables de détecter et de caractériser automatiquement 117 947 pleurs. Lors d'une comparaison avec la littérature, nous avons montré que nos résultats sont cohérents avec deux études qui observent la fréquence fondamentale i) des prématurés en fonction de leur âge gestationnel ou de leur poids de naissance [11] et ii) des nourrissons prématurés et des nourrissons nés terme à un même âge post-menstruel [14]. Ensuite, grâce aux enregistrements longitudinaux réalisés auprès des bébés tout au long de leur hospitalisation, nous avons présenté les évolutions de la durée et de la fréquence fondamentale des pleurs en fonctions des âges post-menstruels et postnatals. Enfin, pour la première fois l'évolution de la fréquence fondamentale pour une population de nourrissons prématurés d'évolution normale est décrite et tracée. Ces résultats sont une avancée majeure pour l'évaluation de la maturation des nouveau-nés prématurés pendant leur hospitalisation.

En conclusion, si ce travail de thèse apporte les outils pour l'évaluation de la maturation et des tendances d'évolution des paramètres des pleurs en fonction de l'âge dans un contexte de soin courant en unités de soins intensifs, il n'en reste pas moins que beaucoup d'améliorations sont à apporter. Les perspectives s'inscrivent naturellement dans la dynamique déjà étudiée et auront comme volonté de traiter le plus de données pour confirmer, renforcer les tendances observées et couvrir la plus large période d'hospitalisation possible dans l'objectif d'apprécier les déviations éventuelles liées à des infections ou des pathologies. Ces travaux seront alors le socle des développements futurs afin de progresser vers une solution entièrement automatique pour une nouvelle génération de systèmes non-invasifs de surveillance en temps réel des nouveau-nés prématurés par l'intermédiaire de l'analyse audio.

BIBLIOGRAPHY

- [1] WORLD HEALTH ORGANIZATION. Who: Recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. *Acta Obstetrica et Gynecologica Scandinavica*, vol. 56, 247–253 (1977).
- [2] SOSPREMA. La prématurité.
Accessed on 25/10/2021 from: <https://www.sosprema.com/la-prematurite/definition/>
- [3] MICHELSSON K. Cry analyses of symptomless low birth weight neonates and of asphyxiated newborn infants. *Acta Pædiatrica*, vol. 60, 9–45 (1971).
- [4] 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*, vol. 56, 975–80 (1974).
- [5] MICHELSSON K., JÄRVENPÄÄ A., AND RINNE A. Sound spectrographic analysis of pain cry in preterm infants. *Early Human Development*, vol. 8, 141–149 (1983).
- [6] THODÉN C.J., JÄRVENPÄÄ A.L., AND MICHELSSON K. Sound spectrographic cry analysis of pain cry in prematures. In *Infant Crying*, 105–117. Springer (1985).
- [7] 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*, vol. 52, 201–208 (1993).
- [8] STEVENS B.J., JOHNSTON C.C., AND HORTON L. Factors that influence the behavioral pain responses of premature infants. *Pain*, vol. 59, 101–9 (1994).
- [9] 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*, vol. 42, 850–61 (1999).
- [10] WERMKE K., MENDE W., MANFREDI C., AND BRUSCAGLIONI P. Developmental aspects of infant's cry melody and formants. *Medical Engineering & Physics*, vol. 24, 501–14 (2002).
- [11] MANFREDI C., BOCCHI L., ORLANDI S., SPACCATERRA L., AND DONZELLI G.P. High-resolution cry analysis in preterm newborn infants. *Medical Engineering & Physics*, vol. 31, 528–32 (2009).
- [12] ANDRÉ V., DURIER V., HENRY S., NASSUR F., SIZUN J., HAUSBERGER M., AND LEMASSON A. The vocal repertoire of preterm infants: Characteristics and possible applications. *Infant Behavior and Development*, vol. 60, page 101463 (2020).
- [13] ORLANDI S., BOCCHI L., DONZELLI G., AND MANFREDI C. Central blood oxygen saturation vs crying in preterm newborns. *Biomedical Signal Processing and Control*, vol. 7, 88–92 (2012).
- [14] 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*, vol. 10 (2014).
- [15] ORLANDI S., REYES GARCIA C.A., 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*, vol. 30, 656–663 (2016).
- [16] 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*, vol. 8, 799–810 (2013).

- [17] 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*, vol. 38, 207–213 (2017).
- [18] WEBER R., SIMON A., PORÉE F., AND CARRAULT G. Deep transfer learning for video-based detection of newborn presence in incubator. In *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2020, Montreal, QC, Canada, July 20-24, 2020*, 2147–2150. IEEE (2020).
- [19] 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*, vol. 20, 85–99 (2015).
- [20] ORLANDI S., BANDINI A., FIASCHI F., AND MANFREDI C. Testing software tools for newborn cry analysis using synthetic signals. *Biomedical Signal Processing and Control*, vol. 37, 16–22 (2017).
- [21] 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*, vol. 45, 174–181 (2018).

Remerciements

Je remercie tous les membres de mon jury de thèse d'avoir apporté leurs expertises à ces travaux. Je remercie Jean-Marc VESIN maître d'Enseignement et de Recherche à l'École Polytechnique Fédérale de Lausanne pour ses remarques concernant les limites de certaines parties des méthodes abordées dans ce document et pour toutes les corrections suggérées. Thomas HUEBER chargé de Recherche CNRS au laboratoire Grenoble Images Parole Signal Automatique de Grenoble pour toutes ses remarques très pointues sur l'ensemble du manuscrit et particulièrement sur l'utilisation d'un apprentissage par transfert dans le cadre d'une classification de spectrogrammes de sons. Hugues PATURAL Professeur des Universités à l'Université Jean Monnet et Praticien Hospitalier au CHU de Saint-Etienne pour ses remarques sur la nature très spéciale des relations parents-enfants à la naissance. Enfin, Alain BEUCHÉE président du jury et Professeur des Universités à l'Université de Rennes 1 ainsi que Praticien Hospitalier au CHU de Rennes pour m'avoir ouvert une fois de plus les yeux sur la spécificité et l'unicité de chacune des naissances et des développements des nouveau-nés prématurés.

Je remercie également Patrick PLADYS, Professeur des Universités à l'Université de Rennes 1 ainsi que Praticien Hospitalier au CHU de Rennes pour l'accompagnement, les nombreuses discussions et la quantité d'idées apportées à chacune de nos rencontres. Ton engouement pour mes recherches et tes mots m'ont convaincu de l'utilité de mon travail. Je tiens également à remercier tout le personnel médical que j'ai croisé, je pense notamment à Florence GESLIN, Céline CITTE, mais aussi Sabrina VALENTIN ou encore Murielle DUFAU Cadre de Santé à l'hôpital Sud de Rennes qui m'ont toutes accueillies chaleureusement et qui m'ont aidé à mieux comprendre les services de néonatalité. Je remercie également toutes les infirmières impliquées dans le projet Européen Digi-NewB qui ont réalisé les enregistrements audio qui sont un atout majeur pour mon travail.

Je remercie ensuite mes directeurs de thèse, Fabienne PORÉE et Guy CARRAULT pour leur présence, leurs conseils et leur soutien. Je vous remercie pour vos encouragements et pour toute la confiance que vous m'avez donnée durant ces dernières années. Vous m'avez également accordé une grande liberté de recherche et développement et je vous remercie de m'avoir soutenue dans mes démarches. Cette thèse a été la meilleure formation que j'ai réalisée étant à la fois maître et élève et étant accompagnée d'une équipe compréhensive et accueillante. Si le parcours en lui-même s'est révélé ardu et m'a demandé beaucoup de sacrifices, je crois que les compétences acquises pendant ces dernières années me suivront tout au long de mon parcours futur et je me sens aujourd'hui capable d'accomplir ou d'entreprendre n'importe quel projet.

Pour continuer avec l'équipe, je souhaite transmettre un immense merci à Sandie. Tu as été une mentore exceptionnelle et je te remercie infiniment pour tout le temps et l'énergie que tu m'as accordés. Tu m'as donné confiance en moi et tu m'as toujours soutenue. Merci de ta présence bienveillante à mes côtés. Merci aussi à tous les membres de l'équipe Digi-NewB, je pense particulièrement à Édouard, Raphaël, Thomas, Crysthine et Gustavo, que ce soit pour le côté professionnel ou personnel, j'ai passé de bons moments en vos compagnies.

Je remercie également chaleureusement 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 mon arrivée. Merci à Soizic, Patricia et Murielle pour tout ce qui concerne le côté administratif de la thèse et aussi pour toutes les conversations pleines d'énergie positive et de sourires qu'on a pu partager. Merci également à Huazhong SHU, directeur du Laboratoire Image, Sciences et Technologies de l'université du Sud-Est de Nankin (Chine). En effet, grâce à la collaboration de ces deux laboratoires, j'ai eu l'incroyable opportunité de réaliser une partie de mon stage de Master à Nanjing où j'ai eu la chance de découvrir une population généreuse, une cuisine délicieuse et une culture merveilleuse ! Je remercie également les étudiants chinois qui m'ont accompagné sur place et notamment ma tendre amie Pan Tan ainsi que mes partenaires de discussion Neil et Sylvia. Je remercie affectueusement Murun de m'avoir acceptée dans ta chambre à la cité internationale de Nanjing et aussi pour m'avoir donné la chance de découvrir ton pays natal : la Mongolie. Je n'oublierai jamais ces voyages alors un grand merci à tous ceux que j'ai croisé.

Je remercie également l'équipe pédagogique de l'IUT GEII Rennes qui m'a également réservé un accueil doux et chaleureux et qui m'a fait découvrir le monde de l'IUT. Merci particulièrement à mes collègues Emmanuel et Mickaël pour les grandes discussions sur l'électronique et la programmation ainsi que pour les récits d'aventures époustouflants. J'ai eu la chance de pouvoir voyager à travers des souvenirs précieux alors merci pour le partage !

Un grand merci aussi à Jean-Claude pour toutes les conversations courtes ou longues qu'on a pu échanger au détour d'un bureau, d'un couloir ou d'un café. Notre rencontre à Nanjing a été formidable et je suis ravie d'avoir pu continuer de te croiser et d'échanger avec toi ses dernières années. Ton sourire est lumineux et tes moqueries m'ont permis de lâcher prise dans des moments laborieux. Merci pour ta bonne humeur et ton humour authentique ! Je te souhaite beaucoup de bonheur et peut-être un peu plus de voyages ou d'aventures ? On a qu'une vie alors profitons en ;)

Je remercie ensuite tous les membres de mon bureau, le 408 aura été le centre de discussions intenses en petit ou grand comité, mais il aura surtout toujours été un lieu de partage et de soutien. Un lieu où je me suis sentie entourée d'amis sincères. Alors, merci Houda pour tes questions et tes compliments sur mes illustrations. Merci Valentin pour les discussions à n'en plus finir sur le deep-learning, tu m'as rassuré quand j'en avais besoin et ça m'a fait beaucoup de bien ! Je vous souhaite à tous les deux beaucoup de courage, il en faudra à certains moments, soyez-en

sûr ! Merci ensuite Pablo pour toutes nos conversations où on a refait le monde tard le soir, ce fut des moments suspendus. Une petite parenthèse ici pour Karim, merci pour nos rendez-vous "réguliers" aux restos ou autour d'un verre, je suis contente qu'on ait réussi à garder le contact entre Rennes et Lyon. Ton sourire fait chaud au cœur tout comme les encouragements que tu m'as offerts. Merci Yannick pour ton énergie et ton positivisme permanent avec ta phrase fétiche du : "T'inquiètes, ça va aller !". Merci pour tout le temps passé à m'aider à mettre mon réseau de neurones en place, merci pour toutes tes réponses, ta patience, ta transmission, et tout le courage que tu m'as donné pendant toutes ces années. Enfin, merci Soumaya pour ta bienveillance, ton sourire, toutes les gourmandises que tu m'as données ;) Réaliser cette thèse en ta compagnie m'a rendu la vie bien plus douce à des moments parfois bien compliqués ... Tu as été un phare plein de lumière et de joie me guidant dans les nuits sombres et dans les chemins parfois tortueux que peut prendre une thèse. Je suis sincèrement heureuse d'avoir partagé tous ces moments avec toi et je te souhaite un bonheur éternel ma tendre amie.

Merci à toutes les personnes qui ont annoté des extraits sonores ce qui m'a permis de constituer, ce que je pense être la première grande base de données de pleurs de prématurés. Cette constitution est un atout majeur de ma thèse et je tiens à remercier particulièrement Paul pour les très nombreuses écoutes et ta grosse contribution à ce travail si minutieux. Merci, aussi, de ta présence et ton soutien au laboratoire, tu es un acharné du travail et je te souhaite le meilleur pour la suite, n'oublies pas de vivre un peu pour toi quand même ;)

Je remercie ensuite les différents maîtres sportifs qui m'ont accompagné tout au long de ce périple. Comme l'a dit Juvénal "Mens sana in corpore sano" ! Je sais que sans ces efforts physiques, je n'aurais jamais tenu psychologiquement, les séances m'ont permis de me défouler et de me vider la tête là où nulle autre activité ne me le permettait. Pendant ces années de thèse, je me suis donnée à fond sur les terrains de badminton et j'ai progressé comme jamais auparavant. Merci Marine pour tes conseils avisés, je n'oublierai pas ces rendez-vous hebdomadaires qui m'ont fait tant de bien ! Merci aussi à Jaff et Hélène de m'avoir fait découvrir l'Aïkido, ai : l'harmonie, ki : l'énergie et dō : la voie. À travers vos enseignements, j'ai découvert un art qui m'a transporté à la fois physiquement, par la technique, et spirituellement. Merci Hélène pour les riches explications et nombreux questionnements qui ont profondément attisé ma curiosité. Tu m'as enseigné l'art du mouvement, que rien n'est jamais fixe et qu'il ne tient qu'à nous d'en prendre conscience. Grâce à toi, j'ai ouvert mon esprit et j'ai découvert mon corps. Merci à vous trois pour vos enseignements et vos pratiques et surtout merci pour le cadre convivial et toutes les rencontres que vous m'avez offert. Je pense notamment à toute la fine équipe d'aikidoka du Budo Raji, merci pour les séances à vos côtés et pour les pizzas aussi !

Merci aussi à François pour ta générosité, ta douceur et ta grande sensibilité. Tu fais partie de ceux qui offrent des compliments dans l'instant présent et je tenais à te remercier pour tes mots. Ils m'ont encouragé et m'ont fait beaucoup de bien. Merci pour toutes les fois où l'on a pratiqué ensemble et pour toutes les conversations qui s'en sont suivies. Ta présence dans le bâtiment fut un cadeau plus grand encore que tu ne l'imagines.

Merci à Lucille, avec qui j'ai partagé de chouettes moments à Rennes, que ce soit les boîtes de nuit improvisées à la maison ou les sorties au bar, on a bien rigolé ! Merci d'avoir été présente aux deux grandes étapes de la thèse à savoir le rendu du manuscrit et le repas de soutenance. Je suis contente d'avoir pu profiter de ton expérience et de tes mots pendant tout ce long parcours. Je te souhaite beaucoup de bonheur pour la suite !

Merci à l'équipe Tout En Vélo qui m'a généreusement offert le café le midi quand je cherchais un peu de lien social. Merci notamment à Corentin pour tous nos échanges très constructifs, sur le café de spécialité, le bon vin et surtout merci de m'avoir offert une oreille attentive dans des moments parfois difficiles. Merci aussi de m'avoir convaincu, malgré toi, de placer ces remerciements après le résumé tel un groupe de Rock demandant à la foule "COMMENT ÇA VA CE SOIR ?" après avoir joué la première chanson. Puisse ton vélo te mener vers de beaux et joyeux chemins !

Je remercie aussi mes parents d'être restés curieux tout au long de ma thèse, de m'avoir posé tant de questions et de m'avoir aidé aussi à me remettre en question ! Merci d'être si présents dans ma vie et de me soutenir dans toutes les situations. Je me sens extrêmement fière et heureuse d'être votre fille. Merci aussi de m'avoir donné la plus chouette des petite sœur qui m'impressionne et qui me donne du courage dès que j'en ai besoin. Il nous en aura fallu du temps, mais je suis comblée par nos échanges, notre complicité et tout le reste. Merci pour ton illustration qui fait la première page de ce manuscrit et merci de tous les compliments dont tu m'abreuves à chacune de nos rencontres. Tu m'as rendue plus forte. Je tiens aussi à te remercier ma très chère Grand-Mère, pour ta présence et ton soutien, mais surtout pour l'amour incommensurable que tu m'offres. Je mesure la chance de t'avoir dans ma vie et je te remercie pour tout. Merci aussi à vous mon Papé et ma Mamé de m'avoir enseigné l'art de la promenade en forêt ainsi que celui de la couture. Grâce à vous, j'ai développé une curiosité de tout et j'ai pu créer mes propres vêtements pour ma soutenance de thèse.

Merci à toi Arthur, d'être venu et d'avoir parcouru toute cette distance, littéralement, pour être revenu jusqu'à moi. Le chemin est encore long, mais tous ensemble rien ne pourra nous arrêter sois en sûr.

Je voudrais ensuite remercier chaudement tous les fourbes escapadés d'être mes amis de longue date maintenant. Entre la Turballe, Bordeaux, Montaimont, Paris, Le Mans, Agen, Lyon, Rodez, Montpellier, Oléron on en a vécu des aventures ... Ces escapades m'ont fait le plus grand bien et ont été une source d'énergie à l'état pur. Merci d'être là les copains et merci pour tous les encouragements et le soutien que vous m'avez donné ! Merci surtout à Ninouche qui a fait le déplacement le jour J et dont la présence est d'un immense soutien. Babylone ma sœur, je n'oublierai jamais notre périple à vélo et tous nos débats existentiels sur le monde actuel ! Bien sûr, je tiens à remercier plus que quiconque my Jane pour tous les messages d'encouragements quotidiens qu'il m'a fallu pour amorcer cette fichue rédaction de manuscrit... Merci d'avoir été

là pour moi et merci de m'avoir malmené sur les dead-lines, rien de tout cela n'aurait pu être possible sans toi ! Sans oublier les nombreux week-ends et excursions surtout en bord de mer qui ont fait, à chaque fois, souffler un vent nouveau sur mes sentiments embrumés. Tu fais partie de celles qui m'offrent un amour et un réconfort sans faille et tu m'impressionnes par l'énergie que tu déploies dans chacun de tes projets. Pour la source d'inspiration que tu es, je te souhaite de ne jamais te tarir et de tracer ta route. Quelle qu'elle soit, je suis convaincue qu'elle sera belle et qu'elle est vouée à croiser sans cesse la mienne.

Merci aussi à tous mes amis cévenols avec qui j'ai partagé ces dernières années des morceaux d'étés et des fins d'années. Merci à Ulysse et Aude d'être ceux qu'ils sont, vous m'avez donné votre cœur et je suis sincèrement heureuse de vous connaître. Merci à Rose et Vincent pour tous les moments de complicité aux repas de famille et ailleurs, avec vous, je rigole librement et je me sens vivre. Bien que la distance nous sépare, je savoure chaque instant en votre compagnie ! Merci aussi à vous, Sylvie et Pascal de m'avoir acceptée telle que je suis. Votre maison est chaleureuse et je m'y sens bien, merci de m'offrir un environnement aussi convivial et généreux fort de discussions et de gourmandises !

Merci aussi à mes deux meilleures amies qui sont physiquement les plus loin, mais qui m'ont guidée jusqu'à la personne que je suis aujourd'hui. Merci Philippine pour tout ce qu'on a partagé depuis notre tendre adolescence, si nos chemins s'écartent parfois, je sais qu'ils finissent toujours par se recroiser. Je te souhaite encore beaucoup d'aventures et de rencontres que nous pourrons ensuite partager toute une nuit sous quelques étoiles. Merci aussi à Myway d'être ma sœur de vie depuis dix ans déjà, si notre rencontre était loin d'être gagnée, je sais qu'elle est le fruit du destin et qu'avec toi aussi nos chemins sont loin de s'écarter. Après tout, nos routes se sont déjà croisées à Angers, Rendsburg, Aigle, Paris, Prague, Malmö, Hamburg, Berlin, Toulouse, et si on ne s'est pas croisé à Shanghai, c'est que la suite nous réserve encore bien des surprises ! Malgré la distance, avec toi je partage tout, mes doutes, mes erreurs, mes envies, mes rires, ... Ensemble rien ne nous arrête, ensemble on apprend à s'envoler ! Merci d'être dans ma vie et merci pour tous les encouragements que tu m'as offerts. Nos discussions m'ont été très précieuses. Je te souhaite d'être libre et heureuse ma très chère amie.

Enfin, je remercie tous les colocs avec qui j'ai partagé mon quotidien à Saint-Hélier à savoir, dans l'ordre, Mathilde, Jordan, Léa, Rafael, Justin et Lucille. Les Français Libres c'est un lieu où il fait bon vivre et où je me sens vraiment à la maison. Un grand merci à ma famille rennaise !

Merci Lucille pour la douceur et la sérénité que tu apportes à la maison depuis ton arrivée. On ne s'est pas rencontré au meilleur moment de ma vie, mais sache que ta présence m'a fait beaucoup de bien dans ces moments de tensions intenses. Je te souhaite beaucoup de joie et d'expéditions pour la suite. Où que ce soit, je suis sûre que tu trouveras ton moment pour te lancer dans des aventures solaires et merveilleuses.

Mon Cher Ghassan, je suis si heureuse de t'avoir rencontré et d'avoir vécu tous ces moments forts en ta présence. Si ton histoire est dure, tu m'impressionnes par ton énergie et les ondes si positives que tu dégages. Tu es celui qui m'a chanté l'encouragement pendant ces longs mois de rédaction et ta voix résonne encore dans ma tête : "Bertille, Bertille, Bertille". Merci pour ton soutien, merci pour tes mots, merci de m'avoir rassuré encore et encore. Je te souhaite beaucoup de bonheur et d'épanouissement, tu le mérites sincèrement.

Justin, ces mots sont les plus durs à écrire, car je n'arrive pas à trouver assez de recul pour mesurer l'envergure de ce que tu représentes pour moi. Je te remercie infiniment de toujours m'avoir soutenue pour tous les projets que j'ai entrepris. Avec tes encouragements, j'ai non seulement réussi à finir cette thèse, mais je suis aussi partie en Erasmus à Prague, en stage en Chine puis en voyage en Corée, au Japon, en Mongolie et en Russie. Merci de m'offrir la liberté d'être moi-même et de m'accepter entièrement. Merci d'écouter attentivement mes peurs, mes envies, mes pulsions et de m'accompagner dans mes réflexions et dans mes projets. Ton calme légendaire attendrit mes tempêtes intérieures et faisant état du chemin parcouru, je me dis qu'on a encore beaucoup de choses à s'apprendre et d'excursions à mener. Jusqu'ici, j'admire nos progrès, chacun avec nos forces et nos faiblesses, et je suis fière et heureuse de ce que nous avons construit. Si la distance a fait partie intégrante de notre vie ces dernières années, je suis comblée de te savoir venir "éprouver ton corps et ta tête dans les champs de bananes d'Océanie où l'on se lève à l'aube". J'ai hâte de savourer chaque instant de vie et d'aventure en ta présence. Merci de tout le soutien que tu m'apportes et de m'offrir à la place de beaucoup de mots, un regard qui en dit long, lorsque tu me souris.

Léa, les mots sont durs à trouver tant il y a de sujets à aborder. Merci pour tout. Merci d'avoir été la plus incroyable colocataire dont je puisse rêver un jour. Tu es le soleil de ma vie dans cette région si réputée pour sa météo. Merci pour tout ce qu'on a partagé. Je repense aux concerts, aux cinés, aux restos, aux promenades, aux repas, aux activités manuelles, aux rangements, aux questions existentielles, aux doutes, aux réconforts, aux larmes, aux joies, aux rires. Tu as partagé mon quotidien depuis trois années maintenant et tu es mon âme, tu es mon cœur et je t'ai dans la peau comme personne. Merci d'être toi, merci pour tout ce que tu m'as appris pour toutes les remises en question et tous les mots rassurants que tu m'as donné. Je me sens comblée de cette rencontre par tout ce que tu m'as apporté et que je n'aurai jamais imaginé... Alors c'est ça ce qu'ils appellent grandir ? Merci d'avoir un si grand cœur et une telle sensibilité. Comme tu l'as dit, mon départ de la coloc n'est pas une fin, mais le début d'une autre histoire. J'ai hâte de te retrouver pour nos futures soirées pyjamas où l'on n'aura pas fini de refaire le monde !

Merci encore à tous ceux que j'ai croisés au cours de ma vie et qui m'ont permis d'atteindre mes objectifs. Aujourd'hui, j'ai reçu le titre de docteur et si l'obtention du diplôme marque la fin d'une époque, il marque surtout pour moi le début d'un nouveau voyage au cours duquel mes années d'études me permettront de chercher celle que j'ai envie d'être.

*Atteindre le sommet leur prit encore deux heures.
Deux heures durant lesquelles Ellana batailla ferme pour avancer.
Batailla contre la montagne et contre ses chaînes.
Deux heures de combat épuisant où elle prit des risques incroyables.
Deux heures passées sans échanger le moindre mot avec Jilano.
Deux heures de bonheur.
Après une ultime traction, elle se retrouva à plat ventre dans la neige.
Il n'y avait plus rien au-dessus d'elle que l'infini du ciel.
Elle se leva lentement.
Le pic qu'ils venaient de gravir se dressait isolé, comme unique prétendant à l'absolu et,
debout à son sommet, Ellana eut soudain l'impression qu'elle pouvait tutoyer le soleil.
Elle ouvrit la bouche pour une exclamation ravie...
La referma.
Jilano se tenait près d'elle et dans ses yeux bleu pâle brillait une lumière nouvelle.
Intense et feutrée, forte et douce, rayonnante et triste. Humaine et tellement plus que cela.
Il s'approcha d'Ellana, la contempla comme s'il la découvrait pour la première fois,
puis, doucement, il lui ôta ses chaînes.
Les jeta au loin.
- Tu es libre, annonça-t-il.
...*

Le Pacte des MarchOmbres, Tome 2 : Ellana, l'envol
Pierre BOTTERO

Contents

Introduction	9
1 About infant crying	13
1.1 Introduction	13
1.2 Prematurity	13
1.3 About crying	15
1.3.1 Anatomy and physiology	15
Mechanical production	15
Neurological production	17
1.3.2 Prosodic feature definitions	18
1.4 State of the art	20
1.4.1 Clinical investigations of crying	20
Full-term newborns and infants	20
Premature newborns	21
1.4.2 Methods for acoustic signal processing	21
Crying data acquisition and databases	22
Audio signal processing	22
Crying feature extraction	23
1.5 Our strategy	26
Bibliography	27
2 Work context	31
2.1 Introduction	31
2.2 Description of the NICU	31
2.3 The Digi-NewB proposal	36
2.4 Acoustic environment in the NICU	40
2.4.1 Sounds in Digi-NewB recordings	40
2.4.2 Noise quantification in a 15-hour annotated recording	42
2.4.3 Sounds variability in recordings	45
2.5 Conclusion	47
Bibliography	48

3	Audio-Video segmentation	51
3.1	Introduction	51
3.2	State of the art	51
3.2.1	Methods for cry segmentation in short recordings	52
3.2.2	Methods for cry segmentation in long and noisy recordings	52
3.2.3	Discussion	53
3.3	Audio segmentation method	54
3.3.1	Orlandi's method	54
	Thresholds computation on Digi-NewB data	58
	Issues related to poor boundaries detection	59
	Issues with replication	60
3.3.2	Improvement of the method	61
	Re-segmentation (RS)	61
	Narrowing STE frequency band (NFB)	61
	Long-term threshold (LTT)	62
3.4	The use of motion for audio segmentation	63
3.4.1	Video segmentation	64
3.4.2	Database	66
3.4.3	Motion quantification	66
3.4.4	Sounds within infants' movement	66
3.4.5	Cries within infants' movement	68
3.4.6	Discussion	69
3.5	Evaluation strategy	70
3.5.1	Segment comparison	70
3.5.2	Duration comparison	71
3.6	Results	72
3.6.1	Database	72
3.6.2	Audio segmentation improvements evaluation	73
	Reproduction of Orlandi's method	73
	Narrowing Frequency Band (NFB) improvement	75
	Re-Segmentation (RS) improvement	75
	Long-term threshold (LTT)	75
	Audio-Video segmentation	77
3.7	Conclusion	77
	Appendix A - Otsu's method	79
	Bibliography	80

4	Classification for cry detection	83
4.1	Introduction	83
4.2	State of the art	83
4.2.1	Feature extraction	84
4.2.2	Classification methods	86
4.2.3	Evaluation metrics	86
4.3	Proposed method	88
4.3.1	Spectrogram computation	88
4.3.2	Transfer learning using ResNet architectures	90
4.3.3	Model training	92
	Best candidate combinations selection	92
	Final combination selection	93
4.4	SoundAnnoT: database creation	93
4.4.1	Interface	94
4.4.2	Labels	95
4.4.3	Annotations	95
4.4.4	User procedure	96
4.5	Results	97
4.5.1	Annotated data	98
4.5.2	Best candidate combinations selection	99
4.5.3	Final combination selection	101
4.5.4	Deployment of the final combination	102
4.6	Conclusion	104
	Appendix A - Best candidate combinations selection	106
	Appendix B - Final combination selection	107
	Bibliography	108
5	Fundamental frequency characterization	111
5.1	Introduction	111
5.2	State of the art	111
5.2.1	Methods	112
	Time domain	112
	Spectral domain	112
	Wavelet domain	113
	Image domain	113
5.2.2	Softwares	113
5.2.3	Limitations on the fundamental frequency estimations	114
5.3	Proposed method	115
5.3.1	Spectrogram computation	115

5.3.2	Automatic frequency band selection	116
5.3.3	Contour detection	117
5.3.4	Fundamental frequency tracking	117
5.4	Evaluation strategy	117
5.4.1	BioVoice software	118
	Software execution	118
	Cry characterization for comparison	119
	Limitations	119
5.4.2	Qualitative comparison with BioVoice	119
5.4.3	Statistic parameters comparison	120
5.5	Results	120
5.5.1	Annotated database	121
5.5.2	Qualitative comparison with BioVoice	121
	Both estimations are equivalent	122
	BioVoice method is better	123
	The proposed method is better	124
5.5.3	Performance and parameters comparison	126
5.6	Conclusion	127
	Bibliography	128
6	Automatic processing for cry analysis: deployment	131
6.1	Introduction	131
6.2	State of the art	132
6.2.1	Pain induced cries	132
6.2.2	Spontaneous cries	133
6.3	Deployment of the proposed methods	133
6.3.1	Database	134
6.3.2	Audio-Video segmentation	134
6.3.3	Classification for cry detection	134
6.3.4	Data management	136
6.3.5	Fundamental frequency characterization	136
6.4	Reproduction of existing studies	137
6.4.1	Fundamental frequency according to GA and birth weight	137
6.4.2	Fundamental frequency at term-equivalent age	139
6.5	New cry characterization insights	142
6.5.1	Fundamental frequency comparison at two postnatal ages	142
6.5.2	Crying evolution with age	144
	Fundamental frequency evolution with PMA	144
	Duration evolution with PMA	146

Duration evolution with PNA in extreme preterm newborns	146
6.5.3 Fundamental frequency longitudinal evolution	147
6.6 Conclusion	149
Bibliography	150
Conclusions & perspectives	151
List of publications	157

Acronyms

AUC	Area Under the Curve
CHU	University Hospital
CNN	Convolution Neural Network
CU	Cry Unit
EP	Extreme Preterm
FN	False Negative
FP	False Positive
FT	Full-term
GA	Gestational Age
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IUGR	IntraUterine Growth Retardation
KNN	K-Nearest Neighbors algorithm
LFB	Linear-Filter Bank
LPCCs	Linear Prediction Cepstral Coefficients
MFB	Mel-Filter Bank
MFCCs	Mel-Frequency Cepstral Coefficients
MLP	Moderate to Late preterm
NICU	Neonatal Intensive Care Unit
NIDCAP	Newborn Individualized Developmental Care and Assessment Program
PCA	Principal Component Analysis
PMA	Post Menstrual Age
PNA	Post Natal Age
PR-AUC	Area Under Precision-Recall Curve
PT	Preterm
SIFT	Simple Inverse Filter Tracking
STE	Short Time Energy
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VP	Very Preterm
ZCR	Zero Crossing Rate

Introduction

Preterm births are defined by the World Health Organization as babies born alive before 37 weeks of pregnancy are completed [1]. Each year in the world, approximately 15 million babies are born prematurely, that is to say more than one baby out of ten. In France, it is 165 births per day with a total of 60,000 births which represent 8% of the births each year. Moreover, due to the increase in the average age of pregnant women, the evolution of lifestyles or the use of medically assisted reproduction, this number is rising [2].

Prematurity is the leading cause of newborn death worldwide and the second leading cause of child death after pneumonia. Most premature infants who survive face a lifetime of disability [3]. All newborns are vulnerable, but premature babies are even more fragile because early birth has prevented complete organ development. Thus, these babies have immature functions such as digestive, cardiorespiratory, immunological or neurological and require special care to stay alive. Therefore, these babies are cared for in Neonatal Intensive Care Units (NICU), where high medical supervision is provided by the medical staff to ensure their optimal development.

Although each infant's development is unique, the journey of a very premature baby begins in an incubator where she/he is usually given various aids such as respiratory intubation, central intravenous infusion or feeding tubes. These invasive devices are removed as the newborn develops and becomes more independent. Unfortunately, infants born very early have a very immature immune system and are therefore more exposed to nosocomial infections from these invasive procedures [4, 5].

However, scientific and clinical advances in perinatology and neonatology have improved the chances of survival of preterm infants. In order to detect markers of possible developmental deficits, clinical and ethical demands have emerged regarding the early assessment of these newborns. Hence, the evaluation of the development of the extreme (i.e., born before 32 weeks) and very preterm (i.e., born before 34 weeks) newborns by the monitoring their unique behavioral communications was proven to be relevant to adapt the care and the caregiving environment [6, 7]. In addition, the continuous monitoring of sleep stages, vocal, motor, or facial activities was shown to be relevant for the detection of various neurological disorders [8–10]. Thus, nowadays, nurse observations are performed in the presence of the newborn as part of the Newborn Individualized Developmental Care and Assessment Program (NIDCAP) [11]. However, several limitations hinder the generalization of these procedures since they are very time-consuming and only a small proportion of newborns can benefit from it. Furthermore, although it is performed by specially trained nurses, these observations remain subjective.

In light of this information, it seems obvious that new solutions for monitoring neurobehavioral development could improve the care of newborns. In regard to the already very intrusive care machines, it is important to consider non-invasive monitoring methods.

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. The analysis of acoustic parameters development of infant cries might offer a non-invasive tool since these characteristics reflect the development and possibly the integrity of the central nervous system. Indeed, crying is a functional expression of basic biological needs, and emotional or psychological conditions and requires a coordinated effort of several brain regions, mainly brainstem and limbic system and is linked to the breath and the lung mechanisms. Thus, acoustic analysis of newborn infant cry appeared to be a good indicator to assess neurophysiological parameters. Moreover, being easy to perform, cheap and completely non-invasive, it can be easily applied in many circumstances.

This thesis was conducted in the context of the European Project Digi-NewB started in March 2016 which proposed a new approach of monitoring based on the acquisition of three sources (electrophysiological, clinical and audio-video data) to help clinicians in their diagnosis. During four years, seven teams worked on a decision support system proposed to gather composite indices collected from clinical data and multi-signal analysis, including heart rate, respiration rate, video, and sound signals. The two main aspects of neonatal health targeted were sepsis and neuro-behavioral maturation.

From an audio perspective, this is the first time that such a device has been implemented and so much data has been recorded. This is why the objective of this work was to develop an automatic processing chain for the detection and characterization of premature newborns spontaneous cries recorded in routine care environments. The resulting manuscript is divided in six chapters.

CHAPTER 1 - We review the basic concepts and terms related to prematurity used throughout the manuscript. Next, we present the anatomy and physiological phenomena involved in the production of crying as well as the definitions of the acoustic parameters of crying. Finally, we review the literature of different clinical and methodological studies on the topic.

CHAPTER 2 - We present the neonatal intensive care units where the recordings are performed as well as the usual cares provided to preterm newborns. Then, we describe in more details the Digi-NewB project, in which all subsequent studies presented in the thesis are framed. Finally, we describe the complex noise environment we have to face and prove the interest of our strategy.

CHAPTER 3 - We propose a segmentation step used to separate the useful sound segments containing audio information from the background noise. Originally based on the study proposed by [12], it was then improved to better process our data. In addition, we propose to use video signal processing to extract only sounds occurring in infant motion periods.

CHAPTER 4 - We propose a recurrent neural network model, which uses sound segment spectrograms to detect whether it contains crying. This method was developed with a population of 43 neonates and 21 340 sounds that were annotated by SoundAnnoT software that was designed for this.

CHAPTER 5 - We propose a new method for fundamental frequency characterization based on contouring techniques on spectrogram after an automatic frequency band of analysis detection step.

CHAPTER 6 - We deploy the complete processing chain combining the three methods described in Chapter 3, 4 and 5 and we present the clinical results computed for 57 newborns. First, we propose to compare our results with existing studies, then, we provide new approaches of data visualizations especially with longitudinal studies that have never been done before.

Finally, in Conclusion, we give some final remarks about the outcome of the research presented in this dissertation. Furthermore, we present a summary of our findings regarding the three proposed methods, as well as the strength and limitations of our study. At last, we introduce new insights regarding possible future directions to continue this line of research.

BIBLIOGRAPHY

- [1] WORLD HEALTH ORGANIZATION. Who: Recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. *Acta Obstetrica et Gynecologica Scandinavica*, vol. 56, 247–253 (1977).
- [2] SOSPREMA. La prématurité.
Accessed on 25/10/2021 from: <https://www.sosprema.com/la-prematurite/definition/>
- [3] WORLD HEALTH ORGANIZATION. Born too soon: the global action report on preterm birth (2012).
- [4] MCGUIRE W., CLERHEW L., AND FOWLIE P.W. Infection in the preterm infant. *Bmj*, vol. 329, 1277–1280 (2004).
- [5] RAMASETHU J. Prevention and treatment of neonatal nosocomial infections. *Maternal health, neonatology and perinatology*, vol. 3, 1–11 (2017).
- [6] BRAZELTON T.B. AND NUGENT J.K. *Neonatal behavioral assessment scale*. 137. Cambridge University Press (1995).
- [7] PRECHTL H. The behavioural states of the newborn infant (a review). *Brain Research*, vol. 76, 185–212 (1974).
- [8] PRECHTL H.F. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Human Development*, vol. 23, 151–8 (1990).
- [9] 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*, vol. 349, 1361–3 (1997).
- [10] 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*, vol. 132, 300–6 (1998).
- [11] ALS H., LAWTON G., DUFFY F., MCANULTY G., GIBES-GROSSMAN R., AND BLICKMAN J. Individualized developmental care for the very low-birth-weight preterm infant. medical and neurofunctional effects. *JAMA : the journal of the American Medical Association*, vol. 272, 853–8 (1994).
- [12] 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*, vol. 8, 799–810 (2013).

About infant crying

1.1 Introduction

The purpose of this chapter is to present the background and interest in the analysis of infant crying. After an overview of prematurity in terms of definitions and care in real-life conditions, we give a brief description of the infant's mechanical and neural activities responsible for the production of crying. Next, we define key terms and features used in this work. Then, through a state of the art, we review the clinical analyses of crying in term and preterm newborns, as well as the methods commonly used in acoustic signal processing. Finally, we present our strategy.

1.2 Prematurity

This section provides several definitions that will be used throughout this document. Therefore it is important to introduce these terms first.

Pregnancy is the term used to describe the period in which a fetus develops inside a woman's uterus. From a medical perspective, it is defined as the period measured from the first day of the last normal menstrual period until delivery, it is measured in weeks of amenorrhea. When lasting about 40 weeks, or 9 months, infants are considered full-term (FT). However, pregnancy can be shortened for various reasons. In the case of birth occurring before 37 weeks of gestation, it is defined by the World Health Organization as preterm birth (PT) [1].

TERMINOLOGY - In this work, we use the standard terminology proposed by the American Academy of Pediatrics [2], which we define in the following and in [Figure 1.1](#).

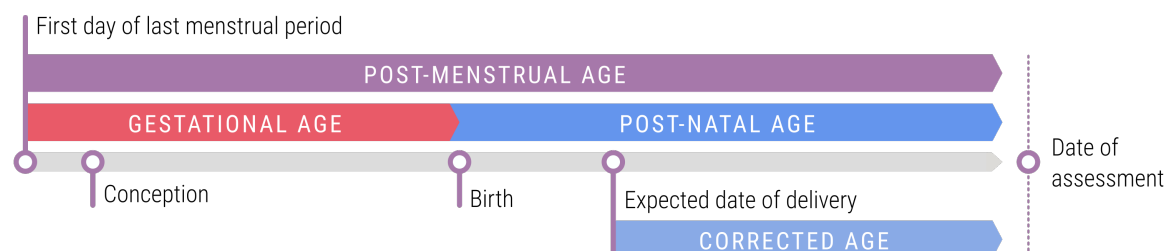


Figure 1.1: Age terminology during the perinatal period according to the American Academy of Pediatrics definitions [2].

- **Gestational Age (GA):** the duration (or term) of pregnancy measured from the first day of the last normal menstrual period until the birth date (in weeks of amenorrhea). It is a valuable definition since it proposes a fixed age to refer to and to identify premature babies independently from their current age at the assessment.
- **Post-Natal Age (PNA):** the duration elapsed since birth (in days, weeks, or months).
- **Post-Menstrual Age (PMA):** the duration between the first day of the last normal menstrual period and the date of assessment (usually in weeks + days). It can be seen as the summation of the GA and PNA.
- **Corrected age:** the duration elapsed between the expected date of birth and the date of assessment (in days, weeks, or months). In fact, this term only exists in the case of premature birth, otherwise corrected age and postnatal age are identical.

PREMATURITY BASED ON GA - According to severity, prematurity is subdivided into three categories. They are defined according to the pregnancy duration (in weeks) and represent respectively 5, 10, and 85% of the total premature births [3]:

- Extremely Preterm (EP), newborns born before 28 weeks;
- Very Preterm (VP), newborns born between 28 and 32 weeks;
- Moderate to Late Preterm (MLP) newborns born between 32 and 37 weeks;

CAUSES - A birth can be premature for many causes that can be classified into two main triggers:

- *provider-initiated*, defined as the induction of labor or elective cesarean due to maternal or fetal indications or other non-medical reasons,
- *spontaneous*, with spontaneous onset of labor or premature rupture of membranes. The main factors, in that case, are multiple pregnancies, infections, chronic maternal conditions (diabetes, hypertension, anemia, asthma, thyroid disease), nutrition, lifestyle, maternal psychological state, genetics, or even age at pregnancy.

However, in up to half of all cases the cause remains unidentified [4].

RISK OF COMPLICATIONS - Premature birth interrupts the newborn's in utero development resulting mainly in the immaturity of four essential organs: the brain and brainstem, the lungs, the digestive tract, and the ductus arteriosus [3].

Thus, preterm infants are exposed to severe cardiorespiratory events (associating apnea with bradycardia and oxygen desaturation) and to an excessive risk of unexpected sudden infant death. The more severe the prematurity, the greater the risk of health problems or sequelae. For instance, the chances of survival of an extremely preterm newborn vary greatly (i.e., between 0 and 90%), whereas an infant born after 29 weeks GA has a much better chance of survival (i.e., 95%) [5].

In addition, prematurity sequelae can also have several long-term effects on both physical and neurological developments such as visual, hearing, or learning impairments, cardiovascular or respiratory disorders, and, global developmental delay.

CARES - However, most of these vulnerabilities are resolved with a good maturation (i.e., development during hospitalization). This is why, since the 1970s, infants are cared for in specialized units called Neonatal Intensive Care Units (NICU). There, they can benefit from thermal, respiratory, and nutritional assistance. Tracking physiological conditions in the perinatal period is of utmost importance to provide the appropriate care and clinical setting to each newborn. Therefore, during their complete hospitalization, infants benefit from careful monitoring of their vital and physiological constants [6].

In addition, in some cases, human observations are performed in the presence of the newborn by trained nurses in the Newborn Individualized Developmental Care and Assessment Program (NIDCAP) [7]. This program aims to improve infant development through behavioral monitoring. Further details on NICU care and configurations are provided in **Chapter 2**.

1.3 About crying

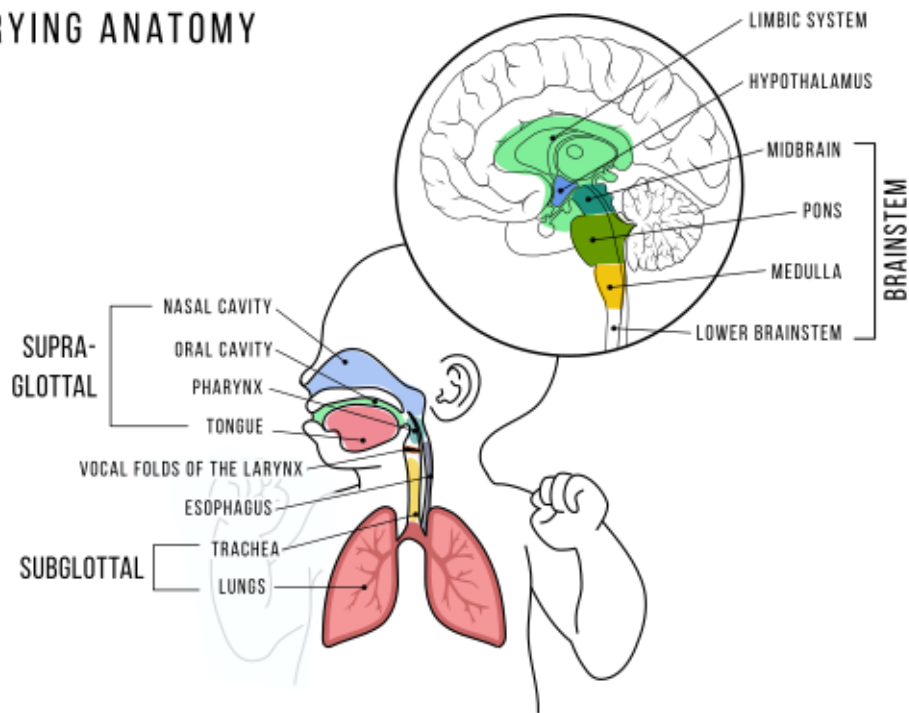
1.3.1 Anatomy and physiology

Contrary to most mammals, newborns remain dependent on adults for a while to eat, move, care, ... This is why babies produce distress signals in order to warn their caregivers. Crying is their primary mode of communication. Shortly after birth, this innate survival mechanism should not be interpreted as a demand for emotional attention but rather for someone to meet their basic needs (absence of caregivers, weariness, colic, fear, fatigue, hunger) [8, 9]. The following paragraphs address the crying mechanical and neurological production which are topics explained in more detail in [10] and [11].

Mechanical production

Breathing is the first step of crying. During the expiration, air comes out of the lungs and travels through the trachea into the larynx located in the throat (see the anatomy of cry in **Figure 1.2**). This organ, composed of the vocal folds and the glottis, is involved in the swallowing, breathing, and voice production functions. Vocal folds are muscular organs composed of two membranes that can be completely relaxed (as for free-breathing), totally blocked, or in an intermediate position (see illustration in **Figure 1.3**). In this case, the increase in air velocity in the narrow passage between the folds results in a drop in air pressure causing them to open and close rapidly. This vibration is responsible for phonation and has a fundamental frequency defined as F_0 .

BABY CRYING ANATOMY



HUMAN NERVOUS SYSTEM

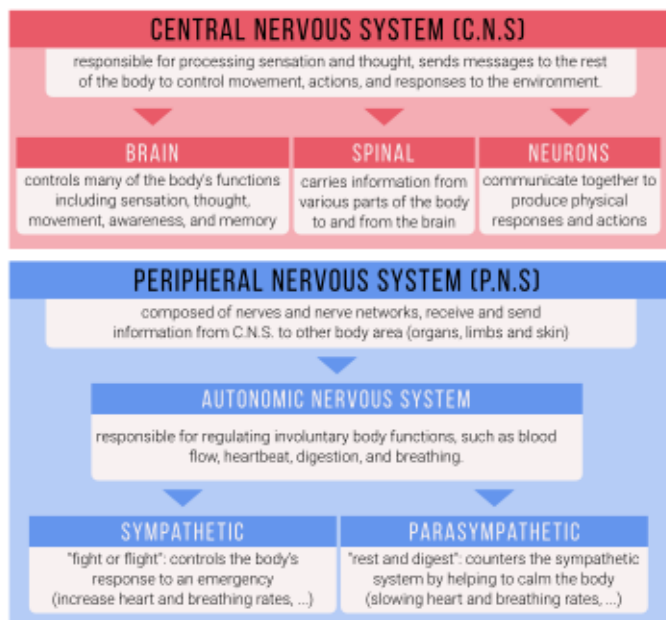
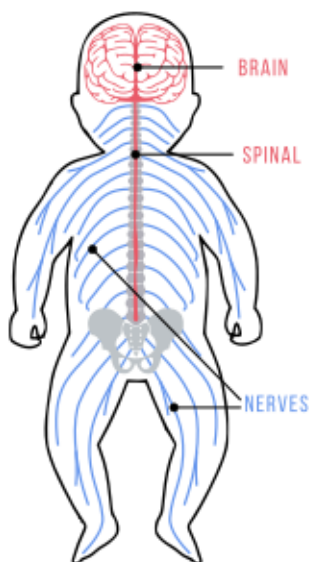


Figure 1.2: Infant crying mechanisms. Body and brain anatomy and nervous system parts responsible for crying.

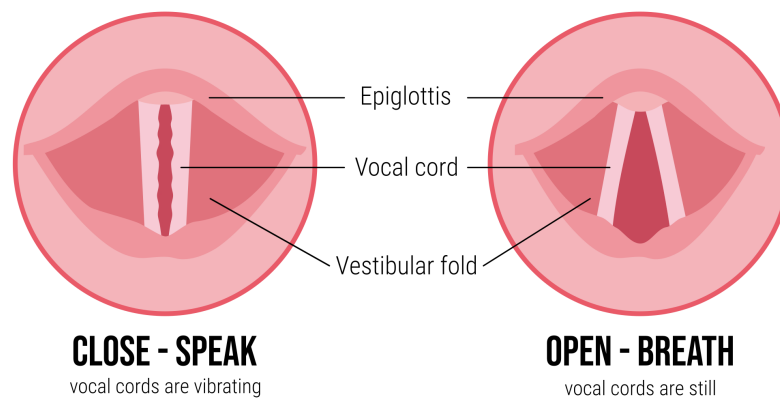


Figure 1.3: Open and close vocal chords positions.

Then the sound is shaped by the different areas it crosses. Thus, after having met the aerodigestive crossroads between the airways and the digestive tract at the level of the pharynx, it reaches the sub-glottal or vocal tract area. The latter is divided into the oral and nasal tracts and it is their size and contour that carve the sound to produce resonant frequencies or formants.

Neurological production

Neonatal crying is triggered by internal or external stimulation and is produced by the coordination of several brain regions (see [Figure 1.2](#) for part of the brain's anatomy).

While cry initiation has been associated with the limbic system, hypothalamus, and sympathetic arousal, the crying configuration is controlled by the midbrain.

Indeed, the lower brainstem controls the muscles involved in sound production through the network of neurons called the reticular activating system. It is the tension's variation of these various muscles (i.e., larynx, pharynx, chest) that is responsible for the fundamental frequency and crying modes. The brainstem also controls the size and shape of the supraglottal system (upper vocal tract) which carves formant frequencies.

Finally, a cry can occur thanks to the nervous system responsible for processing sensation and controlling movement, action, and response to the environment (see the illustration in [Figure 1.2](#)). In particular, crying is controlled by the autonomic nervous system which manages the coordination between the vagal innervation and the central nervous system.

Furthermore, modulation of the overall contour of F_0 as well as the amplitude or intensity of the cry reflects autonomic mechanisms. Thus, atypical F_0 patterns, rapid changes, or high variability suggest neural control system instability or cranial vagal nerve complex lesions (carrying information for the parasympathetic system that help to calm the body).

A more exhaustive review of the crying characteristics with the associated biological mechanisms is presented in [\[11\]](#).

1.3.2 Prosodic feature definitions

This section is intended for the definition of terms and crying prosodic features with which the reader is likely to be unfamiliar and for common terms that have a specific definition in the context of this manuscript.

- *Cry unit*: sound resulting from the passage of air through the vocal folds during a single inspiratory/expiratory cycle.
- *Cry*: total sound response, which may contain many cry units.
- *Fundamental frequency* (F_0): a physical characteristic of all periodic waveforms. It is measured in hertz (Hz) and refers to the number of times a complex waveform repeats itself in one second.
- *Pitch*: the hearing subjective tone perception of highness or lowness that depends on the number of vibrations per second produced by the vocal cords. Its unit is the Mel.
- *Cry type or mode*: identifiable acoustic output an infant can produce, based on the vibration of the vocal cords. There are three expiratory and one inspiratory modes which are illustrated in [Figure 1.4](#) and described below.

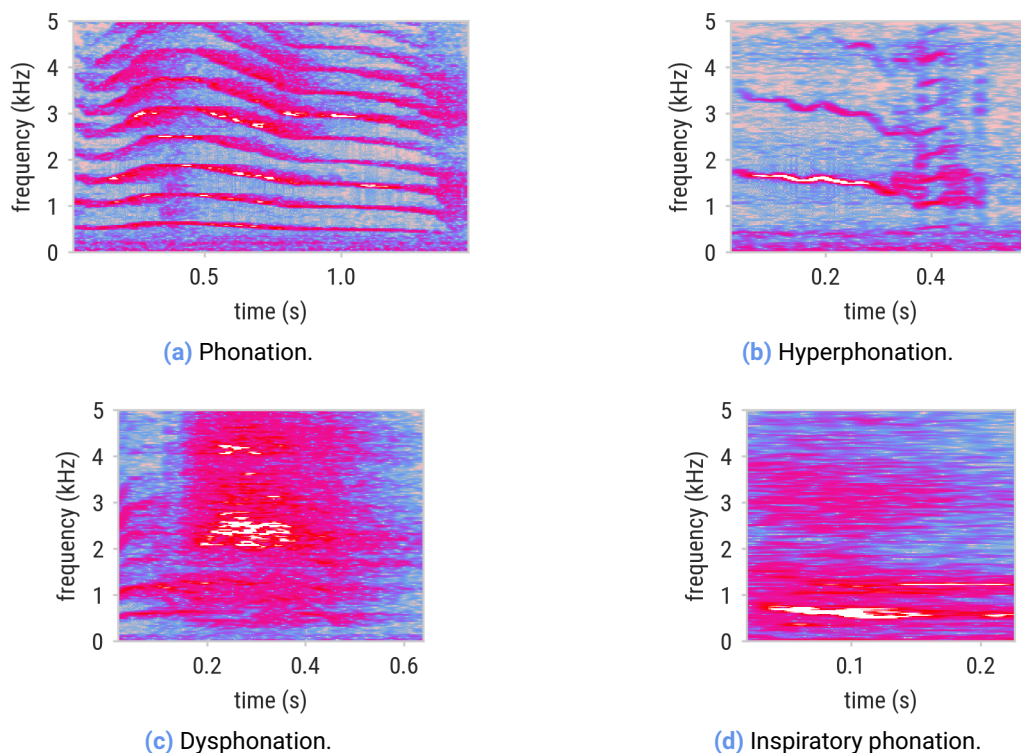


Figure 1.4: The four crying modes illustrated using spectrograms.

- *Phonation* or basic cry: resulting from periodic vocal fold vibration occurring with a F_0 between 250–750 Hz and produced thanks to neural control of muscular tension and airflow.
- *Hyperphonation* or high-pitched cry: caused by a sudden upward shift with F_0 greater than 1000 Hz due to a neural constriction of the vocal tract.
- *Dysphonation* or turbulent cry: caused by noisy or inharmonic vibration of the vocal folds due to unstable respiratory control. Such cry unit is not periodic.
- *Inspiratory phonation*: any sound produced during inspiration.

Phonation and hyperphonation cries have additional frequency characteristics related to their periodic acoustic content, we can mention the essential ones:

- *Harmonics*: multiples of the fundamental frequency. For example, if the fundamental frequency is 100 Hz, then the first harmonic would be 200 Hz, the second 300 Hz, etc.
- *Formant frequencies*: the resonance frequencies of the vocal tract. Formant frequencies are usually independent of the fundamental frequency and its harmonics. Only the first two formants are typically measured.
- *Melody*: identifiable variation of the fundamental frequency along with a cry unit.

Some of the mentioned prosodic features are depicted in [Figure 1.5](#) along a basic cry unit. The signal is represented in time, frequency and time-frequency domains. The usual basic cry formant frequencies are:

- *First formant* (F_1) between 1000 and 1500 Hz;
- *Second formant* (F_2) between 2500 and 3500 Hz.

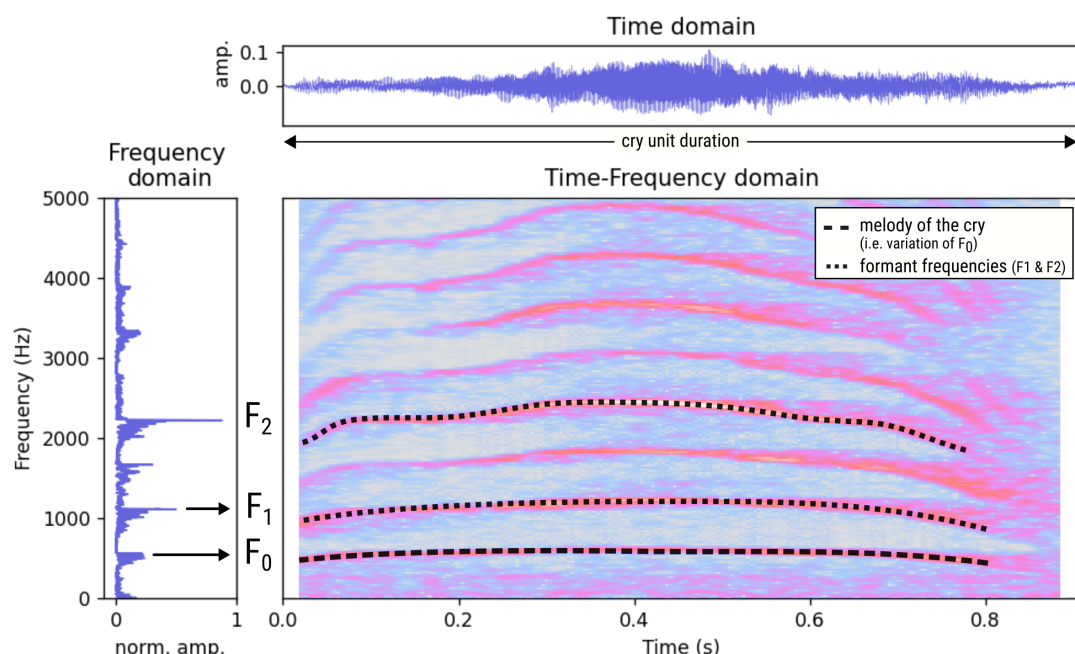


Figure 1.5: Crying features along a cry unit.

1.4 State of the art

1.4.1 Clinical investigations of crying

Research on infant crying started with auditory analysis in the 1960s thanks to the Finnish Wasz–Hockert research group when it was shown, by spectrographic analysis, that four distinct types of cries could be distinguished as birth, pain, hunger, and, pleasure [12]. Then, crying analysis was studied in newborns and small infants with good or poor health conditions, but also in premature newborns (see [13] for a historical review). From there, two other research groups largely contributed to this topic: Lester’s team in Providence (USA) and Manfredi’s team in Firenze (Italy). Led by these three groups, several studies have later shown that cry signals hold valuable information in the infant health status evaluation according to the clinical context, both, for children and full-term or premature newborns (see [6] for a review).

Full-term newborns and infants

Infant cries were studied for the differentiation between normal and pathological cries. For instance, the similarity between the cry of a malnourished infant and the cry of a brain-damaged one suggested that malnutrition might affect the regulatory function of the central nervous system [14]. Moreover, equivalent results suggested that heavy marijuana use also affects the neurophysiological integrity of the infant [15].

In [16] and [17], cries of newborns with prenatal and perinatal complications (such as low birth weight, respiratory symptoms, jaundice, apnea, ...) were detected and acoustical properties presented differences when compared to that of healthy newborns. Furthermore, a comparison was performed between normal and high-risk subjects to find possible early signs of autism [18]. Differences were seen in the fundamental frequency value, the number and the length of episodes, and in their melody.

Then, cries were evaluated either to discriminate, with facial expressions, behavioral reactions between invasive and non-invasive procedures [19] or to measure pain after a heel-prick stimulus. In the latter case, the conclusion was that crying can be used to measure pain in newborn infants only when the cause of crying is known [20].

Finally, it is worthwhile to notice that most of the previously mentioned studies were based on the analysis of pain-induced cries, which were easier to analyze because no processing to detect them was needed. Therefore, the investigation of infants’ spontaneous cries was only recently studied in several contexts, such as profound hearing loss and/or perinatal asphyxia [21–23], early detection of autistic signs [24], monitoring [25] or comprehension of vocal development and early communication [26].

Premature newborns

Characterization of crying episodes in preterm infants was also largely explored either solely or in comparison with full-term newborns where neurophysiological maturity differences were observed as well as a later impact on speech development.

Once again, early studies focused on the analysis of pain-induced cries. Although differences were shown between the cries of premature and full-term infants at the time of birth, it was also shown that as preterm newborns grew, their cries became more like those of full-term infants [27]. The same kind of conclusion was reached in a pain evaluation study, based on facial expressions and crying, when comparing newborns to 2- and 4-month-old infants [28].

As for full-term, analysis of spontaneous cries of preterm infants has been less investigated and is recent. The comparison between spontaneous cries of six premature children (three pairs of twins) recorded at different ages showed essential changes in the cries from the 8th–9th week of life up to the 23rd–24th week of life, and were interpreted as an intentional articulatory activity [29].

In a study, Orlandi et al. presented a correlation between central blood oxygenation and the distress occurring when crying [24]. For a similar decrease in oxygenation levels in both groups, results showed that, after the crying episode, full-term had a faster and more stable recovery time than preterm newborns.

Eventually, effects of gestational age, body size at the recording, and intrauterine growth retardation (IUGR) were investigated in [30]. Cries were recorded before feeding in both healthy preterm and full-term newborns at term-equivalent ages and showed that shorter gestational age was significantly associated with higher F_0 regardless of the smaller body size at recording or IUGR.

1.4.2 Methods for acoustic signal processing

Crying analysis involves three steps which are data acquisition, signal processing, and feature extraction (Figure 1.6). In this section, we review the literature on these steps to give the methodological background and thus propose our strategy.

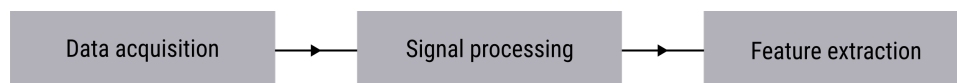


Figure 1.6: Framework of the acoustical processing chain used in cry analysis.

Crying data acquisition and databases

Collecting infant cries is a challenging task since it is difficult to create and implement an audio acquisition protocol as well as to find infants to record since it requires parental consent. In addition, crying analysis requires data annotation to assess the automatic methods. Therefore, in this paragraph, we review several procedures and databases cited in the literature.

So far, most of the cry analysis studies have been conducted on real audio signals recorded in a hospital or at home with microphones placed near infants. While recordings used to be performed occasionally (i.e., to capture crying events one by one), they are now used for long time to record all sound events, this is called monitoring.

Then, these recordings are usually split into small cry/sound signals (a few seconds) that are then annotated by the doctors, nurses, or parents. However, this task is very time-consuming and remains annotator-dependent due to the perceptual aspect and the lack of crying type definitions. Hence, every author constructed their own specific clinical annotated database.

Moreover, due to resource limitations and the sensitivity of the infant data, which have to be anonymized during the collection process, there are few available databases. A review of those is proposed in [31].

To date, the most commonly used database in cry analysis is the Baby Chillanto database with data collected by the National Institute of Astrophysics and Optical Electronics, CONACYT Mexico [22]. Initially developed by Reyes-Garcia et al., it divides the cries according to their cause and proposes five types, including pain, hunger, normal, asphyxia, and deafness. Cries are equally segmented into 1-s duration with a total number of 2268 cries recorded from infants ranging from newborns to nine months of age.

Otherwise, cohort size varied from few infants crying to 12 914 cries in [32] (collected in hospital from 127 babies) or 19 691 cries in [33] (recorded by parents using a smartphone). However, very few crying databases have been recorded in the NICU [34–38], and there is little information about the recordings and the subjects processed, which are gathered in [Table 1.1](#).

Finally, synthetic signals have also been used, and sometimes compared with a real dataset, to increase the number of processed data [36, 39–41].

Audio signal processing

Once the recordings have been made, the signal processing step consists in extracting the cries from the recorded signal. Initially called cry segmentation, because the cries were manually extracted from recordings made in quiet controlled environments, this step was then enhanced as the recordings became more complex.

Paper	Fs	Cohort	Data
[34]	16kHz	5 premature babies 28 to 34+2 GA 2 to 208 days PMA	535 audio segments cry units dur. 2 to 150 s CU tot. duration: 45 min and 55 s
[35]	44.1kHz	26 babies 3 days to 6 months PMA	48 audio segments 971 cry units
[36]	16kHz	1 baby	10 audio segments 234 cry units, tot. dur. 122.95 s
[37]	44kHz	38 babies (28 FT, 10 PT) 23+5 to 42 GA FT: 2 days PMA PT: 35+1 to 43+1 PMA	38 audio segments 6844 cry units, dur. > 0.26s (<i>extracted with Biovoice</i>)
[38]	-	more than one baby 0 to 9 months PMA	175 cry units

Table 1.1: Databases recorded in NICU.

With the development of technologies, the duration of recordings increased, with a consequent increase in the amount of data to be handled. Therefore automated processing methods emerged with the objective to extract the cries from the background noise. They are usually based on energy computation techniques derived from speech processing.

Then, more recent studies investigated spontaneous cries of infants in real-life monitoring context at home or in the NICU. This uncontrolled environment led to new issues, such as cries occurring at unknown times, as well as unpredictable sounds occurring in the recordings (e.g., voices, doors, ...).

To date, two strategies are commonly used to process such data ([Figure 1.7](#)):

- Based on cry segmentation methods, the first strategy extracts in the audio signal all sound events from background noise; then classifiers are used to detect sound segments containing cries [[42–46](#)].
- The second strategy only relies on classifiers that run through the windowed signal and detect the windows containing crying [[36, 47–51](#)].

Crying feature extraction

Feature extraction is the stage to extract discriminative components from audio signals to perform cry analysis. It can be local features extracted from short frame intervals of cry signal or global features computed over the whole cry unit. Due to the high instability of a cry, it is better to use local features to be robust enough to cover variation within the signal. Although the human voice is a subject that has been widely studied, specific cry acoustic and prosodic features need to be defined since infant and adult voice productions differ in terms of energy, intensity, and formants.

Audio signals are usually represented in time and frequency domains, and, recently several studies investigated the time-varying frequency features along with a cry unit. The characteristics derived

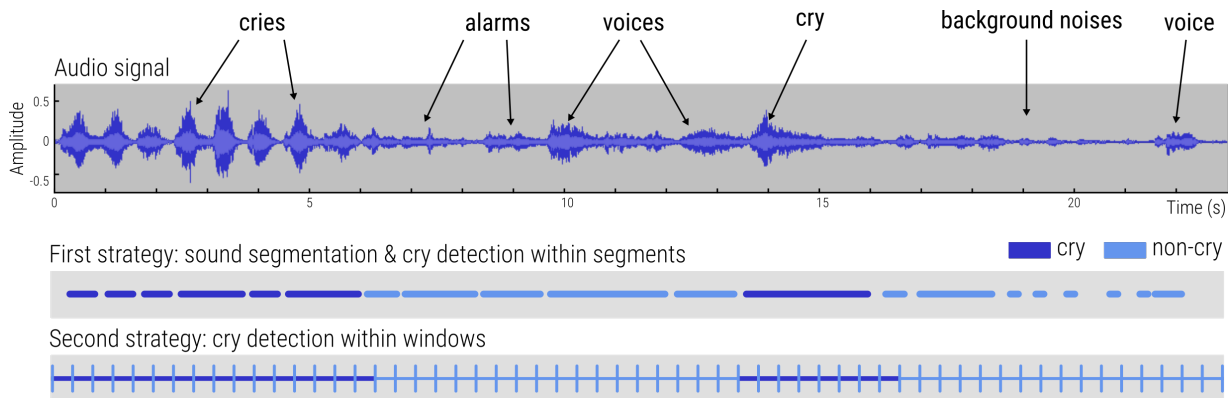


Figure 1.7: Scheme of the two strategies used to automatically extract cries from noisy recordings.

from the three domains are described hereafter.

TIME FEATURES - Duration is the most common time feature which has been investigated. It has been derived in several definitions such as cry unit duration [52] with its mean [53, 54], total cry duration (including one or more cry units) [11, 20, 53] and, the ratio of cry duration within audio signal [53, 55]. Pauses between cry units have also been examined through similar duration metrics [52, 53].

Another commonly used parameter is the latency time, described as the time from known stimulus [11] or pain stimulus [19, 20, 56, 57] (in case of pain-induced cry) to the first cry. In addition intensity, zero-crossing rate, amplitude, and energy-based features have also been proposed in [11, 17].

However, even if time-domain features are easy and straightforward to compute, they are not robust enough to cover the variations within infant cry signals because of their sensitivity to background noises.

FREQUENCY FEATURES - On the contrary, frequency-domain features have a strong ability to model the characteristics within infant cry signals. The spectral energy features have been computed through different approaches such as the overall spectral energy of the signal [28, 58] or the energy only induced by low or high frequencies [55].

Yet, the most relevant clinical parameter to date is the fundamental frequency (F_0) which was investigated in virtually all the mentioned works. Indeed, as explained before, this prosodic feature offers a direct measurement of vocal development since it corresponds to the rate of glottal opening and closing in the vocal tract (see [Section 1.3.1](#)). The fundamental frequency is usually studied through statistical parameters calculated on several cries, such as mean [25, 57], maximum and minimum [30], standard deviation [37] or variation coefficient [54]. Moreover, resonance frequencies were usually investigated through the first two formants F_1 and F_2 in [25, 29, 54], but some authors also proposed to assess the third one (F_3) [37, 54, 59].

Moreover, the common and well-known acoustic features Mel-frequency cepstral coefficients (MFCCs) [48, 60–62] and Linear Prediction Cepstral Coefficients (LPCCs) [23] have proven to be efficient to detect cry within the signal [47]. MFCCs are obtained through a signal projection on the Mel-scale with frequency bands equally spaced inspired by the human auditory system, whereas LPCCs are based on the vocal tract modelization.

TIME-VARYING FREQUENCY FEATURES - Actually, due to the highly non-stationary cry signal characteristics, it is better to represent the energy contents of a signal in a joint time-frequency domain. In practice, it means that frequency features are extracted locally (from short frame intervals of the cry signal) and displayed with respect to time.

Thanks to this representation, one can see the melodic shape of a cry which describes the pattern of F_0 as it varies with time (see Figure 1.5). It is the most common time-varying frequency descriptor and four main melodic shapes were firstly defined in [63]: falling, rising, falling–rising (or rising–falling) and flat.

In [64], Várallyay reduced these shapes to three fundamental units (i.e., falling, rising, and flat) that were further used as the basis for the definition of 77 melodic shapes. Later, the “complex” shape was introduced in [40, 41, 65] to cover all melodic patterns composed of more than two fundamental units. Nowadays, six basic melodic forms of crying are retained and are illustrated schematically in Figure 1.8.

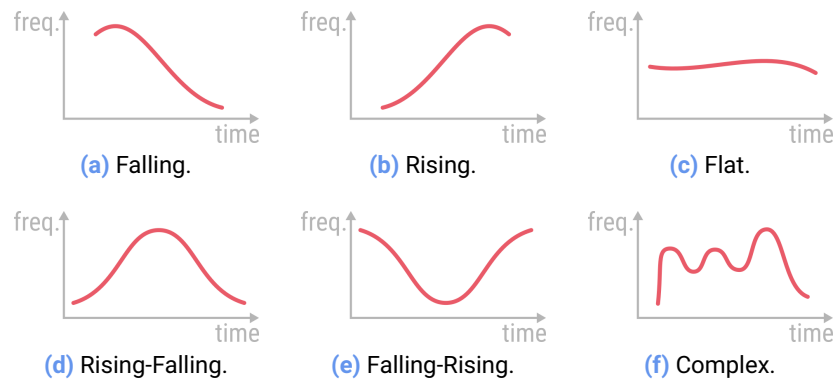


Figure 1.8: Basic and schematic melody shapes of infant cries in the time-frequency plane.

In addition, several other features were defined to assess variations in F_0 along a cry unit or during a cry event (succession of cry units). To mention a few, there is jitter (cycle-to-cycle variations of F_0) [59, 66, 67], shift (sudden change in F_0) [11, 20, 57] and glide (rapid variations in F_0) [20, 57, 68].

1.5 Our strategy

In this chapter, we addressed the subject of prematurity as well as the relevance of acoustic monitoring in the infant's maturation assessment. We showed that the coordination between mechanical and neurological systems is necessary for cry production. Moreover, we saw that the crying analysis has been widely reported in the literature in both term and preterm infants. However, spontaneous crying analysis, especially recorded in NICU, is quite recent and we reported a few studies.

The lack of spontaneous cry analysis is due to several major obstacles related to data. First, we mentioned the sensitivity of human data, where anonymization is crucial, making it difficult to create large databases. Secondly, the long recordings with unpredictable cry onsets led to the use of automated methods. Finally, recording real data in a clinical environment, such as the NICU, remains a real challenge and requires robust signal processing since transitory random noises can occur in the signal (doors, voices, machines, ...).

Although some teams have already proposed methods [25, 34, 36, 50, 69], to date and to our knowledge, no one has realized a continuous processing chain for long recordings performed in such a noisy environment. Therefore, the thesis's objective is to address this topic by proposing a workflow for automated cry analysis and the evaluation of maturation from long recordings made in the NICU.

Due to the many issues mentioned above, we developed a three-step strategy, detailed step by step in the following chapters, illustrated in [Figure 1.9](#) and briefly described hereafter:

1. sound segment extraction;
2. cry detection among the extracted sound segments;
3. fundamental frequency characterization.

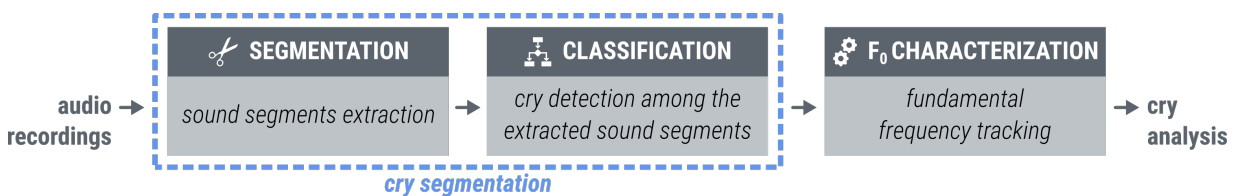


Figure 1.9: Workflow of the proposed cry analysis automated processing chain.

In order to achieve efficient automatic sound processing, it is essential to have a good knowledge of the acoustic environment. Therefore, in the following chapter, we describe the intensive care given in NICU through a review of the assistance and monitoring generally performed during sick or premature infants' hospitalization. Then, we present the European project Digi-NewB in which the acquisition system and the data collection protocols were designed. Finally, we present the acoustic environment that is valuable to understanding our choice of processing strategy.

BIBLIOGRAPHY

- [1] WORLD HEALTH ORGANIZATION. Who: Recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. *Acta Obstetricia et Gynecologica Scandinavica*, vol. 56, 247–253 (1977).
- [2] AMERICAN ACADEMY OF PEDIATRICS. Age terminology during the perinatal period. *Pediatrics*, vol. 114, 1362–1364 (2004).
- [3] INSERM. Prématurité, ces bébés qui arrivent trop tôt (2015). Accessed on 25/10/2021 from: <https://www.inserm.fr/dossier/prematurite/>
- [4] WORLD HEALTH ORGANIZATION. Born too soon: the global action report on preterm birth (2012).
- [5] 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*, vol. 358, 1672–1681 (2008).
- [6] CABON S., PORÉE F., SIMON A., ROSEC O., PLADYS P., AND CARRAULT G. Video and audio processing in paediatrics: A review. *Physiological Measurement*, vol. 40 (2019).
- [7] ALS H., LAWHON G., DUFFY F., MCANULTY G., GIBES-GROSSMAN R., AND BLICKMAN J. Individualized developmental care for the very low-birth-weight preterm infant. medical and neurofunctional effects. *JAMA : the journal of the American Medical Association*, vol. 272, 853–8 (1994).
- [8] LESTER B. AND LAGASSE L. Crying. In *Social and Emotional Development in Infancy and Early Childhood*, 80–90. Elsevier (2009).
- [9] ROTHGÄNGER H. Analysis of the sounds of the child in the first year of age and a comparison to the language. *Early Human Development*, vol. 75, 55–69 (2003).
- [10] GOLUB H.L. AND CORWIN M.J. A *Physioacoustic Model of the Infant Cry*, In B.M. Lester and C.F. Zachariah Boukydis, editors, *Infant Crying: Theoretical and Research Perspectives*, 59–82. Springer US, Boston, MA (1985).
- [11] 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*, vol. 11, 83–93 (2005).
- [12] WASZ-HÖCKERT O., PARTANEN T., VUORENKOSKI V., MICHELSSON K., AND VALANNE E. The identification of some specific meanings in infant vocalization. *Experientia*, vol. 20, 154–154 (1964).
- [13] WASZ-HÖCKERT O., MICHELSSON K., AND LIND J. Twenty-five years of Scandinavian cry research. In *Infant Crying*, 83–104. Springer (1985).
- [14] LESTER B.M. Spectrum analysis of the cry sounds of well-nourished and malnourished infants. *Child Development*, 237–241 (1976).
- [15] LESTER B.M. AND DREHER M. Effects of marijuana use during pregnancy on newborn cry. *Child Development*, 765–771 (1989).
- [16] ZESKIND P.S. AND LESTER B.M. Acoustic features and auditory perceptions of the cries of newborns with prenatal and perinatal complications. *Child Development*, 580–589 (1978).
- [17] GOLUB H.L. AND CORWIN M.J. Infant cry: A clue to diagnosis. *Pediatrics*, vol. 69, 197–201 (1982).

- [18] 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*, 2953–2956. IEEE (2012).
- [19] GRUNAU R.V., JOHNSTON C.C., AND CRAIG K.D. Neonatal facial and cry responses to invasive and non-invasive procedures. *Pain*, vol. 42, 295–305 (1990).
- [20] RUNEFORS P., ARNBJÖRNSSON E., ELANDER G., AND MICHELSSON K. Newborn infants' cry after heel-prick: Analysis with sound spectrogram. *Acta Paediatrica*, vol. 89, 68–72 (2000).
- [21] REYES-GALAVIZ O.F. AND REYES-GARCÍA C.A. Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system. In *Mexican International Conference on Artificial Intelligence*, 949–958. Springer (2005).
- [22] 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*, 408–415. Springer (2004).
- [23] WAHID N., SAAD P., AND HARIHARAN M. Automatic infant cry pattern classification for a multiclass problem. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 8, 45–52 (2016).
- [24] ORLANDI S., BOCCHI L., DONZELLI G., AND MANFREDI C. Central blood oxygen saturation vs crying in preterm newborns. *Biomedical Signal Processing and Control*, vol. 7, 88–92 (2012).
- [25] 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*, vol. 20, 85–99 (2015).
- [26] 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*, vol. 26, 321–333 (1993).
- [27] MICHELSSON K., JÄRVENPÄÄ A., AND RINNE A. Sound spectrographic analysis of pain cry in preterm infants. *Early Human Development*, vol. 8, 141–149 (1983).
- [28] 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*, vol. 52, 201–208 (1993).
- [29] WERMKE K., MENDE W., MANFREDI C., AND BRUSCAGLIONI P. Developmental aspects of infant's cry melody and formants. *Medical Engineering & Physics*, vol. 24, 501–14 (2002).
- [30] 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*, vol. 10 (2014).
- [31] JI C., MUDIYANSELAGE T.B., GAO Y., AND PAN Y. A review of infant cry analysis and classification. *EURASIP Journal on Audio, Speech, and Music Processing*, art. 8, 1687–4722 (2021).
- [32] BGNICG I.A., CUCU H., BUZO A., BURILEANU D., AND BURILEANU C. Baby cry recognition in real-world conditions. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, 315–318 (2016).
- [33] CHANG C.Y. AND TSAI L.Y. A CNN-based method for infant cry detection and recognition. In L. Barolli, M. Takizawa, F. Xhafa, and T. Enokido, editors, *Web, Artificial Intelligence and Network Applications*, 786–792. Springer International Publishing (2019).
- [34] SEVERINI M., FERRETTI D., PRINCIPI E., AND SQUARTINI S. Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation. *IEEE Access*, vol. 7, 51982–51993 (2019).

- [35] LIU L., LI W., WU X., AND ZHOU B.X. Infant cry language analysis and recognition: an experimental approach. *IEEE/CAA Journal of Automatica Sinica*, vol. 6, 778–788 (2019).
- [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 *26th European Signal Processing Conference, EUSIPCO 2018*. European Signal Processing Conference, EUSIPCO (2018).
- [37] ORLANDI S., REYES GARCIA C.A., 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*, vol. 30, 656–663 (2016).
- [38] BHAGATPATIL M.V. AND SARDAR V. An automatic infant's cry detection using linear frequency cepstrum coefficients (lfcc). *International Journal of Scientific & Engineering Research*, vol. 5, 1379–1383 (2014).
- [39] 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*, vol. 8, 799–810 (2013).
- [40] ORLANDI S., BANDINI A., FIASCHI F., AND MANFREDI C. Testing software tools for newborn cry analysis using synthetic signals. *Biomedical Signal Processing and Control*, vol. 37, 16–22 (2017).
- [41] 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*, vol. 45, 174–181 (2018).
- [42] ZABIDI A., MANSOR W., KHUAN L.Y., SAHAK R., AND RAHMAN F. Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In *2009 5th International Colloquium on Signal Processing & Its Applications*, 204–208. IEEE (2009).
- [43] COHEN R. AND LAVNER Y. Infant cry analysis and detection. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, 1–5. IEEE (2012).
- [44] VÁRALLYAY JR G. Z.B. AND ILLÉNY. A. Automatic infant cry detection. page 11–14 (2009).
- [45] 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*, vol. 142, 1318–1331 (2017).
- [46] XIE J., LONG X., OTTE R., AND SHAN C. Convolutional neural networks for audio-based continuous infant cry monitoring at home. *IEEE Sensors Journal*, vol. 21, 27 710–27 717 (2021).
- [47] 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*, 1–5. IEEE (2016).
- [48] TORRES R., BATTAGLINO D., AND LEPAULOUX L. Baby cry sound detection: A comparison of hand crafted features and deep learning approach. In G. Boracchi, L. Iliadis, C. Jayne, and A. Likas, editors, *Engineering Applications of Neural Networks*, 168–179. Springer International Publishing (2017).
- [49] COHEN R., RUINSKIY D., ZICKFELD J., IJZERMAN H., LAVNER YIZHAR" E.W., AND CHEN S.M. *Baby Cry Detection: Deep Learning and Classical Approaches*, In *Development and Analysis of Deep Learning Architectures*, 171–196. Springer International Publishing, Cham (2020).
- [50] SEVERINI M., PRINCIPI E., CORNELL S., GABRIELLI L., AND SQUARTINI S. Who cried when: Infant cry diarization with dilated fully-convolutional neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (2020).
- [51] 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*, vol. 2018, 1–14 (2018).

- [52] ANDRÉ V., DURIER V., HENRY S., NASSUR F., SIZUN J., HAUSBERGER M., AND LEMASSON A. The vocal repertoire of preterm infants: Characteristics and possible applications. *Infant Behavior and Development*, vol. 60, page 101463 (2020).
- [53] VÁRALLYAY G. Future prospects of the application of the infant cry in the medicine. *Periodica Polytechnica Electrical Engineering*, vol. 50, 47–62 (2006).
- [54] 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*, vol. 83, 204–11 (1994).
- [55] 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*, vol. 42, 850–61 (1999).
- [56] STEVENS B.J., JOHNSTON C.C., AND HORTON L. Factors that influence the behavioral pain responses of premature infants. *Pain*, vol. 59, 101–9 (1994).
- [57] MICHELSSON K. AND MICHELSSON O. Phonation in the newborn, infant cry. *International Journal of Pediatric Otorhinolaryngology*, vol. 49 Suppl 1, S297–301 (1999).
- [58] 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*, vol. 56, 1416–1428 (2013).
- [59] MANFREDI C., BOCCHI L., ORLANDI S., SPACCATERRA L., AND DONZELLI G.P. High-resolution cry analysis in preterm newborn infants. *Medical Engineering & Physics*, vol. 31, 528–32 (2009).
- [60] GARCIA J. AND GARCIA C.R. Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 4, 3140–3145 vol.4 (2003).
- [61] ABOU-ABBAS L., TADJ C., GARGOUR C., AND MONTAZERI L. Expiratory and inspiratory cries detection using different signals' decomposition techniques. *Journal of Voice*, vol. 31, 259–e13 (2017).
- [62] XIE J. *Baby cry detection based on audio signals using deep neural networks*. Master's thesis, Eindhoven University of Technology, Eindhoven, Netherlands (2019).
- [63] 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*, vol. 38, 1–11 (1996).
- [64] VÁRALLYAY G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, vol. 71, 1699–1708 (2007).
- [65] WERMKE K. AND MENDE W. Musical elements in human infants' cries: In the beginning is the melody. *Musicae Scientiae*, vol. 13, 151–175 (2009).
- [66] FULLER B.F. AND HORII Y. Differences in fundamental frequency, jitter, and shimmer among four types of infant vocalizations. *Journal of Communication Disorders*, vol. 19, 441–447 (1986).
- [67] MORELLI M.S., ORLANDI S., AND MANFREDI C. Biovoice: A multipurpose tool for voice analysis. *Biomedical Signal Processing and Control*, vol. 64, page 102302 (2021).
- [68] KHEDDACHE Y. AND TADJ C. Identification of diseases in newborns using advanced acoustic features of cry signals. *Biomedical Signal Processing and Control*, vol. 50, 35–44 (2019).
- [69] 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*, vol. 39, 390–395 (2018).

2.1 Introduction

Developing an automated cry analysis processing chain from audio monitoring requires the knowledge of the neonatal intensive care unit. These units, designed to provide specialized medical care for premature newborns or sick infants, have different configurations that depend on the newborn's developmental status. Therefore, in this chapter, we discuss the medical equipment available in NICU and its usefulness during the infant's maturation process. Then, we present the audio-video acquisition system designed for the newborns' contactless monitoring as well as the database created thanks to the European project Digi-NewB. Finally, we address the subject of the acoustic environment in NICU that makes the data automatic processing particularly challenging. After identifying the sound sources heard in the bedrooms, we present two analyses performed on manually annotated recordings. The first one is the quantification of voices, and alarms in a 15-hour recording performed on a premature baby staying in an incubator. The second one shows the sound content's great variability within the recordings. This context is essential for understanding the strategies and methods developed in this work.

2.2 Description of the NICU

Formerly, mothers used to give birth and care for their infants at home without any medical assistance. Physicians started to take an interest in providing care to reduce mortality due to prematurity during the 1880s, especially with the creation of the first incubator by Dr. Stephane Tarnier at Paris's Maternité Hospital (illustration in [Figure 2.1](#)). This new technology aimed to prevent many premature newborns from succumbing to hypothermia (low body temperature). However, at this point caring for premature babies was expensive and, many thought, pointless.



Figure 2.1: Tarnier's incubators in the Maternité Hospital, Paris, 1884. Source: Illustrated London News, 8 March 1884, p. 228.

Thus, despite the technological progress brought by the Industrial Revolution, it was necessary to wait for Dr. Hess's intervention in 1920 so that these techniques are finally recognized. With his chief nurse's help, Evelyn Lundeen, he initiated the establishment of trained nurse teams following specific protocols. Finally, it was in the 1970s that neonatal intensive care units flourished and became hospitals' integral part of the developed world, which helped to drastically decrease neonatal mortality (see [1] for a historical review).

Nowadays these units are designed and used to provide specialized medical care for sick and premature newborns to ensure their development. During hospitalization, newborns go through different configurations depending on their physiological state. To be discharged home, infants must meet the following criteria:

- be thermally independent to maintain the body at a normal temperature;
- have self-sufficient respiratory control;
- be able to feed by mouth to support appropriate growth.

When infants do not yet meet these criteria, and according to their disabilities and development, they may benefit from thermal, respiratory, and nutritional assistance in different bedroom configurations. Depending on their functional immaturity, at birth but also throughout the hospitalization, they join the care process at the appropriate step. These steps are detailed below and then the physiological and neuro-behavioral monitoring in NICU are presented.

THERMAL ASSISTANCE - A baby born extremely premature (i.e., before 28 GA) is not ready to face the extra-uterine life. Her or his thermal system is not able to regulate body temperature properly and her or his skin is still too thin to ensure its protective function. Therefore, at this stage, the baby is always placed in an incubator in the NICU for several weeks. An incubator is a bed enclosed by a plastic shield in which the environment is controlled to keep the baby at the right temperature and humidity level (see [Figure 2.2 - 1](#)). In addition to avoiding hypothermia and dehydration, it minimizes exposure to germs and external noise. The use and parameters of an incubator depend on each infant's specific needs. Thermal regulation is based either on the temperature measured on the baby's skin (large variations) or a configured targeted ambient temperature. Then, according to the thermal regulation capacity of the baby, he or she can be transferred to different environments, from a radiant warmer (see [Figure 2.2 - 2](#)) to a cradle (i.e., without thermal regulation, see [Figure 2.2 - 3](#)).



Figure 2.2: Three examples of beds in NICU.

RESPIRATORY ASSISTANCE - Several respiratory support techniques are used to meet the oxygen needs. Very premature infants may have respiratory distress, such as apnea or bradycardia (slow heart rate), that requires immediate intervention. For the most dependent, an invasive procedure such as intubation supplemented by a ventilator assistance device may be used. Then, depending on respiratory autonomy level, intubation is gradually replaced by less invasive devices such as nasal masks or cannulas (see [Figure 2.3](#)). Throughout the hospitalization, clinicians daily assess the newborn's needs by various means, such as evolution analysis of respiratory distress or blood tests (twice a day).

NUTRITIONAL ASSISTANCE A similar strategy is applied to feeding. Initially, premature infants are not able to digest food. They are therefore fed (i.e., given essential nutrients) through a central venous catheter, which is connected to the heart through the arm or leg. Next, thanks to a naso-gastric tube (i.e., a tube connecting the nose to the stomach, see [Figure 2.4](#)), they are fed with very small milk quantities. As they develop, infants consume more milk and less infusion. Then, when they are sufficiently developed and after they began to suckle, food administration is gradually replaced with breastfeeding. In general, the feeding capacity remains the last step before discharge.



Figure 2.3: Non-invasive respiratory assistance: air (enriched or not with oxygen) is delivered to the infant's airway through a nasal mask. The mask is attached to the nose/face with headgear.



Figure 2.4: The naso-gastric tube is connecting the nose to the stomach. Milk is delivered by automatic syringes.

A journey example of an extremely preterm infant who was cared for at the Rennes University Hospital (i.e., Rennes CHU) is illustrated in [Figure 2.5](#). This baby was born at 27+6 GA with a weight of only 930 grams, and first spent 47 days in the intensive care department during which he became more independent. Then, he finished his full development in the neonatology department and was discharged at 38+5 PMA with a weight of 2.590 kg. This picture chronologically depicts the required medical equipment evolution, as well as the interactions with the parents.

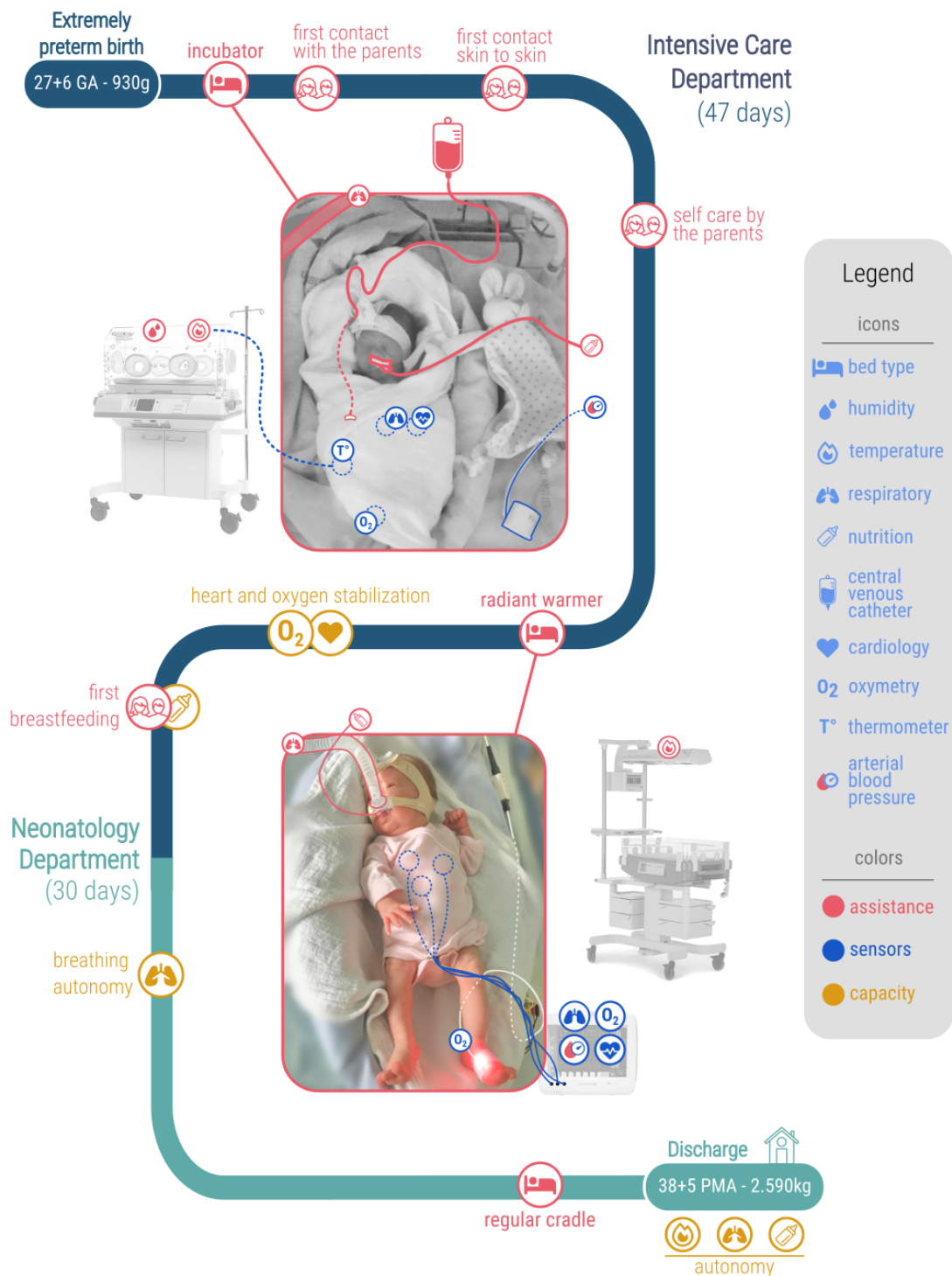


Figure 2.5: The journey of an extremely premature baby in the NICU. **TOP** - In an incubator with feeding assistance, i.e., a naso-gastric tube (coming out of the nose) and a venous catheter (with cable surrounding the head) connected to an automatic syringe. The cardiac, respiratory, and pulse oximeter sensors are invisible but present and the respiratory assistance was removed at the time of the picture. The arterial blood pressure armband is visible in the bottom right corner. **BOTTOM** - In a radiant warmer with non-invasive ventilation and naso-gastric tube. The pulse oximeter is visible with the cardiac and respiratory sensor wires.

PHYSIOLOGICAL MONITORING - The newborn is monitored continuously throughout the hospitalization. This monitoring is essential to understand the infant's needs and to provide the best care. Cardiac, respiratory, oxygen, and temperature sensors are used to collect physiological information. A review of the physiological signals advanced analyses performed in NICU is given in [2].

As mentioned earlier, the newborn's skin temperature can be measured to regulate the temperature of the incubator or the radiant warmer. Heart and respiratory rates are continuously acquired through electrodes set on the infant's body, while arterial blood pressure is measured with an armband. In case of cardiorespiratory distress, alarms are triggered for nurses who can perform immediate interventions.

Non-invasive pulse oximetry is also used to measure blood oxygen saturation and pulse rate using a photodetector. It is used in most neonatal intensive care units as a detector of de-saturation (sudden loss of oxygen in the blood).

These sensors are kept for most newborns throughout their hospitalization. An example of a radiant warmer infant bedroom configuration is shown and described in [Figure 2.6](#). Sometimes at the end of the stay, when the infant goes into the parents' arms or in the parents' presence, the sensor's wires can be disconnected from the scope.



Figure 2.6: Intensive care medical equipment at the Rennes CHU for a radiant warmer from two different points of view. LEFT - Bedroom global equipment overview. RIGHT - Zoom on the crib. CENTER - Physiological signal representations on a scope.

NEURO-BEHAVIORAL MONITORING - Unlike physiological monitoring, neuro-behavioral monitoring is performed on a more punctual basis at the doctors' request. It is performed to evaluate possible brain lesions, and it is mainly based on sleep analysis. Since premature babies have low electrical activity in the brain, the neuro-behavioral monitoring through electroencephalography can only be achieved after a good development. To perform such a measure, an ambulatory system is deployed in the newborn's room and electrodes are placed on her/his head to acquire the signal.

Moreover, in some cases, human observations are performed in the presence of the newborn as part of the Newborn Individualized Developmental Care and Assessment Program (NIDCAP) [3]. Since it has proven to be relevant to adapting the care and the caregiving environment, this program aims to evaluate the development of the extreme and very preterm newborns by monitoring their unique behavioral communications [4, 5].

In practice, a one-hour examination is performed by a trained nurse who visually annotates, within 2-minute steps, several components such as sleep stages, vocal, motor, or facial activities. These components have also been shown to be relevant for the detection of various neurological disorders [6–8].

However, several limitations hinder the generalization of these procedures. Indeed, this operation is very time-consuming and only a small proportion of newborns can benefit from this monitoring. Furthermore, although it is performed by specially trained nurses, these observations remain subjective.

2.3 The Digi-NewB proposal

The Digi-NewB project aimed to improve neonatal care thanks to the development of a new generation of non-invasive monitoring systems in neonatology. Particularly, this setup was intended to assist clinicians in the early sepsis diagnosis and the analysis of newborns' cardiorespiratory and neuro-behavioral maturation. On a larger scale, the main purpose was to decrease the mortality, and morbidity rate, reduce the risk of neurodevelopmental disorders as well as diminish the hospitalization health cost and duration. To fulfill these objectives, the French clinical network Hôpitaux Universitaires du Grand Ouest (HUGO), and our laboratory (LTSI-INSERM) collaborated to recruit five partners from four European countries composed of two companies and four university groups with multidisciplinary expertise. The Figure 2.7 shows the key figures of the Digi-NewB project, which received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 689260 and was carried out between March 2016 and May 2020. The aim was to propose a decision support system gathering composite indices collected from clinical data and multi-signal analysis, including heart rate, respiration rate, video, and sound signals. Audio and video were chosen, on the one hand, because they have proven their relevance in the evaluation of the problems addressed by the European project (see [9] for a

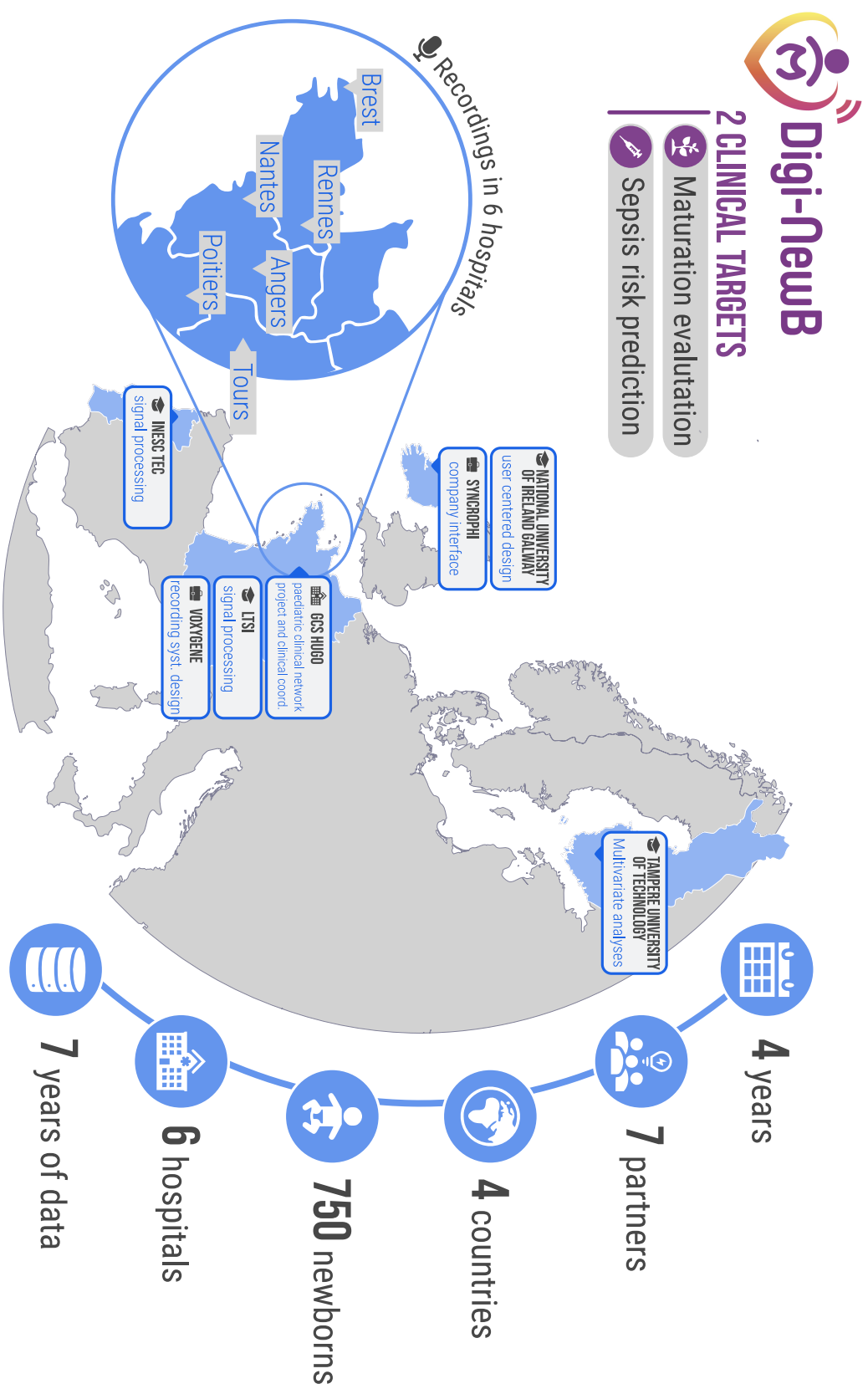


Figure 2.7: Overview of the Digi-NewB project with keys, tasks, and locations.

review), on the other hand, because their non-invasive acquisition does not disturb the medical staff or the baby by being contactless. However, unlike traditional signals that can be received from the machines already present in the hospital rooms, audio and video signals require the design of a specific acquisition system with protocols.

ACQUISITION SYSTEM - The proposed audio-video acquisition system designed for the project was developed by the Voxygen company in collaboration with Feichter Electronics and is composed of two devices (see [Figure 2.8](#)) with a total of 4 video streams recorded at a rate of 25 frames per seconds (1 colored, 1 thermal, and 2 black and white cameras, see [10] for more details) and 2 audio channels. Regarding acoustics, omnidirectional microphones (FG-23329-P07) were chosen from Knowles Acoustics, with recordings made at a sampling rate of 24 kHz.

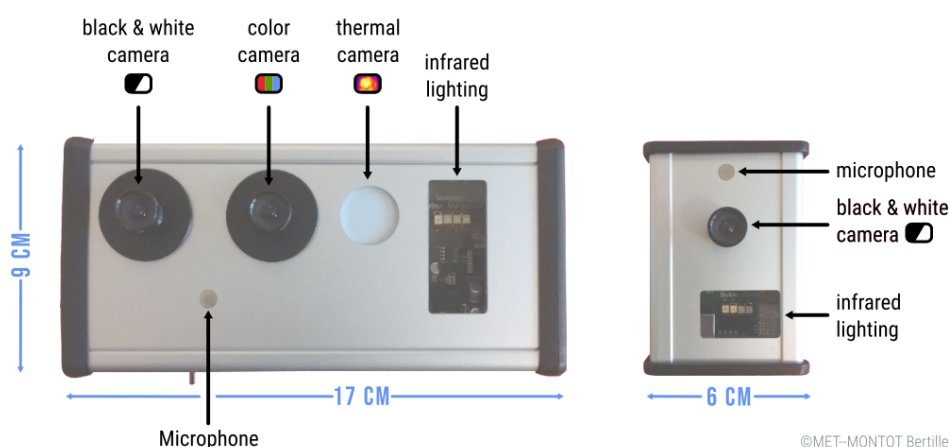


Figure 2.8: Digi-NewB acquisition device components.

DATA - Audio and video data were stored independently in 30-minute files, respectively in "WAV" and "MP4" formats. It is worthwhile to remind that the prototype of the recording device was created during the initial phase of the project and some problems were encountered on the first recordings. Thus, despite the availability of two microphones, only one of the two channels is used in this work. In addition, several video modalities were explored at first to select the best ones for the sepsis and maturation purposes.

RECORDING PROTOCOL IN NICU - Concerning the recording protocol, nurses from the six partner hospitals received training to place the devices on either side of the infants. Due to the room layout diversity and the available equipment, it has not been possible to establish a strict protocol for the system position and distance. Moreover, microphones were placed differently depending on the type of bed: in a closed bed, they were placed inside the incubator at the newborn's feet, while in an open bed, they were set near the head in a perimeter ranging from 30 cm to 1 meter. Examples of the Digi-NewB devices used to collect data in real context are shown in [Figure 2.9](#) that presents the system installed in the NICU at the Rennes CHU.



(a) Incubator.



(b) Radiant warmer.

Figure 2.9: Digi-NewB data acquisition system in real-life settings in NICU.

- ① Digi-NewB main recording device, including microphone and video cameras,
- ② auxiliary microphone and camera,
- ③ scope, monitoring the physiological signals (heart rate, respiration and oxygen saturation),
- ④ Digi-NewB computer, handling acquisition and monitoring systems.

In the pictures, babies are connected to the traditional physiological monitoring systems while cameras and microphones are set up to record the infants' movement and sound without any contact. This continuous real-time monitoring has the advantage of not affecting the infants' environment in the NICU, which could be detrimental to their maturation, neither imposing additional difficulties for health care staff or parents to interact with the newborns.

COHORT - In this work, we focused on the second objective of the Digi-NewB project, i.e., the evaluation of maturation therefore we assess only healthy infants. The protocols established for these newborns consisted in recording for several consecutive days, between birth and the date of central line removal, and then every 10 days for approximately 24 hours until discharge (**Figure 2.10**).

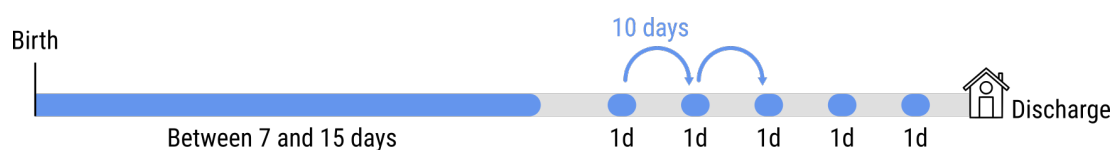


Figure 2.10: Digi-NewB maturation recording protocol.

Although during the Digi-NewB project 750 babies were recorded, only a small cohort is analyzed in this work. This is because a careful and time-consuming selection was made by the medical team to identify healthy newborns who had no complications during their entire hospitalization. As a result, a base of 57 healthy babies including 24 girls and 33 boys born between 27+1 and 41+6 GA recorded between 27+5 and 42+1 PMA were involved in this thesis work.

2.4 Acoustic environment in the NICU

To ensure the proper newborn development, the NICU environment should be similar as far as possible to what it would have been in an intrauterine pregnancy. However, as mentioned before, a lot of medical devices are needed to take care of the newborn's health. Alarms are activated when the infants' states are unstable (i.e., cardiac, respiratory distress, ...) or when the machines require human intervention (i.e., warnings for empty syringes, empty ventilation water tanks, missing equipment connection, ...). In addition, several adults are also present around the baby, mainly the medical staff and the parents. As a consequence, the acoustic environment surrounding newborns in the NICU is quite noisy and has already been investigated [11–13].

Hence, the sound environment contains many disturbing noises of short duration and at irregular intervals that deeply corrupt the audio recordings. Furthermore, several of these sound sources are overlapping in time, which makes the automatic processing of recordings in the NICU very challenging. In this context, Raboshchuk et al. presented the acoustic environment of a preterm infant in NICU (see Figure 2.11), and addressed extensively the problem of acoustic alarm detection [14–16].

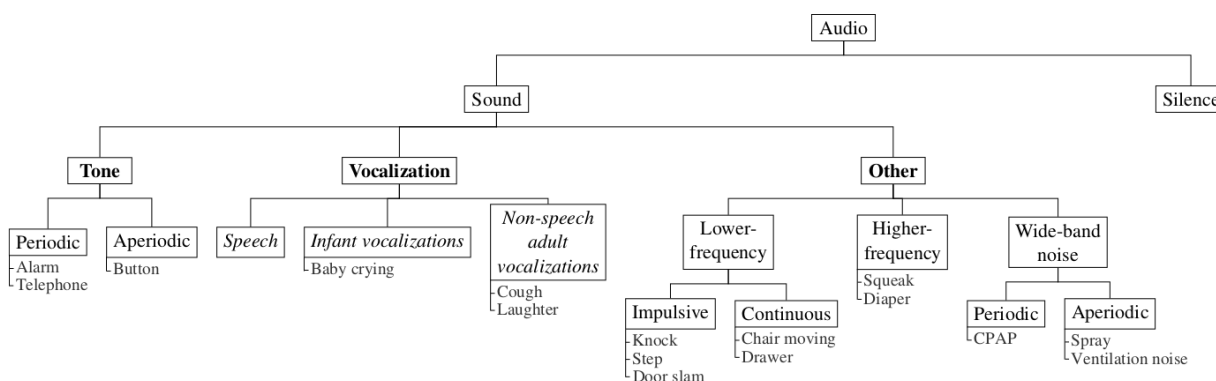


Figure 2.11: A general sound taxonomy of a typical NICU. Taken from Raboshchuk et al. [17].

The recordings made in the framework of the Digi-NewB project are no exception. Therefore, in the following sections, we review the main sound sources occurring in the audio signal that we identified with the help of the medical team. Then, we quantify some of these noises through the annotation of a 15-hour recording performed in the NICU for a very premature infant staying in an incubator. Finally, we introduce the high variability of sound content within a single recording by annotating all the sound events occurring in three WAV files.

2.4.1 Sounds in Digi-NewB recordings

After listening to many audio recordings made during the Digi-NewB project, we gathered the sound sources into six categories that are illustrated in Figure 2.12 and detailed hereafter. Since very premature require more medical assistance than full-term newborns, the sound environment

is generally more prominent in incubator recordings and becomes less important as the infants develop.

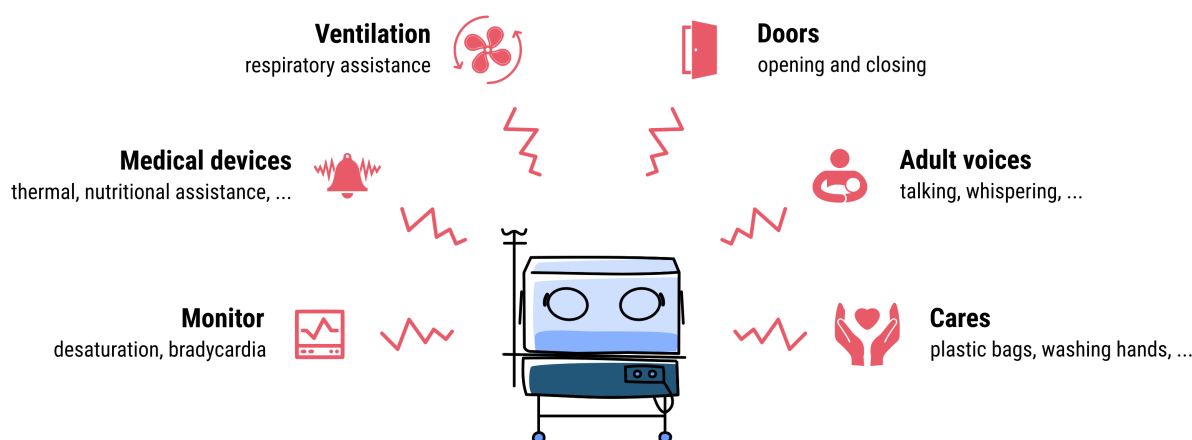


Figure 2.12: The noise sources occurring in the NICU and heard in the Digi-NewB recordings.

MONITOR - During hospitalization, newborns are monitored through sensors measuring their heartbeat, oxygen saturation, and respiratory frequency signals. When an abnormality is detected, such as an irregular heartbeat (bradycardia) or lack of oxygen (de-saturation), the monitor starts ringing through different alarm levels to inform of either an important or a critical situation.

MEDICAL DEVICES - The medical devices surrounding the baby's cradle are designed to recreate the intrauterine environment and provide all the care needed. Furthermore, the more premature is the baby, the more machines are needed for her/his good development and the noisier is the room. The machines aim to warm up the water infusion, warm up the bed, help the infant's respiratory system, feed, etc.

In addition to the fact that some of these machines produce noise during their use, they all have one or more specific alarms to inform nurses of their status when they require human action (i.e., problems or maintenance such as filling the water infusion or the feeding syringe driver). While noises are usually wide frequency bands, alarms are narrow-frequency bands with each of them having a different tone, duration, and time of repetition.

VENTILATION - There are several types of ventilation devices in the NICU that are chosen according to the preterm infant's particular needs. Ventilation produces noise that strongly interferes with the acoustic environment. This noise is usually spread over a wide frequency and is not constant over a recording since it can be turned on or off at irregular intervals. As a result, ventilation introduces a lot of variability to the data.

DOORS - Mainly sliding glass doors that make noise when they open and close and have poor sound insulation properties. A door sound is short in time and has a wide frequency range.

ADULT VOICES - Parents and nurses can be talking or whispering when being in the bedroom. The human voice is a harmonic signal located in the low-frequency range (i.e., 300 up to 3000 Hz).

CARES - Several times a day, care is given to the newborn, for change, wash, feed, ... during these moments noises can be emitted by medical plastic bags, drawers, hand washing, ... Usually very short in time, these noises have a wide frequency band.

For readers who may not be familiar with the NICU environment, a typical neonatal health care unit bedroom is depicted in **Figure 2.13**.

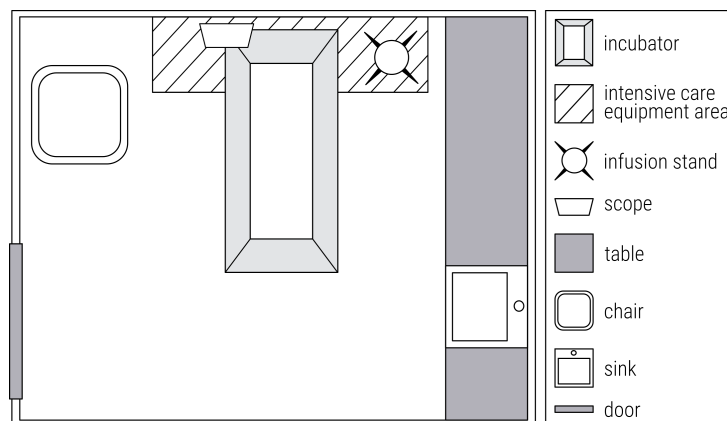


Figure 2.13: A typical NICU bedroom layout at the Rennes CHU.

It is worthwhile to mention that during his/her shift, a nurse is in charge of several newborns. To ensure an intervention when necessary, all the monitors corresponding to those babies are related one to another. Thus, when caring for one baby, the nurse can listen to alarms related to the other newborns she/he is responsible for through the monitor located next to her/him. Therefore, besides the alarms related to the monitored baby, slightly different alarms can occur in the acoustic background coming either from other close bedrooms or from the central reminder (where nurses monitor all newborns). Moreover, although nurses are paying close attention to all the assigned babies, they cannot be everywhere whenever an alarm occurs, therefore, it can lead to long, noisy periods. It is the most harmful and common sound source in the bedroom.

In addition, it is worthwhile to mention that even more complex environments were recorded with co-bedding for twins or shared bedrooms with several babies. Particular attention will be paid to these recordings since we cannot know if the recorded cries are actually produced by the monitored baby.

2.4.2 Noise quantification in a 15-hour annotated recording

To show how noisy the acoustical environment can be in NICU, we studied a 15-hour recording performed in the incubator of a very premature infant ¹.

After listening to the recording (i.e., 15 hours corresponds to 30 WAV files in the Digi-NewB

1. i.e., baby 010049 recorded between 5:00 PM and 8:00 AM the 2017-07-12.

database), we decided to manually annotate two noise categories, i.e., alarms and adult voices which are the most representative. These annotations were performed through a careful subjective listening recognition according to the sound level and spectral content.

Alarms also called beeps are short and often repetitive noises that alert nurses about the status of either the machines or the infant's health. All alarms have different tones allowing the nurses to quickly detect the problem source and fix it. Since the physiological monitor can relay alarms unrelated to the observed baby (see the previous section Sounds in Digi-NewB recordings), we discriminate the alarms into two subcategories:

- *baby alarm*: alarms that directly concern the infant's health status (high intensity level);
- *reminder alarm*: alarms coming from the physiological monitor that correspond to the other neonates under the nurse's care (low intensity level).

We inspected the recording through 1-minute windows. In practice, it means that for each of the 900 minutes, we annotated whether an adult voice, a baby alarm, and/or a reminder alarm occurred. Therefore, when at least one of these sounds occurs in the analyzed minute, we consider it "noisy", whereas when no sound is found, minutes are considered "clear". A resulting timeline is illustrated in [Figure 2.14](#) for a 30-minute WAV file.

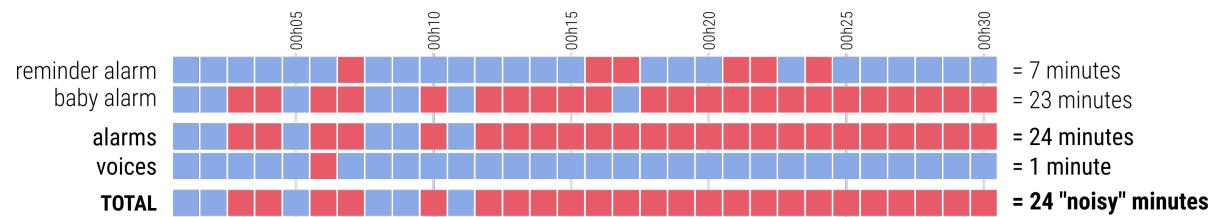


Figure 2.14: Illustration of the annotations performed on the one-minute windowed timeline for a 30-minute audio file. A minute is labeled in red when at least one adult voice or one alarm occurs, otherwise, the minute is labeled in blue. The "reminder" and "baby alarms" lines are combined in the "alarms" line, itself combined with the "adult voice" to form the "total" line.

The annotations' distribution over the 15 hours is presented in [Figure 2.15](#). From a global point of view, results show that 62% of the minutes are considered "noisy", meaning that from the 900 analyzed minutes, 561 of them contain at least an adult voice or an alarm.

Among these noisy minutes, 62% of them correspond to alarms, 19% correspond to adult voices, while 19% are a combination of these two sources.

Since alarms were discriminated through two subcategories, one can see the unexpected distribution. Indeed, only 62% of the minutes are related to the monitored infant, whereas 26% of them come from reminder alarms corresponding to the other neonates under the nurse's care. The remaining 12% corresponds to a combination of both sources.

To go further, we also studied the alarms and identified ten different types. For each of them, the duration between two occurrences is different, as well as the frequency spectrum, which is

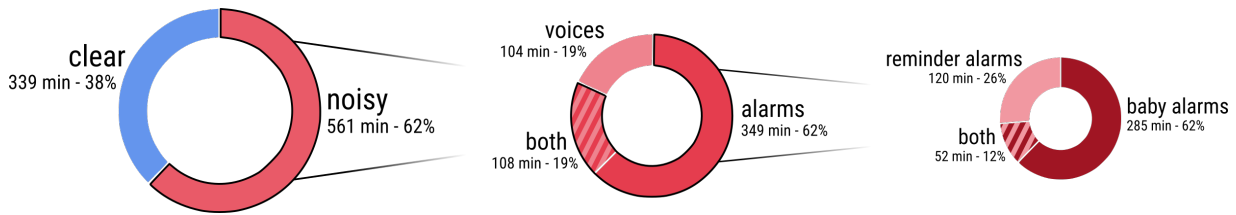


Figure 2.15: Left: Percentage of minutes considered as "noisy" by voices or alarms within the 15-hour recording. Middle: Distribution of the noise categories polluting the sound environment. Right: Alarms distribution after discrimination.

composed of a fundamental frequency that ranges from 400 Hz to 2.4 kHz and may contain several higher components. Furthermore, it is worthwhile to mention that most alarms do not consist of a single beep, but rather of several beeps. Also, this list may not be exhaustive since the study was conducted during the night (5:00 PM - 8:00 AM) for a preterm infant in an incubator. Other alarm types may be encountered during the day and/or in other NICU rooms, depending on the bed and care equipment.

In addition, the author would like to point out the risk of such a noisy environment on newborns. Indeed, it has been recognized that babies are very sensitive to high surrounding variations such as light [18], odors [19], or noise [20] and high exposure can lead to potential neuronal circuits wire damage of the newborn brain. Moreover, premature infants have an underdeveloped auditory system that is not able to adapt to an extrauterine acoustic environment in the same manner as a full-term infant. While the fetus begins to respond to low-frequency sounds after 19 weeks of gestational age [21], the cochlea's response to sound continues to mature between 24 and 35 weeks. As a result, loud noise can create neonate stress responses [22] that may lead to hair-cell damage and subsequent auditory impairments [23–26]. Nevertheless, the problem of the noise level in NICU is well known by the medical staff who suffer from it as well. Hence, improvements are required and the internal review at the Rennes CHU suggested for the future the use of portable alarm systems, the development of new alarm algorithms, or the development of new devices. To date, the neonatology health service in Rennes suggests improving acoustic conditions through alarm management protocols and wishes to consider acoustic improvements during the reconstruction of its site.

In this section, we proposed to quantify the noise that can be present in a NICU bedroom. This characterization, although considering only voices and alarms, showed once again the difficulty of automating crying detection treatments in such an environment.

2.4.3 Sounds variability in recordings

Through this study, we want to show the great variability of the sound content encountered in the different WAV files that constitute a recording. Therefore, three 30-minute sound files were selected from a 20-hour recording made for one baby². Their content is briefly described hereafter:

- 21h25: some sound events with few cries;
- 21h55: a lot of cries with few sound events;
- 01h25: very few sounds, no cry.

Thanks to the knowledge of acoustical environment in the NICU, five labels were chosen according to the possible sound activities and are defined as follows:

- *cry*: infant crying;
- *baby other*: infant vocalization (e.g., cooing), coughing or hiccups;
- *alarm*: alarms produce by medical devices;
- *voice*: adult voices, whispering;
- *other*: background, footsteps, doors, cares, any other noise.

The recordings were manually labeled using the free and open-source digital audio editor and recording software Audacity through start- and end-points identification of all audible sound events in the soundtrack. In practice, it means that even when some sounds occur at the same time and are mixed we annotate each of them individually. The boundaries were set at the points where the sound could no longer be heard. A total of 1 774 sound events have been labeled through the three files. An example of the annotations performed in Audacity software is given in [Figure 2.16](#) and the resulting distribution of segment duration by labels for the three files is illustrated in [Figure 2.17](#).

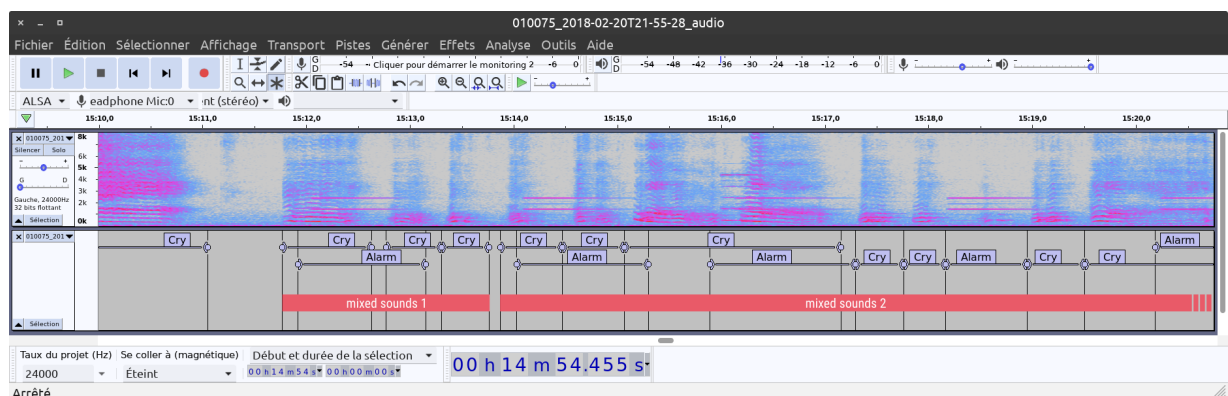


Figure 2.16: Examples of annotations performed in Audacity.

2. i.e., baby 010075 recorded during night time the 2018-02-20.

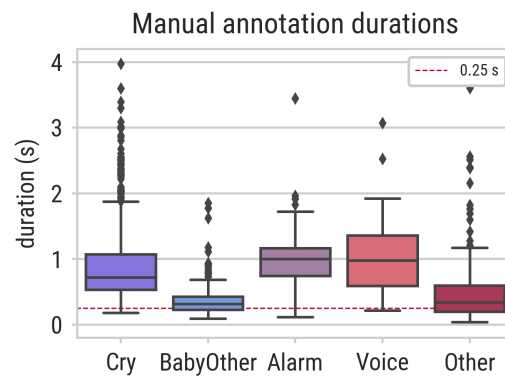


Figure 2.17: Duration of sound events by the label category for all three recordings combined (1 774 sounds). The four annotated sounds, longer than four seconds, are not displayed.

From these distributions, it appears that sounds do not last longer than 4 seconds and that cries are mostly longer than 0.25 s. Although the sounds *Baby other* and *Other* are slightly shorter, all labels have a similar range of duration. Therefore, it is not possible to distinguish crying from other sounds by their duration. Moreover, we remind that we have considered the sounds individually. Therefore, longer durations are to be expected when considering mixed sounds, i.e., when we place limits on the points where all the sounds together can no longer be heard.

During the process, we decided to ignore short sounds with a duration of less than 0.25 s which are more difficult to identify. The resulting 1 593 selected sounds are presented by labels in terms of total segment quantities and duration in [Table 2.1](#).

	21H25		21H55		01H25	
	qty	dur. (s)	qty	dur. (s)	qty	dur. (s)
Cry	155	106.51	776	717.39	-	-
Baby Other	140	64.15	91	37.18	6	2.84
Alarm	50	41.23	255	257.98	4	2.46
Voices	11	11.79	14	17.11	-	-
Other	53	31.13	25	24.56	13	7.06
TOTAL	409	254.81	1161	1054.21	23	12.36

Table 2.1: Labeled sound events for three 30-minute files.

Based on these results one can see the file diversity and the great variability of the sound contained within the same recording. Naturally, the sound environment is not always noisy and if sometimes a 30-minute recording contains almost no sound (i.e., 1:25 a.m. - 23 sounds, 12.36 s), it can also contain a lot (i.e., 9:55 p.m. - 1161 sounds, 1054.21 s). These distributions show once again the complexity of the sound environment and the difficulty to set up a completely automatic processing chain which, however, would be absolutely necessary to perform cry analysis in the NICU.

2.5 Conclusion

The main objective of this chapter was to learn about the neonatal intensive care unit. Indeed, to perform cry analysis, it is necessary to understand this environment, which is designed and used to provide specialized medical care to sick and premature newborns. After a review of the medical equipment that may be used to assist the newborns, we also presented the physiological and neuro-behavioral monitoring usually performed there.

Next, we introduced the European project Digi-NewB in which an audio-video acquisition system was designed and used to record more than 750 infants providing a very large multi-signal database (i.e., heart rate, respiratory rate, video, and audio signals). We also detailed the recording protocols used and illustrated the system's set-up in two NICU configurations.

Then, we presented the acoustic environment in the NICU, which was recorded during the Digi-NewB project, and we reviewed the main sound sources occurring in the audio signal. Through the quantification of the noises occurring within a 15-hour recording on a very premature infant staying in an incubator, we showed the acoustical complexity of the environment. Moreover, to survive the infant needs medical equipment that produces alarms in 62% of the minutes for this recording, which shows that the files can sometimes be very noisy. Finally, with the annotation of three WAV files from a 20-hour recording, we showed the great variability of the sound content between the different 30 minutes files. Indeed, while some contain a few sounds, some others can contain a lot.

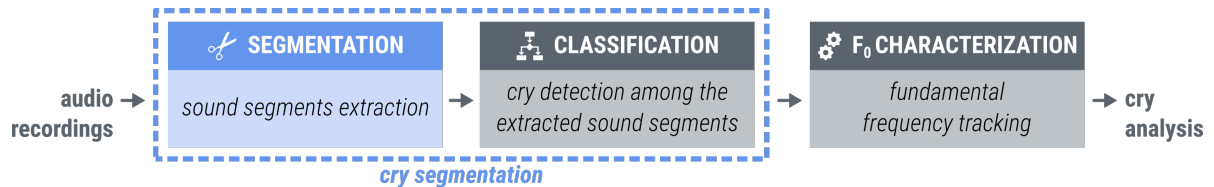
With knowledge of the environmental context and the sounds occurring within the recording, it appears that using the three-step processing chain proposed in the first chapter (see [Section 1.5](#)) is a good approach. Indeed, using a sound event segmentation step seems to be a relevant strategy considering the large amount of data to be processed. Actually, such a procedure is essential to reduce the quantity of data to analyze in order to detect crying. Therefore, in the following chapters, we present the different steps implemented in this work, including the segmentation, classification, and fundamental frequency estimation methods.

BIBLIOGRAPHY

- [1] CHOW S., CHOW R., POPOVIC M., LAM M., POPOVIC M., MERRICK J., STASHEFSKY MARGALIT R.N., LAM H., MILAKOVIC M., CHOW E., AND POPOVIC J. A selected review of the mortality rates of neonatal intensive care units. *Frontiers in Public Health*, vol. 3, page 225 (2015).
- [2] 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*, vol. 38, page R253 (2017).
- [3] ALS H., LAWHON G., DUFFY F., MCANULTY G., GIBES-GROSSMAN R., AND BLICKMAN J. Individualized developmental care for the very low-birth-weight preterm infant. medical and neurofunctional effects. *JAMA : the journal of the American Medical Association*, vol. 272, 853–8 (1994).
- [4] BRAZELTON T.B. AND NUGENT J.K. *Neonatal behavioral assessment scale*. 137. Cambridge University Press (1995).
- [5] PRECHTL H. The behavioural states of the newborn infant (a review). *Brain Research*, vol. 76, 185–212 (1974).
- [6] PRECHTL H.F. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Human Development*, vol. 23, 151–8 (1990).
- [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*, vol. 349, 1361–3 (1997).
- [8] 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*, vol. 132, 300–6 (1998).
- [9] CABON S., PORÉE F., SIMON A., ROSEC O., PLADYS P., AND CARRAULT G. Video and audio processing in paediatrics: A review. *Physiological Measurement*, vol. 40 (2019).
- [10] CABON S., PORÉE F., CUFFEL G., ROSEC O., GESLIN F., PLADYS P., SIMON A., AND CARRAULT G. Voxyvi: A system for long-term audio and video acquisitions in neonatal intensive care units. *Early Human Development*, vol. 153, page 105303 (2021).
- [11] LIVERA M., PRIYA B., RAMESH A., RAO S., SRILAKSHMI V., NAGAPOORNIMA M., RAMAKRISHNAN A., AND DOMINIC M. Spectral analysis of noise in the neonatal intensive care unit. *Indian journal of pediatrics*, vol. 75, 217–22 (04 2008).
- [12] RABOSHCHUK G., NADEU C., GHABABI O., SOLVEZ S., MAHAMUD B.M., DE VECIANA A.R., AND HERVAS S.N. On the acoustic environment of a neonatal intensive care unit: initial description, and detection of equipment alarms. In *Proc. Interspeech 2014*, 2543–2547 (2014).
- [13] SHIMIZU A. AND MATSUO H. Sound environments surrounding preterm infants within an occupied closed incubator. *Journal of Pediatric Nursing*, vol. 31, e149–e154 (2016).
- [14] 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*, 2902–2906 (2015).
- [15] PEIRÓ LILJA. A., RABOSHCHUK. G., AND NADEU. C. A neural network approach for automatic detection of acoustic alarms. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS, (BIOSTEC 2017)*, 84–91. INSTICC, SciTePress (2017).

- [16] 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*, vol. 6, 1–10 (2018).
- [17] RABOSHCHUK G. *Automatic analysis of the acoustic environment of a preterm infant in a neonatal intensive care unit*. Ph.D. thesis, UPC, Departament de Teoria del Senyal i Comunicacions (2016).
- [18] CLAIRE Z., DUFOUR A., PEBAYLE T., DAHAN I., ASTRUC D., AND KUHN P. Observational study found that even small variations in light can wake up very preterm infants in a neonatal intensive care unit. *Acta Paediatrica*, vol. 107 (2018).
- [19] FRIE J., BARTOCCI M., LAGERCRANTZ H., AND KUHN P. Cortical responses to alien odors in newborns: An fnirs study. *Cerebral cortex (New York, N.Y. : 1991)*, vol. 28, 1–12 (2017).
- [20] KUHN P., CLAIRE Z., PEBAYLE T., HOEFT A., LANGLET C., ESCANDE B., ASTRUC D., AND DUFOUR A. Infants born very preterm react to variations of the acoustic environment in their incubator from a minimum signal-to-noise ratio threshold of 5 to 10 dba. *Pediatric research*, vol. 71, 386–92 (2012).
- [21] HEPPER P.G. AND SHAHIDULLAH B. The development of fetal hearing. *Fetal and Maternal Medicine Review*, vol. 6, page 167–179 (1994).
- [22] GRAHAM F., BERG K., BERG W., JACKSON J., HATTON H., AND KANTOWITZ S. Cardiac orienting response as a function of age. *Psychonomic Science*, vol. 19, 363–365 (1970).
- [23] GRAVEN S. Sound and the developing infant in the nicu: Conclusions and recommendations for care. *Journal of perinatology : official journal of the California Perinatal Association*, vol. 20, S88–93 (12 2000).
- [24] HASSANEIN S.M.A., RAGGAL N.M.E., AND SHALABY A.A. Neonatal nursery noise: practice-based learning and improvement. *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 26, 392–395 (2013).
- [25] ZIMMERMAN E. AND LAHAV A. Ototoxicity in preterm infants: Effects of genetics, aminoglycosides, and loud environmental noise. *Journal of perinatology : official journal of the California Perinatal Association*, vol. 33 (08 2013).
- [26] ALMADHOOB A. AND OHLSSON A. Sound reduction management in the neonatal intensive care unit for preterm or very low birth weight infants. *Cochrane Database of Systematic Reviews*, vol. 1 (01 2020).

Audio-Video segmentation



3.1 Introduction

Traditional cry segmentation methods are based on energy thresholding. When applied in a controlled and non-noisy environment these methods lead to accurate cry detection. However, in the context of this work we are dealing with long audio recordings and a large database. In addition, recordings are performed in NICU where the sound environment is very noisy and many sounds occur besides the infant's cries (see [Chapter 2](#)). In this case, it is important to understand that using traditional techniques, all the sounds will be extracted and we should therefore call this approach sound segmentation. Hence, to detect the cries in such signals, a two-step strategy was adopted during the Digi-NewB project, including a sound segmentation step and a subsequent classification step.

This chapter introduces the first step of this strategy. After a review of the state of the art, we describe the method proposed by Orlandi et al. [1] that inspired our own. Then, we introduce the improvements, which have been done, to adapt the method to our data. Finally, we also propose to use the motion information computed by another team of our laboratory (LTSl) during the Digi-NewB project [2, 3]. We suggest collecting only the sounds appearing in newborns' motion intervals to reduce the data quantity to be further processed.

3.2 State of the art

If in the literature, cry segments used to be manually recorded or selected, some recent studies, proposed automated solutions. Indeed, when working with long recordings it is necessary to apply appropriate processing to extract cries from the signal. The preliminary step, defined as the segmentation step, is a must to separate cries from the background soundtrack.

In the context of speech processing, it is common to perform audio segmentation tasks. Indeed,

it is widely used in speech detection in the audio signal or in voiced/unvoiced part detection, resulting in both case in the extraction of relevant parts of the acoustic signal. The same principle is used for infants, it is called the detection of cry units (CU). As in speech, where the initial and final points of a word are located, the objective in cry unit detection is to find the initial and final points of a cry unit. As words in speech, the cry units have higher energy than unvoiced segments.

In the following, we describe traditional segmentation methods and then discuss the new strategies that have emerged to process long-term audio monitoring which has recently grown in popularity.

3.2.1 Methods for cry segmentation in short recordings

At first, cries used to be manually recorded in controlled quiet environments. Thus, most of the traditional methods were based on the computation of Short Time Energy (STE) and Zero Crossing Rate (ZCR). While the first one provides an energy envelope of the sound signal, which helps to distinguish audible sounds from silence, the second one allows to detect the voiced parts.

Methods based on STE thresholding were investigated in [1, 4–8]. Additionally, Orlandi et al. used two thresholds calculated through the Otsu's method to perform the segmentation [1]. Since this segmentation inspired our own, an overview of this method is proposed in this chapter (see **Section 3.3.1**).

Cry segmentation was also performed in combining the two short-time methods, STE and ZCR, [9] and applying a threshold to extract CU. In the continuity of this study, a third step was added to distinguish harmonic and non-harmonic audio segments [10]. Some authors also investigated Simple Inverse Filter Tracking (SIFT) [11] or word reliability [12].

3.2.2 Methods for cry segmentation in long and noisy recordings

Then, with the advancement of technologies, longer recordings were made and required new processing methods. Thus, the recent approaches have considered cry segmentation as a classification problem. In these methods, the whole signal is considered and cut into frames. These frames are then classified into different categories according to the studies. For example, Reggiannini et al. [13] started with a KNN classifier and proposed the three basic classes: voiced part, unvoiced part, and silence.

However, it appeared that in long recordings, besides cries surrounding sounds were also recorded in the signals. Therefore, new sound classes emerged. In the case in [14], where Abou-Abbas et al. considered six classes dividing infant voiced parts (i.e. cries) into expiratory and inspiratory phases with a Hidden Markov Model (HMM). Later, these results were improved by decreasing the

number of classes and gathering the non-cry sounds in a class called "others" [15] or "residuals" [16].

The discrimination of the three classes achieved with the KNN resulted in an Area Under the Curve (AUC) of 0.88 [13]. In comparison to HMM, Gaussian Mixture Model (GMM) gave the best results with a classification error rate of 8.9% [15], while Naithani et al. reached a total accuracy of 89.2% with HMM [16].

Then, researchers began to perform long recordings in real-life settings which are much less controlled. Thus, once again, new methods emerged to deal with the many other sound events occurring, besides the infant's cries, in the recordings.

Most of these methods, based on deep learning approaches, are used to detect cries in domestic environment [17–20] or in the NICU [21, 22]. All these studies worked with a Convolution Neural Network (CNN) and the input layer is computed from Mel-Frequency Cepstral Coefficients (MFCCs) associated with either the Mel-Filter Bank (MFB) [17, 18, 20–22] or the Linear-Filter Bank (LFB) [23].

However, to reach good performances, these processes require large amounts of data, and some authors proposed to introduce normalization and regularization to adapt CNN to a limited data set [18], or to enhance the data set with simulated data [21].

As a result, automatic cry classification in domestic environment led to AUC over 90% in [18] and an averaged Area Under Precision-Recall Curve (PR-AUC) of 90% in [23]. Moreover, the promising results in [17] were confirmed with considerably better performance compared to a traditional machine learning classifiers (SVM and logistic regression) in [20], especially for low false-positive rates. While, in the NICU, an average accuracy of 86.58% was obtained [21] and a PR-AUC of 97.50% was reached on real data in [24].

3.2.3 Discussion

Therefore, when processing recordings performed in controlled quiet environments, an automatic cry segmentation can be easily computed based on energy thresholding methods. However, these techniques are no longer sufficient when it comes to recordings in a routine hospital environment. In fact, in such a noisy environment, all occurring sounds are segmented and must be sorted to find those that contain crying.

To date, only a few studies have achieved cry segmentation methods in a monitoring context, and none have involved long recordings in a routine hospital care setting. Moreover, while frame-by-frame classification methods work well, they are computationally intensive to process long records. As this work focuses on the large Digi-NewB database, we proposed a two-step strategy including an audio segmentation step and a classification step to reduce the data quantity to be processed.

3.3 Audio segmentation method

The first step of our strategy is therefore to segment the sound events that occur in the recordings. We choose to exploit the method proposed by Orlandi et al [1] which is based on energy and threshold calculations, the process is described and discussed below. Subsequently, we propose improvements due to the issues encountered during the application to our data.

3.3.1 Orlandi's method

This traditional segmentation method based on energy thresholding is described in details in [1] where a long term audio analyzer was proposed. The process was compared to existing software tools commonly used in biomedical applications using two synthetic signals sets: the first one was based on adult voice excerpts and the second one was obtained from newborn cries. This method is now deployed in the user-friendly voice analysis software BioVoice also developed by the Italian team [6, 8, 25].

This method, illustrated in an exhaustive workflow in [Figure 3.1](#), is described in the following and can be decomposed into three main steps:

- pre-processing: band pass filtering and down-sampling of the recording;
- automatic segmentation: detection of sound intervals in the signal;
- duration filtering: exclusion of short sounds.

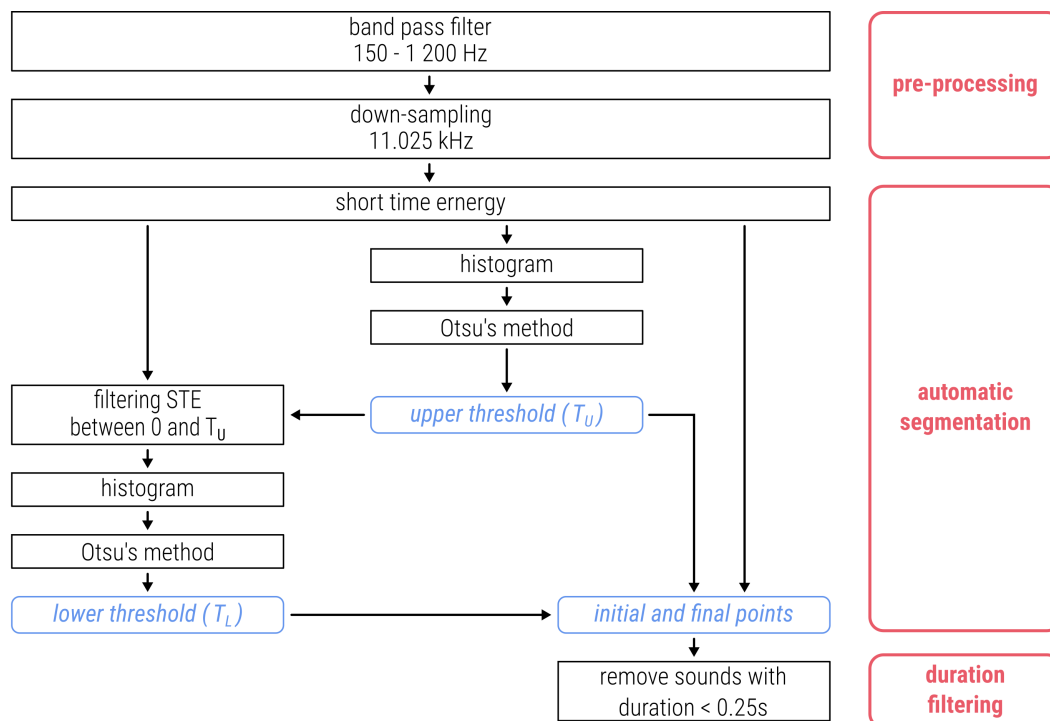


Figure 3.1: Segmentation workflow proposed by Orlandi et al. in [1]

PRE-PROCESSING - First, the recorded signal is band-pass filtered by a 5th-order Butterworth filter and cut-off frequencies set between 50 and 1000 Hz. Then, the resulting signal is also down-sampled to 11.25 kHz to speed up processing.

AUTOMATIC SEGMENTATION - SHORT-TERM ENERGY (STE) The pre-processed signal is then divided into 20-millisecond windows with 50% overlap between adjacent windows. On each window the short time energy is evaluated as:

$$\mathbf{ste} = \log_{10} \left(\frac{\sum_{i=1}^n s(i)^2}{n} + \varepsilon \right) \quad (3.1)$$

where n is the number of samples in the window, s is the signal and ε is a small constant to avoid $\log(0)$. The resulting values of all windows are stored in an energy vector named **ste**. An example of this vector is given in [Figure 3.2](#) for a 3-second signal containing two cry units. We can see that STE values increase during the cries and decreases during silences.

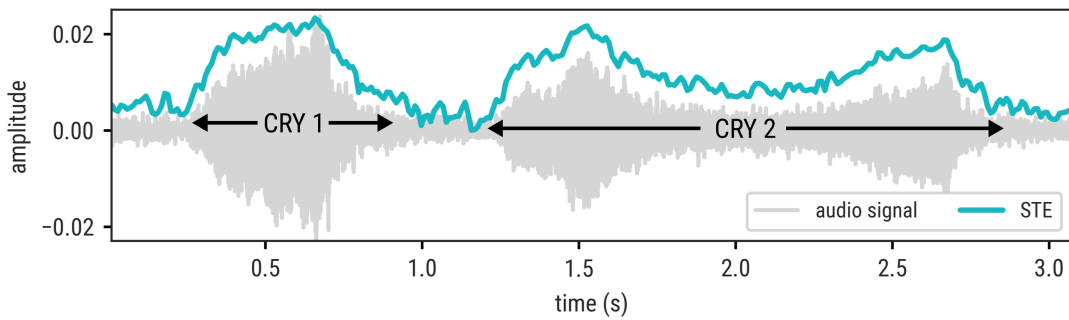


Figure 3.2: STE values (blue) computed on a 3-second signal (grey) containing two cry units.

AUTOMATIC SEGMENTATION - OTSU'S THRESHOLDS - Suitable thresholds are required to determine the boundaries, i.e., sound events start and stop points. They are obtained using a modified version of Otsu's method [26] applied on the STE value histogram. While the original method is described in the [APPENDIX](#) of this chapter, Orlandi's proposal [1] is detailed hereafter.

1. First, the histogram of the STE values is calculated through 2000 levels for a reasonable compromise between sufficient detail and computing speed.
2. Next, the upper threshold (T_U) detects the sound event apparition. It is computed on the whole STE value distribution and is illustrated with the corresponding histogram in [Figure 3.3a](#). The segments of value lower than T_U are considered as silences and those higher than T_U as sound segments. The application of the threshold on the signal containing the two cries is presented in [Figure 3.3b](#). We can see that by considering only the upper threshold, the second cry is mis-segmented and divided into two segments.

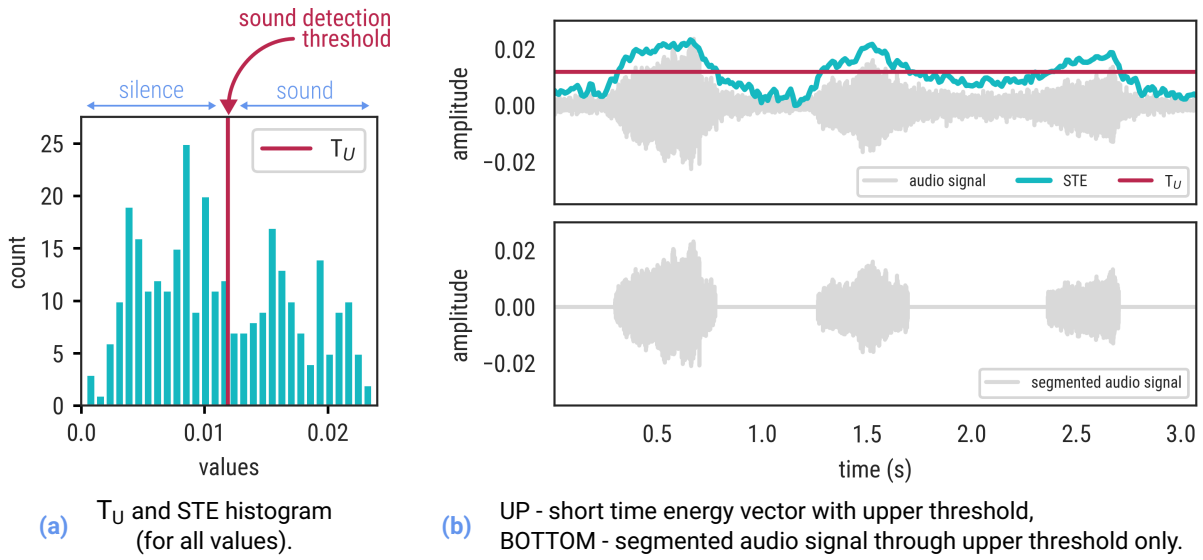


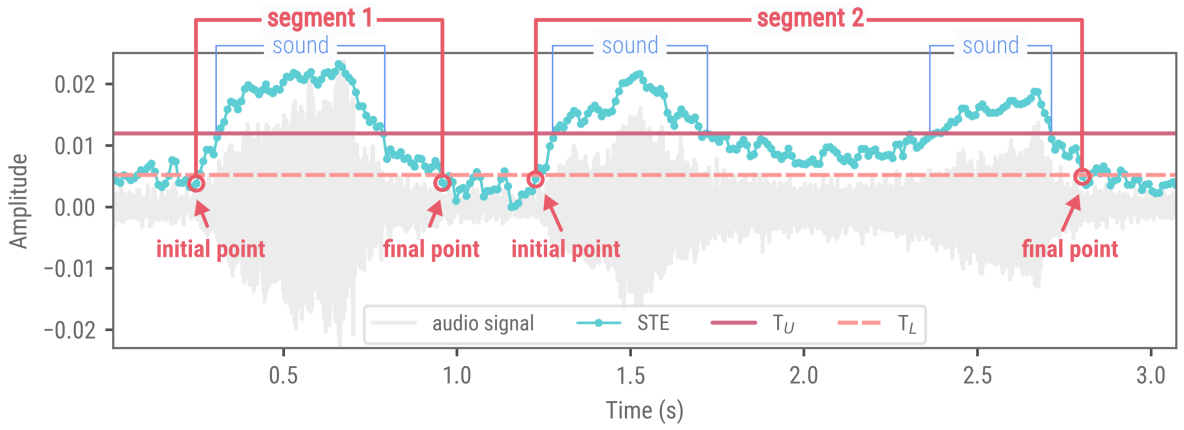
Figure 3.3: Upper threshold (T_U) computation.

3. Therefore, to enhance the segmentation a second lower threshold named (T_L) is computed also based on Otsu's method but with STE values included between zero and the upper threshold T_U . Applying both thresholds requires a segmentation step using the upper threshold to detect sounds and the lower threshold to find their start- and end-points. This segmentation process is illustrated in [Figure 3.4a](#) and detailed hereafter.

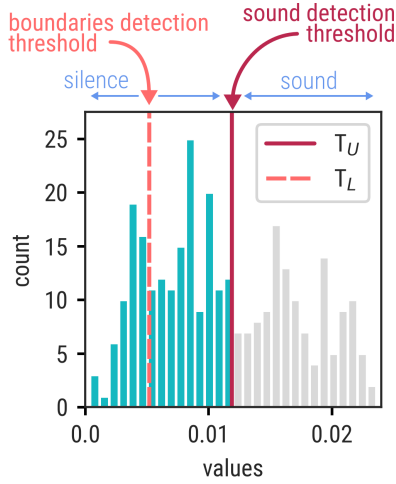
- (a) First, when several successive points of the **ste** have values higher than T_U we call the gathering of these points an *interval* which we consider as a detected sound.
- (b) Then, the initial point is determined as the first point under T_L located to the left of the sound start. In practice, all the elements preceding the first point of the interval, are checked and the first element for which the energy value is lower than T_L is defined as the starting sample of the sound.
- (c) Finally, the same principle is applied to find the final point which is defined as the first element occurring below T_L after the sound offset.

The two thresholds segmentation is illustrated with the the lower threshold computation and the corresponding histogram in [Figure 3.4b](#). The application of both thresholds on the signal containing the two cries is presented on figure [Figure 3.4c](#). This time, we can see that the two cries are correctly segmented.

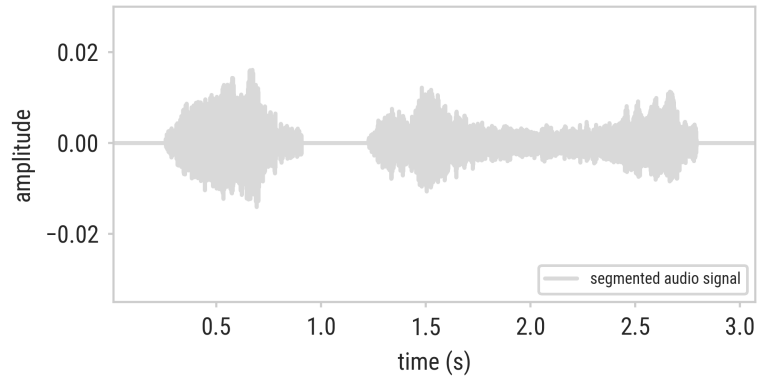
4. Finally the two thresholds are multiplied by a factor d given by the ratio of the differences between the maximum and minimum values of energy and the number of levels of the histogram. This gives the upper T_U and the lower T_L thresholds required to determine if a frame is voiced or not. Then, minimum value of the signal energy is added to T_U to guarantee that T_U is above the minimum.



(a) Segmentation technique illustration where boundary points are detected for all STE intervals that surpass T_U . In the picture, the initial point is determined as the first point under T_L located on the left side with respect to the sound onset. The final point corresponds to the first point occurring below T_L after the sound offset. In this example, the resulting segmentation returns two segments. While *segment 1* matches the first cry, *segment 2* gathers (thanks to the double threshold method) two detected sounds corresponding to the second cry unit that has lower energy in the middle.



(b) T_U , T_L and STE histogram (for values from zero to T_U).



(c) UP - short time energy vector with upper and lower thresholds, BOTTOM - segmented audio signal through both thresholds.

Figure 3.4: Lower threshold (T_L) computation.

Using a double threshold prevents sound split and improves the segmentation by finding better boundaries and extracting the complete sound event.

DURATION FILTERING All detected sounds with a duration of less than 250 ms are removed, so that inspiratory sounds are not taken into account [6]. Moreover, this duration was chosen for its relation to the physiological infant voice properties as four times a second is how fast the vocal cords can change and is what is needed to obtain a complete acoustic profile of the newborn [27].

Thresholds computation on Digi-NewB data

As described in **Chapter 2**, in the framework of the Digi-NewB project the audio signals recorded are performed in a routine hospital care environment, and data are stored in 30-minute WAV files (i.e. 24 hours recordings = 48 wav files). Hence, there are several ways to apply the automatic segmentation on our data with the thresholds that can be calculated:

- *locally* - with new computation for each 30-minute file;
- *longitudinally* - with a single computation for each recording (i.e., several hours);
- *generally* - with a single computation for all recordings.

We decided to use the first strategy with local thresholds because the sound contents are very variable within a single long recording (several hours) and especially between different recordings/babies with distinct configurations. For example, newborns in incubators require a lot of assistance and the sound environment can be disturbed by noisy machines (such as ventilators, heaters, ...) while the surroundings of infants in a regular cradle can be very quiet.

Therefore, since Otsu's method is sensitive to the environmental audio content (based on the fact that thresholds are computed over the signal energy values distribution), it was decided to compute local thresholds to ensure crying detection in any acoustical context. In practice, the method applied to our data consists in computing the short-time energy vector and Otsu's thresholds for each 30-minutes file.

The signal resulting from the segmentation process for a 23-second noisy audio signal is illustrated in **Figure 3.5** with the different sound sources occurring in the recording. We can observe that some extracted segments are very long (e.g., more than 5 seconds).

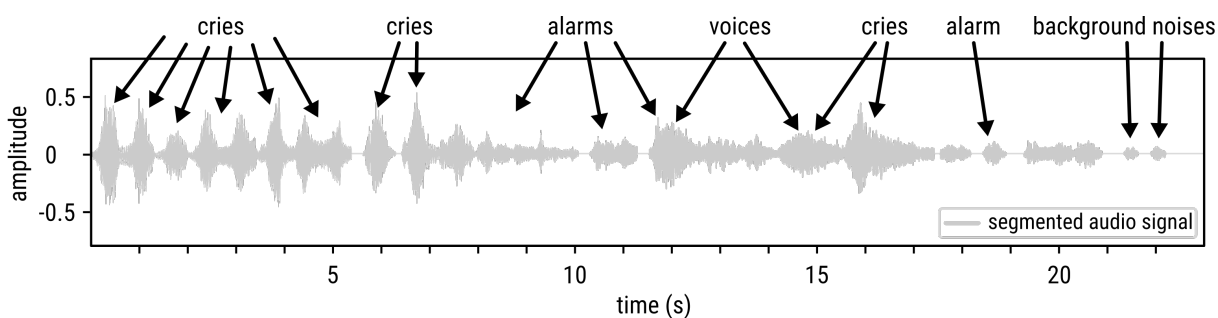


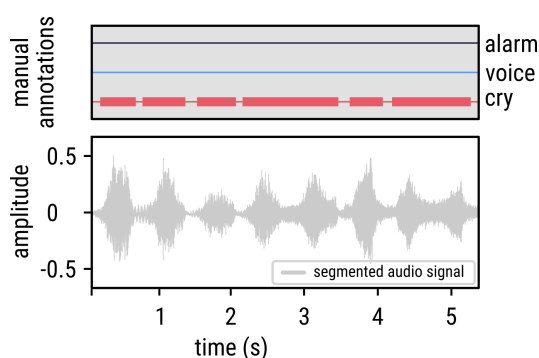
Figure 3.5: Sound segments resulting from the segmentation process for a 23-second noisy audio signal represented on its original time axis.

Indeed, the deployment of the method proposed by Orlandi et al. on the Digi-NewB database resulted in issues related to *i*) a poor boundaries (i.e., initial and final points) detection in different cases as well as *ii*) the replication of the method. Both subjects are addressed in the following sections.

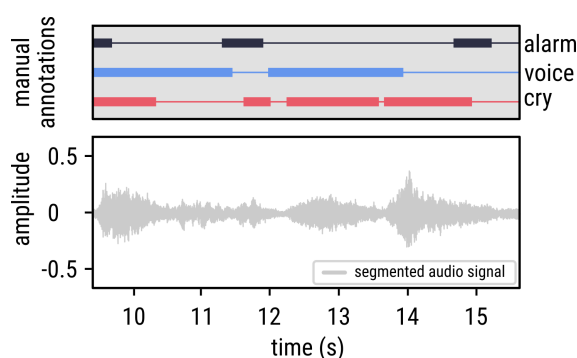
Issues related to poor boundaries detection

After applying the segmentation method on our data, we noticed several errors in the detection of the initial and final points in the following three cases:

- First, in the case of crying bout (several consecutive cries) and when the pause between two cries is very short only one audio segment might be extracted. This is due to the fact that the energy signal is calculated using sliding windows and that the STE values cannot decrease. An example is given in [Figure 3.6a](#) where the resulting audio segment is a signal containing a crying bout composed of six cries.
- Then, once again it is worthwhile to remember that the recordings are performed in the NICU, a routine care environment where many sounds can occur besides infants' cries (i.e. alarms, voices, doors, ...). Therefore, it is normal that in such a noisy environment several sounds mix. Thus, the segments resulting from the segmentation might gathers several overlapping sounds since once again the STE values cannot decrease. An example is given in [Figure 3.6b](#) where the resulting audio segment is composed of alarm, voice and cry signals.



(a) Sound segment containing several cry units.



(b) Sound segment containing several types of sounds.

Figure 3.6: Examples of extracted sounds poorly segmented and lasting more than 5 seconds.

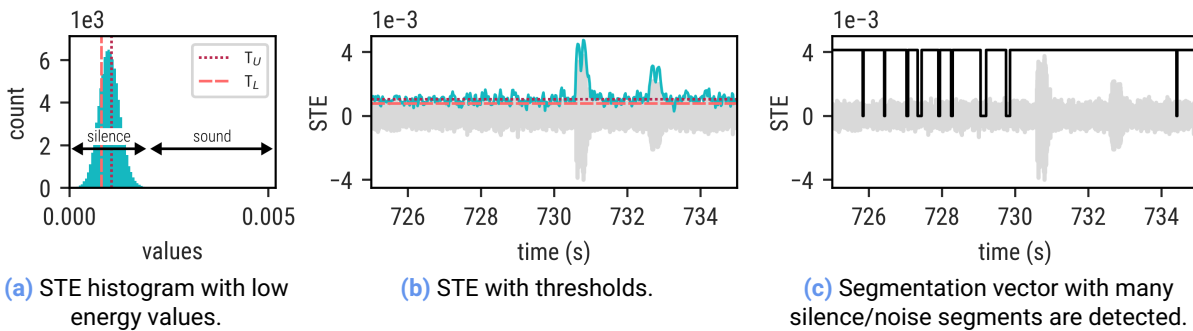
- At last, some noise sources are activated for long period and considerably influence the energy value distribution. This case can occur when a ventilation system is used to assist a newborn in an incubator. The machine produces a constant background noise that is randomly turned on and off during the whole recording. Impacting the energy value distribution can lead to the detection of very long sound events matching the ventilator activation. In that case, the extracted sound events can last several seconds or more.

Issues with replication

In addition, we faced a problem when reproducing step 4 of Orlandi's method (see [Section 3.3.1](#)). Indeed, in this step, the authors propose to adjust the thresholds in particular with the help of a ratio using the minimum of the energy. According to them, this step is used to guarantee that T_U is above the minimum. However, in our case, the minimum of energy is very small, with values close to zero. Therefore, we have not implemented this step in the replication.

In fact, this type of step is necessary when processing signals containing very few sound events, because the majority of the energy values are located in the low amplitudes. Thus, the thresholds, which are computed on the data distribution, do not allow to separate the few sounds from the background noise.

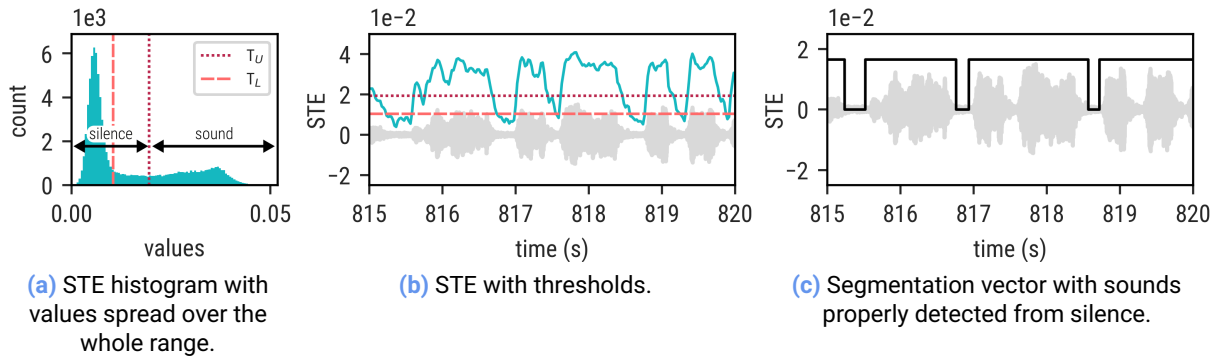
The case of a signal extracted from a recording containing very few sound events is illustrated in [Figures 3.7](#). On the left image, we can see the histogram of energy values that are concentrated around a small amplitude. If the two thresholds calculated with this distribution are too low to separate sound from silence, they are sufficient to segment silence (i.e., noise) such as illustrated in the right picture.



Figures 3.7: Segmentation example with an extract from a quiet recording. Since few sound events occur, STE values are mostly located around low amplitude (a) thus, thresholds cannot help to detect sound events (b) leading instead to the detection of many silences/noises during the segmentation step (c).

Conversely, the example of a signal extracted from a recording containing many sound events is illustrated in [Figures 3.7](#). The energy value are spread over the histogram and the two thresholds allow to separate sound from silence.

Therefore, in the next section we propose improvements so that the method correctly handles *i)* the resulting segments of long duration that are likely to be part of the three cases described above and *ii)* the recordings that do not contain many sounds.



Figures 3.8: Segmentation example with an extract from a recording containing numerous sounds. In this case, STE values are spread over the whole range (a) thus, thresholds help to detect sound events (b) that are correctly extracted from the background silence during the segmentation step (c).

3.3.2 Improvement of the method

Regarding the extracted segments of long duration, we propose a re-segmentation step. Then, in order to manage sound detection in recordings with little sound content, we introduce a modified workflow that better takes into account the sound environment thanks to new frequency filters. Finally, to reduce the amount of data to be processed, we suggest a supplementary step to ignore recordings containing very few sounds. These three improvements are presented below.

Re-segmentation (RS)

As we showed, the segmentation sometimes results in the extraction of long segments containing several cry units (see Figure 3.6a), overlapping sounds (see Figure 3.6b), or noisy periods. In order to process these segments efficiently, we propose a re-segmentation step within the final duration filtering step (see Figure 3.1).

After applying the pre-processing and automatic segmentation steps, the segments of duration longer than five seconds are identified. For each of them, the corresponding pre-processed signal is extracted and new local thresholds are computed. These local thresholds are then applied to the extracted signal to find new initial and final points. Eventually, among the resulting segments, only those with a duration between 0.25 and 5 seconds are retained. This re-segmentation step is illustrated in Figure 3.9.

Narrowing STE frequency band (NFB)

To better handle the recordings that do not contain many sounds, we propose to modify the pre-processing and automatic segmentation steps of the initial workflow by using a double frequency filter.

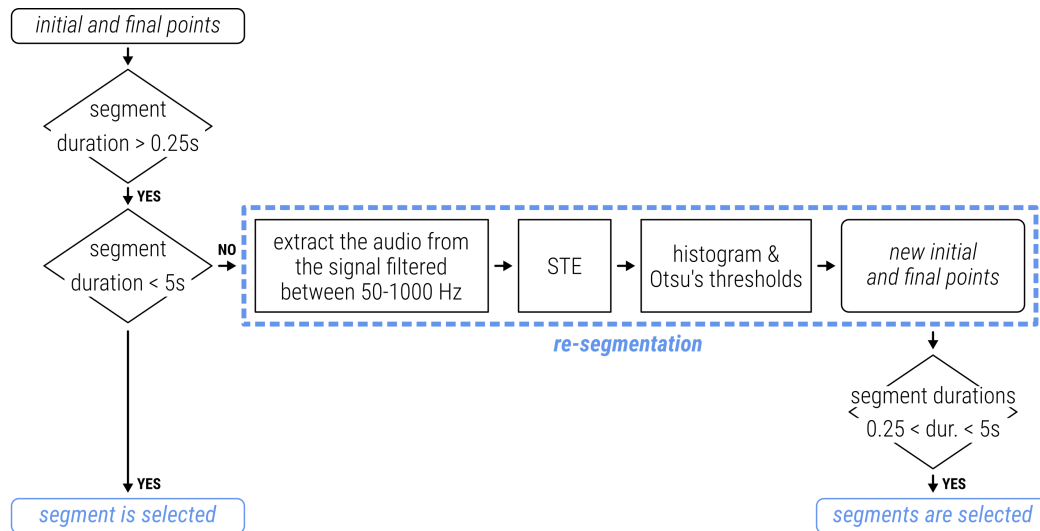


Figure 3.9: Flowchart of the duration filtering step with the re-segmentation procedure for each segment with duration greater than five seconds.

The idea is to filter at the same time the original signals in two different frequency bands:

- the first one is based on the default frequency band used by Orlandi et al. [1]: 50-1000 Hz which is a proper band to consider the surrounding sounds such as human voice and low pitch noises.
- the second band-pass is set between 200 and 1000 Hz, which is a reasonable frequency band where an infant can be expected to cry [28].

Once both signals are pre-processed, the two thresholds can be computed between 50 and 1000 Hz to be more sensitive to the acoustical environment. However, the short-time energy on which the thresholds are applied is computed on the signal filtered between 200 and 1000 Hz. This strategy allows a better segmentation in recordings that do not contain many sounds and has the advantage to reduce the segment detection to sounds with energy located within the infant crying frequency band. Therefore, the final workflow is illustrated in [Figure 3.10](#).

Long-term threshold (LTT)

First, to detect variations in sound content in the recordings, we suggest using a threshold T computed on sliding window of two hours. To do so, T is computed like T_U over the concatenation of 4 **ste** vectors calculated on the pre-processed signal filtered between 50 and 1000 Hz. Then, the files with less than 10 intervals detected above this threshold are discarded. This step allows to reduce the number of resulting segments by ignoring 30-minutes files with very little detected sound content.

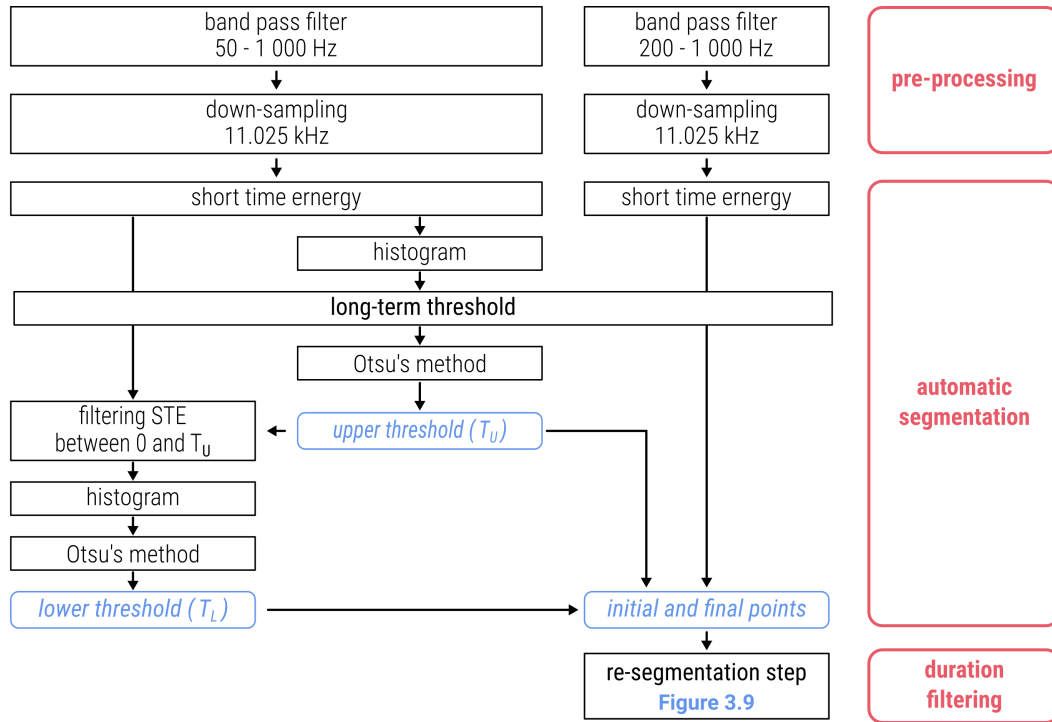


Figure 3.10: Workflow of the updated segmentation method.

3.4 The use of motion for audio segmentation

It is quite natural to consider that a baby is moving when she/he cries. Indeed crying, which requires a deep respiratory activity, helps the newborn to communicate discomfort. Thus, both of these aspects can lead to movement. Studies investigating the correlation between crying and movement in preterm infants are mostly related to the assessment of the behavioral sleep stages [29]. The three usual categories are: active sleep, quiet sleep, and wake. According to a recent study, while the sleep states include reflexive body movements with sobs, sighs, and distressing noises, the wake behavioral state includes high body activity level and crying [30]. Therefore, information about the infant's movements could be used to reduce the amount of data to be processed during automated crying analysis.

In addition, Orlandi et al. proposed a contactless system for audio-video infant monitoring (AVIM) in which both modalities are considered separately [6]. In their study, the motion analysis is semi-automatic since the user needs to select points to track on the video frame, and automatic crying analysis is performed after the manual removal of interfering sounds.

In this work, the automatic sound segmentation allows extracting the sounds with energy located between 200 and 1000 Hz (i.e., crying, adult voices, monitor beeps...) and thanks to the work of another team of the laboratory, the automatic motion segmentation allows identifying the intervals of movement and non-movement.

Therefore, after presenting the video segmentation method in the following section, we propose to use joint audio and video processing to reduce the number of segments to be further processed in the classification step such as illustrated in [Figure 3.11](#). To our knowledge, it is the first time that video signals are used for crying segmentation. In practice, it means to collect sound segments occurring in specific motion segmentation intervals. Nevertheless, in order to evaluate the validity of this approach, we propose a preliminary study to investigate *i*) the amount of motion in the recordings, *ii*) the sound distribution and *iii*) the cries distribution.

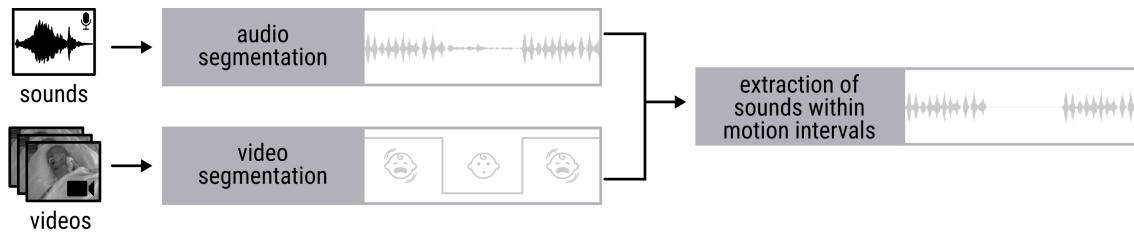


Figure 3.11: Audio segmentation strategy by extracting sounds occurring within infant's motion intervals.

3.4.1 Video segmentation

The video segmentation method was developed during the Digi-NewB project by Cabon et al. [31]. It relies on video analysis to extract different infant movement states and is based on the following steps:

1. First, the infant's movement is calculated by an inter-image difference [31, 32]. An example of a motion signal is presented in [Figure 3.12](#) with frame samples of the movement states.

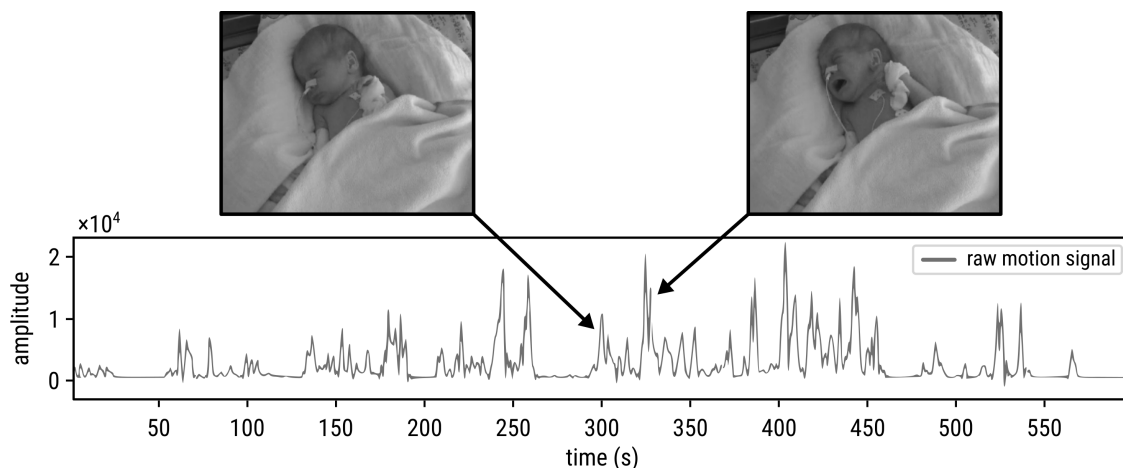


Figure 3.12: Example of a motion signal with two sample frames acquired when the infant is still (left) and in movement (right).

- Then, the intervals when the baby is not present in the bed, as well as those including the presence of adults (parents or caregivers) in the field are automatically excluded. This essential step is performed thanks to a Deep Learning approach (see [3] for more details). The different configurations are presented in [Figure 3.13](#).

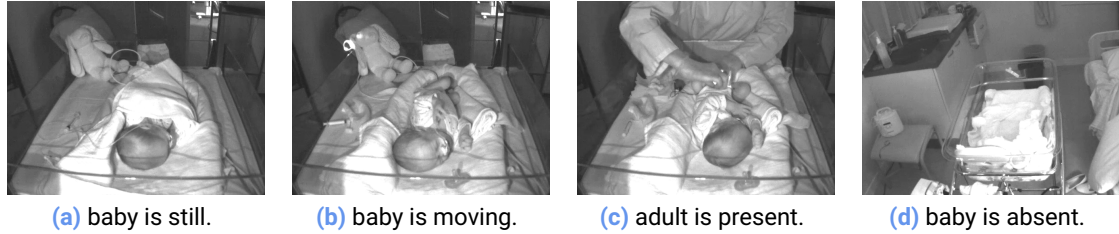


Figure 3.13: Illustration of the possible configurations of images in the recordings. The intervals corresponding to baby presence such as images (a) and (b) are processed, while the intervals with adult presence (c) and/or baby absence (d) are excluded.

- Finally, the movement and non-movement intervals are segmented using an approach based on a Random Forest classification. To do so, a pre-processing step is applied to clean the noise within the raw motion signal. Then movement and non-movement intervals are detected and a synthetic signal is constructed. It is equal to 1 during movement intervals, 0 during non-movement intervals and “NaN” in case of absence of the baby or presence of adults. A last step allows to eliminate or merge the periods where these intervals are very short [31]. The segmentation steps are illustrated in [Figure 3.14](#).

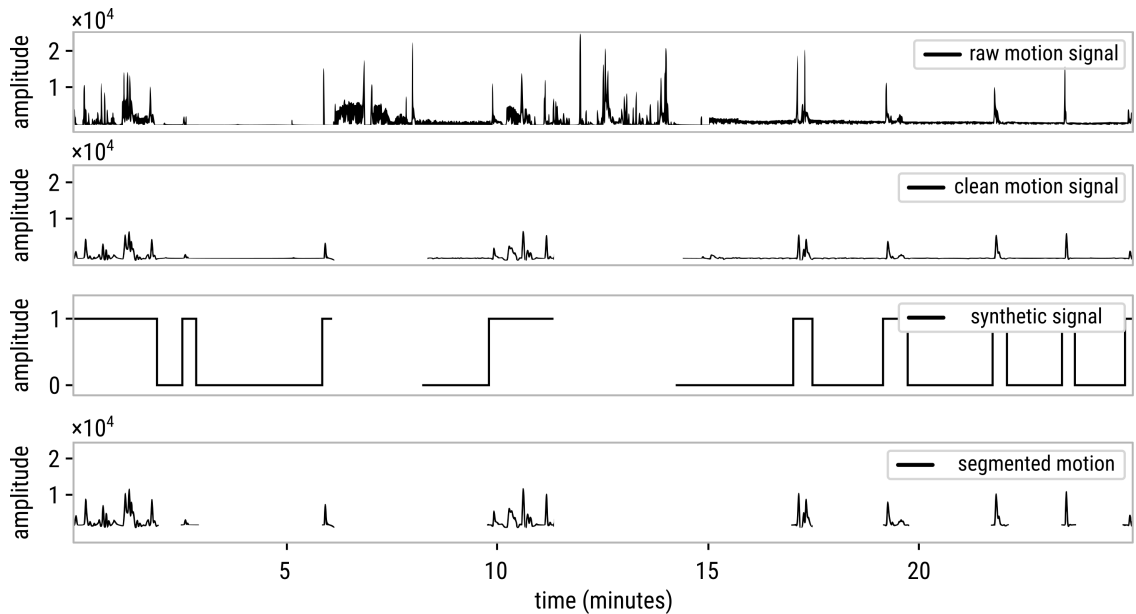


Figure 3.14: Illustration of the motion and non-motion interval segmentation steps (from top to bottom): the raw motion signal, the clean motion signal, the synthetic motion signal resulting from the segmentation, the final segmented motion signal.

3.4.2 Database

We selected a set of data representing a large part of the diversity encountered in the project. Hence, 36 recordings were selected from the Digi-NewB database. They involve 10 girls and 12 boys born between 25+6 and 40+3 GA and recorded between 28+1 and 41+3 PMA.

In total, 243 hours (i.e. 487 audio and 487 video files of 30 minutes) were processed. Recordings were generally performed overnight periods, between 9:00 PM and 6:00 AM, and lasted about 8 hours each.

Using the audio segmentation method (see [Section 3.3.2](#)), 191 533 sound segments were automatically extracted corresponding to 1 day, 10 hours, 26 minutes and 53 seconds duration.

To check the cries distribution within movement, a part of these sound segments was annotated manually. Finally, a total of 4 150 cry segments were identified through human listening, corresponding to a duration of 1 hour, 18 minutes and 46 seconds.

3.4.3 Motion quantification

First, we quantify the percentage of motion in each recording. To do so, we apply motion segmentation to all 36 video recordings of the 22 infants.

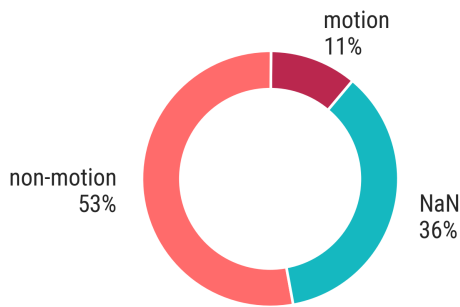
Motion distribution averaged over all recordings is presented in [Figure 3.15a](#) and shows that newborns do not move much (i.e. 11% of time on average) and are mostly immobile (i.e. 53% of time on average). The remaining time corresponds to intervals where an adult is present or when the baby is absent from the image.

While the quantity of non-movement and NaN intervals vary greatly between the 36 recordings (see [Figure 3.15b](#)), the movement intervals are steadier and remain lower than 40%.

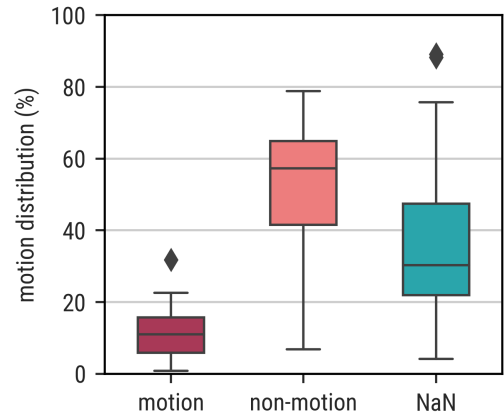
3.4.4 Sounds within infants' movement

In this section we consider all the automatically extracted sound segments resulting from the audio segmentation step and we quantify their distribution within motion segmentation. In practice, it means that for each sound, we observe the corresponding motion segmentation signal and consider the sound with the following conditions:

$$\text{sound is } \left\{ \begin{array}{ll} \text{in motion} & \text{when motion segmentation equals 1,} \\ \text{in non-motion} & \text{when motion segmentation equals 0,} \\ \text{in NaN} & \text{when motion segmentation equals NaN.} \end{array} \right. \quad (3.2)$$



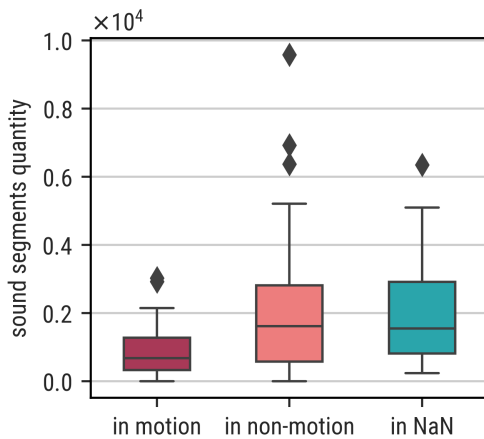
(a) Averaged distribution for the 36 recordings.



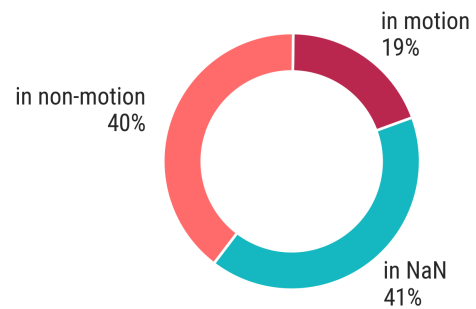
(b) Overall distribution for the 36 recordings.

Figure 3.15: Distribution of motion segmentation (including 22 babies, 36 recordings, 191 533 sound segments).

Since audio and video signals have different sampling rates, we have chosen to apply a simple decision which consists in considering a sound within motion if at least one point of the corresponding motion segmentation is equal to 1 and in the majority category otherwise. Results of the sound distribution in motion are presented in [Figures 3.16](#).



(a) Sound segment quantity in recordings.



(b) All sound segment durations accumulated.

Figures 3.16: Sound segments distribution in motion segmentation (including 22 babies, 36 recordings, 191 533 sound segments).

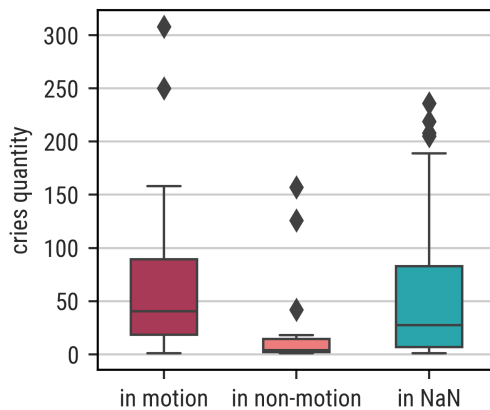
The distribution within motion segmentation of sound segments in each recording ([Figure 3.16a](#)) shows that there are generally more sounds extracted in non-motion and NaN intervals (i.e., baby absence and/or adult presence) than sounds in motion intervals. In addition the quantification of the accumulated sound durations ([Figure 3.16b](#)) show that the extracted sounds appear:

- in non-motion phase during 40% of the time and should correspond mostly to machine noises;
- in NaN intervals during 41% of the time which correspond to the baby's absence and/or the presence of adults in the image field. It is important to note that in the latter case, it is normal to observe a significant sound contribution since adults are usually there to take care of the newborn. Hence sounds occurring in this interval should mostly correspond to adult voice, cares and newborn cries;
- in motion intervals during 19% of the time and should contains infant cries.

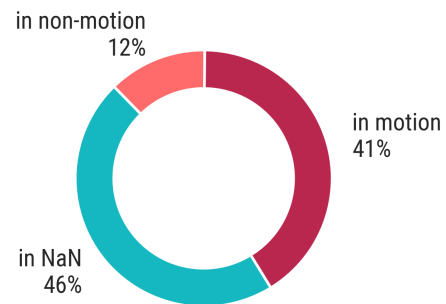
Thus, the purpose of the following section is to assess the cries distribution within motion segmentation.

3.4.5 Cries within infants' movement

In this section we consider the 4 150 cries manually identified. In the same way as before we observe, for each cry, the corresponding motion segmentation signal. Results are presented in [Figure 3.17](#).



(a) Cries quantity in recordings.

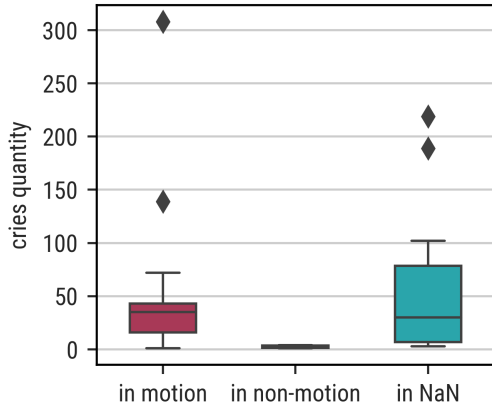


(b) All cry durations accumulated.

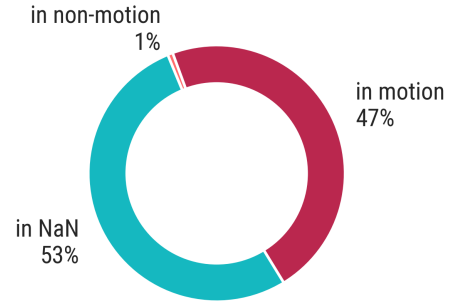
Figures 3.17: Cries distribution in motion segmentation (including 22 babies, 36 recordings, 4 150 cries).

The distribution of cries within motion segmentation in each recording ([Figure 3.16a](#)) shows that most of the cries occur in motion and NaN intervals. In addition, according to the quantification of the accumulated cries duration ([Figure 3.17b](#)), cries were included in non-movement intervals in 12% of the time. Therefore we investigated these recordings and we identified that they were performed in shared-bedroom or co-bedding configurations (see [Section 2.4.1](#)). Hence, some of the detected cries occurring in non-motion intervals are not produced by the monitored baby but rather by a neighboring baby.

Therefore after identifying the recordings matching this condition, we removed them and computed one more time the crying distribution within the motion segmentation. The new results are presented in [Figure 3.18a](#).



(a) Cries quantity in recordings.



(b) All cry durations accumulated.

Figures 3.18: Cries distribution in motion segmentation (including 12 babies, 17 recordings, 1591 cries).

These new results showed that only few cries occur when the baby is not moving (i.e. less than 1% within the non-movement intervals). After a new investigation, we found that it correspond to video segmentation errors especially with the detection of non-movement intervals instead of baby absence intervals.

3.4.6 Discussion

In light of the preliminary study results, we can conclude that cries never occur in the non-motion intervals. Hence, video segmentation can be used to reduce the amount of signal to be processed. Indeed, limiting the sound segmentation within motion intervals reduces by 80% the sounds to be classified afterwards (see [Figure 3.16b](#)). However, if it is easy to ignore the non-movement intervals, special attention should be paid to the NaN intervals.

On the one hand, these data must be analyzed differently depending on the application. When the goal is to detect as much crying as possible, they must be kept, however, when motion information is required, they can be removed. This is the case in sleep stage estimation [29], where signal information are combined to define the infants' sleep states. In this case, the NaN intervals are unusable anyway for motion analysis, hence, audio processing is not performed during these intervals either. On the other hand, these intervals correspond to complex data, especially in the presence of adults. Indeed caring can produce a lot of sounds that may be mixed with the crying.

Hence, in the framework of this work although many cries occur during NaN intervals, we decided to collect the sounds occurring within motion intervals only. In addition, this strategy can be

useful to process recordings made in shared-bedrooms since it should limit the detected crying amount that does not belong to the monitored infant.

In regards to the studied database, considering only infant's movement intervals would reduce to 19% of the total duration of the sounds automatically segmented. In terms of duration, it means processing 5 hours instead of the 34 hours initially segmented. Therefore the video segmentation can be a valuable strategy especially when processing very large database such as the Digi-NewB.

3.5 Evaluation strategy

To assess our segmentation method, three 30-minute files were manually annotated by identifying the start and end points of all audible sounds and their type (crying or not crying). The proposed strategy is to perform a comparison between cries manually identified and the sound segments derived from the automatic segmentation methods.

To perform such a comparison, two detection signals are created for both cases: manual and automatic. Designed with the audio recording sampling rate (i.e., 24 kHz), these signals are filled with 0 and 1 values such as:

$$\text{detection signal} = \begin{cases} 1 & \text{within manual annotations (cries) or} \\ & \text{audio automatic segmentation (sounds),} \\ 0 & \text{otherwise (silence).} \end{cases} \quad (3.3)$$

The annotated cries and segmented sounds are compared in terms of segment quantities and durations. Moreover, for consistency with the method, annotated segments with a duration lower than 0.25 second and greater than 5 seconds are not taken into account. The parameters used to evaluate the segmentation are described hereafter.

3.5.1 Segment comparison

To compare the manual annotation and the automatic segmentation in terms of segment quantity we define four segment parameters, which are illustrated in [Figure 3.19](#).

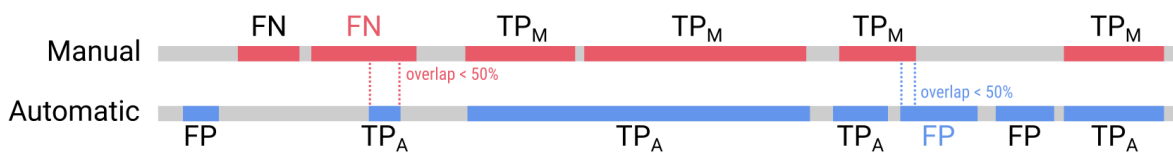


Figure 3.19: Illustration of the segments comparison parameters.

For each manually annotated segment, we evaluate the corresponding automatic detection signal. When at least 50% of the manual annotation samples are equal to 1 within the automatic segmentation signal, the audio segment is considered detected, otherwise it is not.

- True Positive Annotations (TP_M): the number of annotated segments overlapping one or more segment detected by the segmentation,
- False Negative (FN): the number of segments annotated but not detected.

The same process is symmetrically repeated for the segments resulting from the automatic segmentation. In that case, a segment is considered detected when at least 50% of the samples within the annotated detection signal are equal to 1, otherwise not.

- True Positive Segmentation (TP_A): the number of detected segments overlapping one or more annotated segment,
- False Positive (FP): the number of segments detected but not annotated.

Based on the previously defined segment parameters, we can describe the segmentation method performance through the sensibility and precision, defined as:

- Sensibility (S): percentage of annotated segments that have been detected through segmentation. It answers the question: “How much of the annotated segments were detected?”:

$$S = \frac{TP_M}{TP_M + FN} \quad (3.4)$$

- Precision (P): percentage of segmented sounds that are actual annotated cries. It answers the question: “How much of the extracted segments are really cries?”:

$$P = \frac{TP_A}{TP_A + FP} \quad (3.5)$$

3.5.2 Duration comparison

Manual annotation and automatic segmentation are compared in terms of duration through three parameters, which are illustrated in [Figure 3.20](#) and defined as:

- Δ_M - the manual annotation total time;
- Δ_A - the total time of segments resulting from the automatic segmentation.
- $\Delta_{M \cap A}$ - the total manual and automatic segmentations overlapping time.

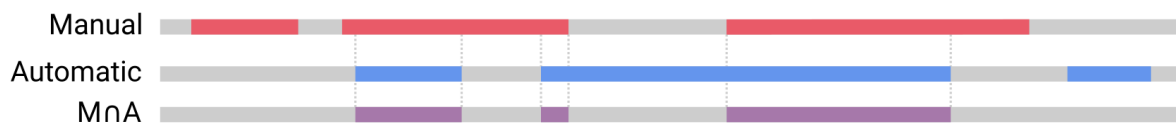


Figure 3.20: Illustration of the duration comparison parameters.

The total annotation or segmentation time is obtained by counting the number of samples equal to 1 in each detection signal while the overlapping time is computed by counting the number of samples equal to 2 in the signal resulting from the two detection signals summation. Then the length of the detected samples is converted to duration using the sampling frequency.

Once again, the segmentation method performance are assess in terms of sensibility and precision, defined for durations as:

- Sensibility (Δ_S): percentage of cries detected through the automatic segmentation:

$$S = \frac{\Delta_{M \cap A}}{\Delta_M} \quad (3.6)$$

- Precision (Δ_P): percentage of segmented sounds that are actual annotated cries:

$$P = \frac{\Delta_{M \cap A}}{\Delta_A} \quad (3.7)$$

3.6 Results

In this section, we deploy our segmentation method on three 30-minute audio files. After a description of the database, we evaluate the various improvements brought to the method initially proposed by Orlandi et al. [1] thanks to the parameters defined previously.

3.6.1 Database

To observe and evaluate the performance of this segmentation method, three 30-minute files were manually annotated using the Audacity software (see [Section 2.4.3](#) for exhaustive acoustic environment annotations). The start- and end-points of each audible cry event in the soundtrack were identified and the annotation boundaries were set at the point where the cries could no longer be heard. It is worthwhile to remind that only cry annotations whose duration is between 0.25 and 5 seconds are considered to be consistent with the segments derived from the automatic methods.

The three 30-minute sound files were selected from a 20-hour recording made for one baby¹. These files were selected for their sound event variety and show once again the acoustic environment variability in a single recording:

- WAV 1 - 01h25 - 23 sounds (0 cries and 23 non-cries);
- WAV 2 - 21h25 - 409 sounds (155 cries and 254 non-cries);
- WAV 3 - 21h55 - 1161 sounds (776 cries and 385 non-cries).

1. i.e., baby 010075 recorded during night time the 2018-02-20.

In the following, we propose to review the results derived from the different improvements provided to the reproduction of the method proposed by Orlandi et al. as well as the results derived from the proposed audio-video segmentation.

The results of the three annotated 30-minute WAV files processed by all steps and methods are presented in the [Table 3.1](#) in terms of segment and the [Table 3.2](#) in terms of duration.

The results are presented according to the methods used, whose acronyms are given below:

- **REP**: reproduction of the method proposed by Orlandi et al.;
- **NFB**: results obtained after applying the Narrowing Frequency Band solution;
- **NFB+RS**: results obtained after applying both, the Narrowing Frequency Band and Re-Segmentation solutions.
- **LTT**: results obtained after identifying audio files with enough sound content. True if more than 10 sounds are detected above the threshold T calculated over two-hour sliding windows, false otherwise.
- **AV**: results obtained with the proposed enhancements (i.e., NFB+RS+LTT) and after collecting the sounds included in infants' motion intervals only.

3.6.2 Audio segmentation improvements evaluation

Reproduction of Orlandi's method

With these results, we can justify the remarks made earlier in [Section 3.3.1](#) when discussing the issues encountered when applying the reproduction of Orlandi et al. method on our data.

First, we saw that Otsu's method does not work on recordings containing few sound events and no cry. This issue is illustrated by processing the file WAV 1 whose audio content is poor and where no less than 1977 segments are extracted.

Then, what can also be observed is the difference between the quantity of manually annotated cries and the number of segments resulting from the automatic segmentation. In the case of the WAV 3 file we can notice that 707 cries are automatically detected (TP_M) by the segmentation (out of the 776 that were manually annotated), however only 591 segments are extracted (TP_A). This means that the method gathers sounds in the resulting segments (i.e., 1 automatically extracted segment = n annotated segments).

		REP	NFB	NFB+RS	LTT	AV
WAV 1 - 01h25 $n_{\text{cries}} = 0$ $n_{\text{non-cries}} = 23$	TP _M	-	-	-		-
	FN	-	-	-		-
	TP _A	-	-	-	false	-
	FP	1977	23	23		0
	TOTAL	1977	23	23		0
	S P	- -	- -	- -		- -
WAV 2 - 21h25 $n_{\text{cries}} = 155$ $n_{\text{non-cries}} = 254$	TP _M	118	109	109		107
	FN	37	46	46		48
	TP _A	121	114	114	true	112
	FP	100	55	55		48
	TOTAL	221	169	169		160
	S P	76% 54%	70% 67%	70% 67%		69% 70%
WAV 3 - 21h55 $n_{\text{cries}} = 776$ $n_{\text{non-cries}} = 385$	TP _M	707	706	706		571
	FN	70	70	70		205
	TP _A	591	621	638	true	466
	FP	42	42	42		31
	TOTAL	633	663	680		497
	S P	91% 93%	91% 94%	91% 94%		74% 94%

Table 3.1: Segmentation comparison in terms of segments quantity.

		REP	NFB	NFB+RS	LTT	AV
WAV 1 - 01h25	Δ_M			0		
	Δ_A	897.81	9.69	9.69	false	0
	$\Delta_{M \cap A}$	-	-	-		-
	Δ_S	-	-	-		-
	Δ_P	-	-	-		-
	Δ_M			106.51		
WAV 2 - 21h25	Δ_A	131.77	101.86	101.86	true	97.36
	$\Delta_{M \cap A}$	74.56	73.17	73.17		71.60
	Δ_S	70%	69%	69%		67%
	Δ_P	57%	72%	72%		74%
	Δ_M			716.57		
	Δ_A	670.40	659.40	656.34	true	477.91
WAV 3 - 21h55	$\Delta_{M \cap A}$	620.80	620.80	618.81		447.99
	Δ_S	87%	87%	86%		63%
	Δ_P	92%	94%	94%		94%

Table 3.2: Sound segment durations comparison (in seconds).

Narrowing Frequency Band (NFB) improvement

As mentioned before, this step improves the threshold computation and helps to better take into account the acoustical environment. Its impact on the three files is detailed below.

- In WAV 1: the number of false detections (which are mostly noise) particularly decreased. Indeed, this step allows extracting 23 segments instead of the 1977 ones extracted with the reproduction method. In terms of duration, it is a matter of extracting 9.69 s instead of the 897.81 s previously extracted.
- In WAV 2: the number of detected segments that are not cries diminished (i.e., FP decreased from 100 to 55 segments). Moreover, we can also notice that the number of detected cries has slightly decreased without impacting the total duration of the extracted cries (i.e., Δ_S decreased from 70% to 69%).
- In WAV 3: the number of extracted segment increased (i.e., TP_A increased from 591 to 621 segments) without impacting the sensibility).

Re-Segmentation (RS) improvement

The re-segmentation step is applied to the resulting segments obtained through the NFB method. First, we saw in section **Section 2.4.3** that the annotated cries did not last more than a few seconds (see **Figure 2.17**). Then, we showed some issues related to the poor segmentation of cry bouts (several consecutive cries) and mixed sounds. Therefore, this is why the re-segmentation step proposes to re-cut the detected segments whose duration is longer than five seconds.

This step allows both to normalize the extracted segment duration and to better segment the sound events. This can be noticed with the WAV 3 file for which the amount of automatically extracted crying segments increased (i.e., TP_A increased from 621 to 638 segments) without affecting the total duration of the extracted cries (i.e., Δ_S decreased from 87% to 86%).

Long-term threshold (LTT)

As a reminder, we proposed to use this step to reduce the amount of data to be processed by ignoring recordings containing very few sounds.

In order to illustrate this procedure, we computed the thresholds for all 30-minute audio files of the recording performed on baby 010075 (i.e., 40 files). They are presented in **Figure 3.21a**. First, one can see the considerable variability with T_U , represented in grey, and T_L represented in red. These variations are due to the signal energy value distribution related to the acoustic environment. Thus, files containing minimal audio content have low thresholds (e.g., WAV 1) while files containing many sounds have higher thresholds (e.g., WAV 2 & WAV 3).

Then, to be able to detect WAV files with poor acoustic content, we suggested to use a sliding threshold T computed over up to a two-hour window. This threshold is illustrated with the black line in **Figure 3.21b** and we can note that it varies less abruptly than the T_U and T_L thresholds. Then, for each audio file, the number of intervals whose STE values exceed the corresponding threshold level T is estimated (such as performed to detect sound event intervals with STE values over T_U). Files with less than 10 intervals are represented by dots while the others are represented by squares. In this 20-hour recording example, seven WAV files are detected as files with minimal audio content (such as WAV 1) and 33 are further processed through the segmentation (such as WAV 2 and WAV 3). The three annotated WAV files used in this study are highlighted in blue. Therefore, when considering the whole recording including 40 WAV files (i.e., 14 166 sounds extracted with NFB+RS), the Long-term threshold step helps to reduce the file quantity to be processed to 33 (i.e., 11 435 sounds extracted with NFB+RS+LTT).

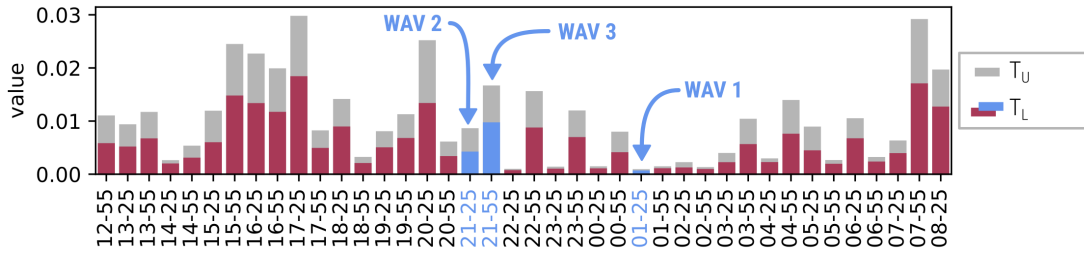
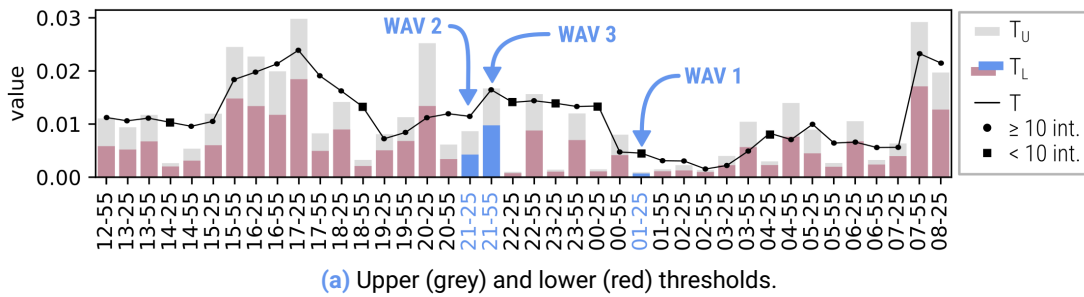


Figure 3.21: Thresholds computed for every 30-minute files over a 20-hours recording¹. The three annotated WAV files used in this study are highlighted in blue.

It should be mentioned that this step is presented in this results section after the NFB and RS steps to show the interest of these improvements on the three annotated files. However, in the processing chain it is applied before these steps.

Audio-Video segmentation

Regarding the results derived from the complete audio-video segmentation by collecting only the sounds occurring in infants' motion intervals, we can see that it reduced the number of extracted segments to be further processed without seriously affecting the total duration of the extracted cries. Indeed, when comparing the REP step with the AV step, we can see that the sensibility decreased in both files (i.e., Δ_S diminished from 70% to 67% for WAV 2 and from 87% to 63% for WAV 3) while the precision increased (i.e., Δ_P increased from 57% to 74% for WAV 2 and from 92% to 94% for WAV 3).

We remind that it is normal that the sensitivity (i.e., the number of extracted cries) decreases since we do not collect sounds contained in NaN intervals (i.e., adult presence and/or baby absence) in which cries can occur.

3.7 Conclusion

We have seen that most of the studies conducted in the literature concern recordings made in specific, non-noisy environments. Thus, the usual pre-processing step called "cry segmentation" cannot be used in our case since the recordings studied contain many other sounds than infant cries. This is due to the NICU environment which hosts care activities and machines required to help premature or sick newborns. Therefore, we proposed a two-step crying extraction method. While the first step segment all the sounds occurring in the signal, the second step will classify the extracted sounds and detect those containing crying.

In this chapter, we have presented the first step, which allows extracting sounds from background noise. This approach is based on the one originally proposed by Orlandi et al. in [1] and improvements were proposed to better process our database. Thus, after removing 30-minute audio file containing a poor audio content, we included a frequency filtering as well as a re-segmentation steps. These enhancements were applied consecutively to three different 30-minute annotated files and we compared the results in terms of segment and duration. We showed that the method is relevant for cry extraction and also helps to reduce the amount of data to be further processed (i.e., sensibility greater than 60% and precision greater than 70%).

Finally, since motion segmentation was also performed during the European Digi-NewB project, we proposed to extract exclusively the sounds occurring within infant's movements intervals. The relevance of this strategy, which naturally suggests that a baby is moving when crying, was confirmed in a preliminary study with an evaluation on a large database including 243 hours from 36 recordings of 22 newborns (see [Section 3.4](#)). First, we showed that infants are most of the time immobile (i.e., 53% on average) and that they don't cry during those periods. However, the cries may be produced in motion intervals or in intervals with adult presence and/or baby absence. However, in the latter case, it is more complicated to process the audio recordings since

it correspond to care periods in which many other sounds are produced. Yet, to minimize and facilitate at most the process, we decided limit the audio segmentation within motion intervals. Considering the studied database, this strategy led to discarding 87% of the sound segment total duration initially extracted.

The use of motion for audio segmentation is however not a mandatory step and has some limitations due to interval detection errors. Nevertheless, this strategy was never performed before and seems relevant for our data processing.

At this stage, all sounds whose energy is included in the newborn's fundamental frequency band are extracted. Thus, the next step consists in classifying these sounds to detect those containing cries. This is the purpose of the following chapter which proposes a binary classifier by a Deep-Learning approach using spectrograms.

APPENDIX A - OTSU'S METHOD

Otsu's thresholding concept is coming from image processing and is used to binarize an image based on pixel intensities. In other words, it manages to convert an image composed of several gray levels into black and white, such as illustrated in [Figure 3.22](#).



Figure 3.22: Otsu's method application.

To do so, the algorithm assumes that the picture is composed of two classes and tries separating the foreground pixels from the background ones. The optimal threshold is determined by minimizing intra-class intensity variance (defined as a weighted sum of the two classes' variances) or equivalently, by maximizing inter-class variance σ_B^2 defined as:

$$\sigma_B^2(t) = \omega_0(t)\omega_1(t) [\mu_0(t) - \mu_1(t)]^2$$

where weights ω_0 and ω_1 are the two class probabilities and μ_0 and μ_1 are the means of these classes separated by the threshold t .

Therefore, by computing iteratively inter-class variance through all possible thresholds based on the image pixel distribution (i.e., histogram), it is possible to determine the optimal threshold located where the inter-class variance is maximum, see [Figure 3.23](#).

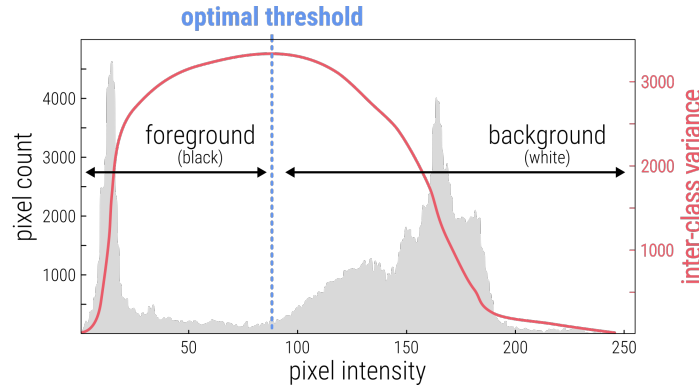


Figure 3.23: Optimal threshold computation through Otsu's method.

Computed with a recursion relation it permits fast calculation and gives an effective algorithm with the advantage of reduced processing time.

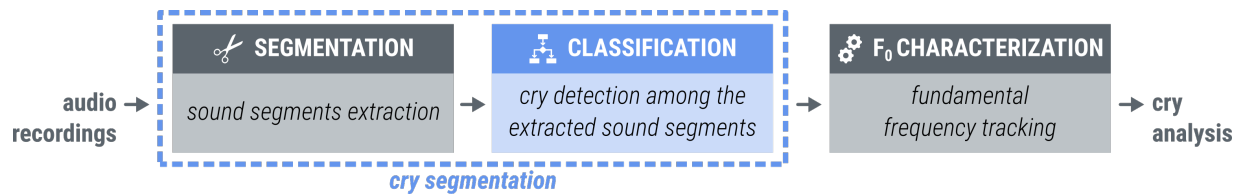
Since the procedure to determine an optimal threshold based on the global histogram properties is simple, automatic, and stable, the method was implemented for mono sound signals in [1].

BIBLIOGRAPHY

- [1] 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*, vol. 8, 799–810 (2013).
- [2] 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*, vol. 38, 207–213 (2017).
- [3] WEBER R., SIMON A., PORÉE F., AND CARRAULT G. Deep transfer learning for video-based detection of newborn presence in incubator. In *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2020, Montreal, QC, Canada, July 20-24, 2020*, 2147–2150. IEEE (2020).
- [4] 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*, vol. 7, 43–49 (2012).
- [5] ORLANDI S., BOCCHI L., DONZELLI G., AND MANFREDI C. Central blood oxygen saturation vs crying in preterm newborns. *Biomedical Signal Processing and Control*, vol. 7, 88–92 (2012).
- [6] 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*, vol. 20, 85–99 (2015).
- [7] 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*, vol. 30, 656–663 (2016).
- [8] 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*, vol. 45, 174–181 (2018).
- [9] VÁRALLYAY G. Future prospects of the application of the infant cry in the medicine. *Periodica Polytechnica Electrical Engineering*, vol. 50, 47–62 (2006).
- [10] VÁRALLYAY G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, vol. 71, 1699–1708 (2007).
- [11] 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*, 2953–2956. IEEE (2012).
- [12] 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*, vol. 10, page 86 (2013).
- [13] 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*, vol. 56, 1416–1428 (2013).
- [14] 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*, vol. 19, 35–43 (2015).
- [15] ABOU-ABBAS L., TADJ C., GARGOUR C., AND MONTAZERI L. Expiratory and inspiratory cries detection using different signals' decomposition techniques. *Journal of Voice*, vol. 31, 259–e13 (2017).

- [16] 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*, vol. 2018, 1–14 (2018).
- [17] 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*, 1–5. IEEE (2016).
- [18] TORRES R., BATTAGLINO D., AND LEPAULOUX L. Baby cry sound detection: A comparison of hand crafted features and deep learning approach. In G. Boracchi, L. Iliadis, C. Jayne, and A. Likas, editors, *Engineering Applications of Neural Networks*, 168–179. Springer International Publishing (2017).
- [19] XIE J., LONG X., OTTE R., AND SHAN C. Convolutional neural networks for audio-based continuous infant cry monitoring at home. *IEEE Sensors Journal*, vol. 21, 27 710–27 717 (2021).
- [20] COHEN R., RUINSKIY D., ZICKFELD J., IJZERMAN H., LAVNER YIZHAR" E.W., AND CHEN S.M. *Baby Cry Detection: Deep Learning and Classical Approaches*, In *Development and Analysis of Deep Learning Architectures*, 171–196. Springer International Publishing, Cham (2020).
- [21] FERRETTI D., SEVERINI M., PRINCIPI E., CENCI A., AND SQUARTINI S. Infant cry detection in adverse acoustic environments by using deep neural networks. In *26th European Signal Processing Conference, EUSIPCO 2018*. European Signal Processing Conference, EUSIPCO (2018).
- [22] SEVERINI M., PRINCIPI E., CORNELL S., GABRIELLI L., AND SQUARTINI S. Who cried when: Infant cry diarization with dilated fully-convolutional neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (2020).
- [23] XIE J. *Baby cry detection based on audio signals using deep neural networks*. Master's thesis, Eindhoven University of Technology, Eindhoven, Netherlands (2019).
- [24] SEVERINI M., FERRETTI D., PRINCIPI E., AND SQUARTINI S. Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation. *IEEE Access*, vol. 7, 51 982–51 993 (2019).
- [25] MORELLI M.S., ORLANDI S., AND MANFREDI C. Biovoice: A multipurpose tool for voice analysis. *Biomedical Signal Processing and Control*, vol. 64, page 102302 (2021).
- [26] OTSU N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, 62–66 (1979).
- [27] LESTER B. AND LAGASSE L. Crying. In *Social and Emotional Development in Infancy and Early Childhood*, 80–90. Elsevier (2009).
- [28] ROTHGÄNGER H. Analysis of the sounds of the child in the first year of age and a comparison to the language. *Early Human Development*, vol. 75, 55–69 (2003).
- [29] 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. *Biomed. Signal Process. Control.*, vol. 52, 362–370 (2019).
- [30] BIK A., SAM C., DE GROOT E.R., VISSER S.S., WANG X., TATARANNO M.L., BENDERS M.J., VAN DEN HOOGEN A., AND DUDINK J. A scoping review of behavioral sleep stage classification methods for preterm infants. *Sleep Medicine*, vol. 90, 74–82 (2022).
- [31] CABON S. *Monitoring of premature newborns by video and audio analyses*. Master of science thesis, Université de Rennes 1 (2019).
- [32] PRECHTL H.F. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Human Development*, vol. 23, 151–8 (1990).

Classification for cry detection



4.1 Introduction

The previous chapter has demonstrated that the audio segmentation is not sufficient to extract cries in recordings performed in a noisy environment such as the NICU. Whatever the efficiency of the segmentation step, due to the real context with many sound sources, the resulting audio segments are not just crying but also voices, alarms, etc. Therefore, it is of great interest to classify the segments derived from the previous chapter to detect the ones containing cries.

In this chapter, we propose a framework based on a feature learning scheme powered by a pre-trained discriminative Convolution Neural Network (CNN) using spectrograms. After a review of the state of the art of methods investigated in the literature, we present the main components of the proposed framework. Then, we introduce our two-step training strategy to fine-tune some of the model hyperparameters. Finally, since supervised neural network approaches expect dedicated training and testing sets of annotated data, we introduce the annotation software created to design such a database.

4.2 State of the art

As mentioned in **Chapter 1**, studies have shown that important information related to infant health status, emotions, and needs can be interpreted by analyzing the acoustics of infant crying. Therefore, in the last decades, many studies have focused on the detection or classification of infant vocalizations. In fact, the different approaches can be divided into four categories which are described below.

- **Pathology detection** which is a binary classification task where a cry is classified as normal or pathological [1, 2].
- **Pathology identification** which aims to determine the type of pathology the infant is suffering from. It has been used, for example to detect deaf newborns [1, 3, 4], those who have suffered from perinatal asphyxia [5], both conditions [6, 7], hypothyroidism [8, 9] or even cleft palate [10].
- **Crying cause identification** which aims to discover the reason that triggered the cry, for example, hunger, pain, sleep, or many other causes [11–15].
- **Crying detection** which consists in identifying crying in the signal, either by determining the temporal limits of vocalization when processing the entire audio signal or by determining the presence of crying in segmented sounds.

In this work, we are interested in this last application where the goal is to detect the infant cry signal efficiently and accurately in a noisy environment. Studies that address this topic investigated data recorded over a long time, either at home to develop systems to detect crying and alert parents [16–20] or in hospital. In the latter case, crying detection is performed to investigate infants' reaction to auditory stimuli of the NICU environment [21], to quantify the amount of time an infant cries [22] or to serve as a pre-processing stage for deeper analysis (i.e., related to pathology or cause identification) [23–27].

A cry detection system is usually composed of two steps: *i*) a pre-processing step that extracts the most suitable features from sound signals and *ii*) a classifier to recognize the cry features in an audio signal. This section provides an overview of existing feature extraction methods and classification strategies (see [28] for an extensive review).

4.2.1 Feature extraction

The challenge in cry detection systems is to select acoustic features that allow clear discrimination between a cry and other sounds. As mentioned in **Chapter 1**, the acoustic and prosodic characteristics of crying signals are often studied in time and frequency domains. However, it is the combination of the two, i.e., the time-frequency, that is most interesting. This domain involves dividing the sound signals into several small chunks called frames and constructing a feature vector for each frame. Thus, it allows following the variations of the frequency characteristics of the signal as a function of time. The most commonly used features are cepstral coefficients (MFCCs and LPCCs), wavelet transforms and Fourier transforms (illustrated in **Figure 4.1**).

Cepstral coefficients are widely used in the literature for audio signal processing and are recognized as performing well for tasks such as speech recognition or music genre classification. In particular, MFCCs represent the short-term power spectrum of an audio clip based on the discrete cosine transform of the logarithmic power spectrum on a non-linear Mel scale.

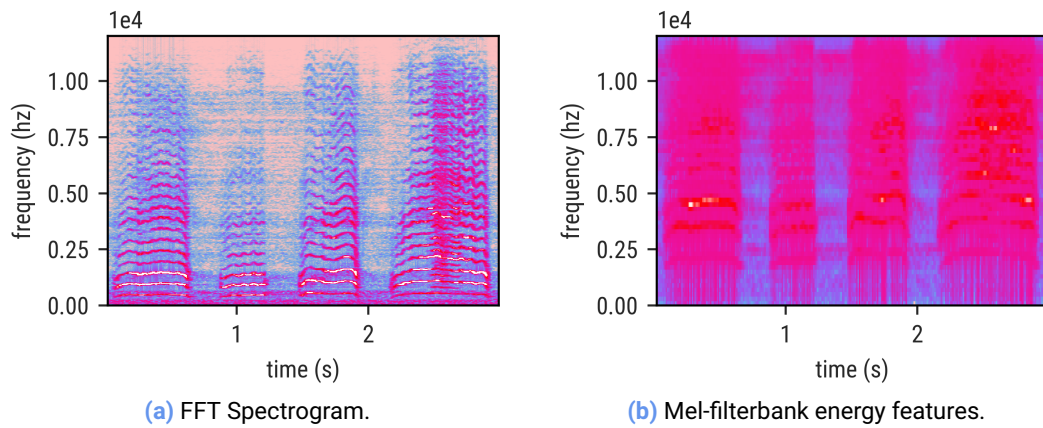


Figure 4.1: Illustration of the usual audio features used in sound classification methods. In both examples, four cry harmonic sequences are separated by unvoiced breath of the baby which produces noisy-like sounds in the lower frequency.

Thus, the frequency bands are equally spaced on the mel scale, which very closely mimics the human auditory system, making MFCCs a key feature in various audio crying detection systems, especially when associated with either the Mel-filter bank [16, 17, 19, 22, 24–26] or linear-filter bank [18]. For their part, Abou-Abbas et al. proposed a solution with empirical mode decomposition and MFCCs associated with Mel-filter bank [23], while a previous work of our team proposed harmonic plus noise modeling before computing MFCCs and associated them with temporal parameters and modeling parameters. The feature set dimensionality was then reduced using principal component analysis [27].

Otherwise many studies were based on Fourier transforms. Indeed, spectrograms are commonly used to represent the whole audio signal spectral decomposition over time. More precisely, a spectrogram is a two-dimensional image in which the x-axis represents time and the y-axis represents frequency. Depending on the brightness of the image, we can observe the energy level of different frequencies as a function of time. Thus the brighter an area is, the more excited the corresponding frequency at that time. In particular, spectrograms were used for infant cry classification [12, 14, 15], infant cry detection [20], infant speech recognition [13], and also in other domains such as bird species identification [29, 30], or even in speech emotion recognition [31–34].

In both cases, i.e., using MFCCs or spectrogram, authors had to calculate the Fourier transform. Hence, the reported windowing parameter values in the literature are: *i*) frames of 20 ms with 50% overlap [22, 25, 26]; *ii*) frames of 25 ms and Hamming window with 50% overlap [24] or Hanning window [34]; *iii*) frames of 30 ms with an 10 ms step [21] or a 21 ms overlap [23]; *iv*) other parameter such as frames of 23.2 ms and 50% overlapping with Hamming windows [21], frames of 32 ms with 50% overlap [17] and frames of 1.5 s with 0.5 s overlap [18]. Therefore, there is no consensus about the window parameters to use when computing the Fourier Transform to date.

4.2.2 Classification methods

As mentioned, only a few studies have addressed the problem of crying detection in a real-life context (i.e., home or NICU) and all of them propose different solutions. Some authors investigated traditional machines learning methods such as hidden Markov model [21], Gaussian mixture model [24], a combination of both [23] or even K-nearest neighbor [27]. In recent years, much research on deep learning has been conducted in image and speech recognition with results sometimes surpassing classical methods. Thus, researchers suggested neural network algorithms based on Convolutional Neural Network (CNN) [15–20] or proposed their own deep neural network architecture [22, 25, 26].

Although the reported results appear reliable, strong limitations regarding the representativeness of the training and assessment datasets prevent them from being considered sufficiently robust for deployment in the clinic [15]. Indeed, recordings can be severely affected by various sound sources from the surroundings, in particular in the NICU where those sounds are very diversified due to the type of room/bed and the required medical equipment [35]. Hence, the previous studies have worked with limited short audio recordings, recording environments, and ranges of neonatal PMA and GA. To overcome the lack of real-world recorded data, Ferretti et al. generated simulated data to improve their deep neural network training model [22]. However, the solution remains limited, since only one room was simulated and the final model was tested on only a few 30-second sequences of real-world data from a single newborn. Furthermore, in the previous work of our team, although a wide variety of real-world sounds were included, limitations regarding the ability of our approach to characterize high-frequency sounds were raised [27].

Thus, a method that takes into account all NICU challenges has yet to be proposed. Such a solution will provide a robust continuous monitoring tool to improve newborn health care through crying analysis.

4.2.3 Evaluation metrics

Performance evaluation is an important aspect of the machine learning process. Metrics are mandatory to compare the results of the different trained models. Moreover, depending on the classification objectives, attention may be focused on different metrics. Therefore, in this section, we review the ones used later in this chapter to assess the model performance. While some of the definitions have already been described in **Chapter 3**, they are reported here in the context of a binary classification for the purpose of cry detection (i.e., the two classes being *cry* and *non-cry*).

CONFUSION MATRIX - is a very practical tool used to present the performance of a supervised learning algorithm. It quantifies the number of correct and incorrect classifications by comparing predictions to actual labels. A confusion matrix for a binary classifier is reported in **Table 4.1**. It

is composed of four numbers described hereafter.

- *True Positive* (TP): number of samples accurately predicted as *cry*;
- *True Negative* (TN): number of samples accurately predicted as *non-cry*;
- *False Positive* (FP): number of samples predicted as *cry* instead of *non-cry*;
- *False Negative* (FN): number of samples predicted as *non-cry* instead of *cry*.

All of the following evaluation parameters are calculated based on the confusion matrix and their values are between 0 (i.e., worst) and 1 (i.e., perfect).

		PREDICTIONS	
		<i>cry</i>	<i>non-cry</i>
ACTUAL	<i>cry</i>	True Positive	False Negative
	<i>non-cry</i>	False Positive	True Negative

Table 4.1: Confusion matrix for a binary classifier.

SENSITIVITY (S_e) OR RECALL (R) describes how well cry sounds were classified as *cry*. It answers the question: "How much of the actual cries were correctly classified ?":

$$S_e = \frac{TP}{TP + FN} \quad (4.1)$$

SPECIFICITY (S_p) describes how well actual non-cry sounds were correctly classified as *non-cry*. It answers the question: "How much non-cry sounds were correctly classified ?":

$$S_p = \frac{TN}{TN + FP} \quad (4.2)$$

BALANCED ACCURACY (BAcc) describes how good the classifier is in predicting if a sound belongs to the cry or non-cry class. It is especially useful when the classes are unbalanced and it is computed as the arithmetic mean of the two previous metrics:

$$BAcc = \frac{S_e + S_p}{2} \quad (4.3)$$

PRECISION (P) describes how well sounds detected as *cry* were actual cries. It answers the question: "How much of non-cry is classified as cry?":

$$P = \frac{TP}{TP + FP} \quad (4.4)$$

F1-SCORE (F_1) describes also how good the classifier is in predicting if a sound belongs to the cry or non-cry class. It is computed as the harmonic mean of the model's precision and recall.

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (4.5)$$

4.3 Proposed method

In this work, we propose an infant cry detector using a framework based on a feature learning scheme powered by a pre-trained discriminative CNN using spectrograms. The latter are computed on sounds collected in the NICU and derived from the segmentation method (see [Chapter 3](#)). Thus to identify cries in all the resulting audio segments we chose the spectrogram feature for its efficiency to represent a wide spectral decomposition in time. The final framework is a binary classifier composed of the two classes: *cry* and *non-cry*. The classification is performed in four steps illustrated in [Figure 4.2](#) and described hereafter:

- for each extracted sound, the spectrogram is computed by a short term Fourier transform;
- since the extracted sounds have variable durations, the resulting spectrogram is cut into frames of the same duration;
- these images are used in the input of a convolutional neural network;
- for each initial extracted sound, the decision taken is the majority prediction on all the images.

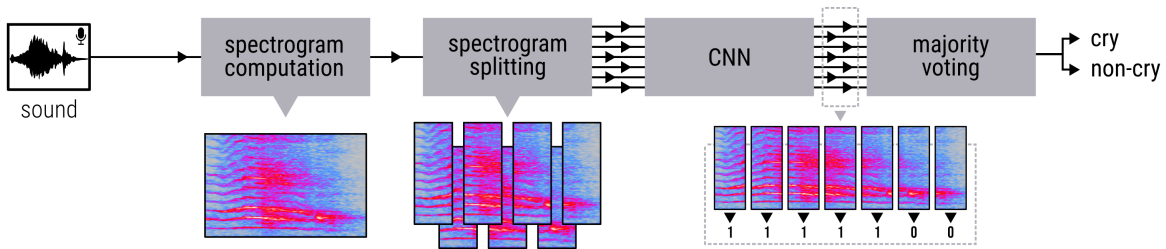


Figure 4.2: Binary cry classification framework based on CNN using spectrograms.

4.3.1 Spectrogram computation

First, spectrograms were computed for each sound file using Short-Time Fourier Transform (STFT) of successive 0.04 ms long (1000 samples) Hamming-windowed frames with 95% overlap. Since signals have a sample rate of 24 kHz, the configuration provides a frequency resolution of 23.4 Hz (ranging from 0 to 12 kHz) and a time resolution of 4.2 ms (illustrated in [Figure 4.3](#)). To have a good image contrast, the magnitude of the spectrogram is converted to a logarithmic scale and an image quantization of 256 levels is performed on a fixed colormap.

Then, the spectrograms were divided into several smaller spectrograms of the same size with a 50% overlap such as proposed in speech emotion recognition in [\[31\]](#). We named a small spectrogram: *frame* and the set of small spectrograms extracted from the big one: *frame group*. Therefore, the number of frames depends on the cry duration. Before the split, the frame group is centered on the spectrogram since the acoustic characteristics are generally more interesting in the middle of it.

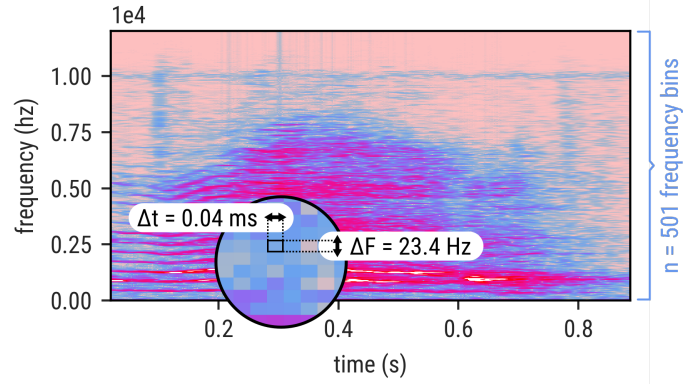


Figure 4.3: Resulting spectrogram with temporal and frequency resolutions for cry unit sampled at 24kHz.

This spectrogram division serves two purposes, on the one hand, it normalizes the size of the images at the input of the CNN (since the sounds all have different durations), on the other hand, it increases the number of spectrograms which allows us to design a powerful model. The spectrogram division process is illustrated with 0.20-second duration frames by a scheme of the frame group centered in [Figure 4.4a](#) and by the resulting frames numbered in [Figure 4.4b](#). Finally, each frame is saved with a resolution of 224x224 pixels with three channels (RGB).

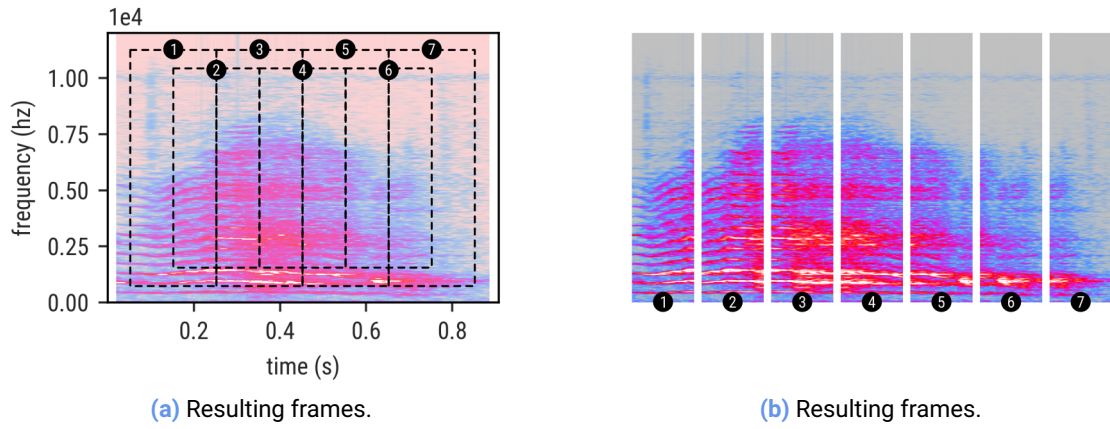


Figure 4.4: Illustration of the spectrogram division process with 0.20s duration frames.

Since there is no consensus in the literature about the Fourier Transform windowing parameters, we decided to explore the spectrogram division with two different frame durations:

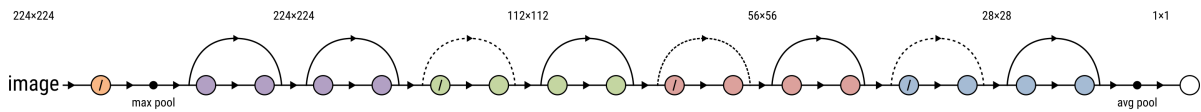
- 0.20 s which corresponds to the most common value found in the literature (see [Section 4.2.1](#)).
- 0.25 s which is the minimum duration of the sound segments resulting from the segmentation (see [Chapter 3](#)).

4.3.2 Transfer learning using ResNet architectures

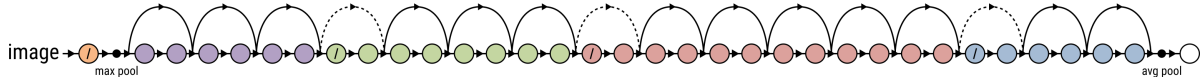
Image classification tasks have seen breakthroughs in terms of performance, thanks to the rise of CNN. These networks are composed of a sequence of filters on the raw pixel data of an image to extract and learn high-level features. The analysis of the visual field is done through a set of overlapping sub-regions, this is called convolutional processing. The model then uses the extracted features to perform classification. The three main components of a CNN are convolutional, pooling, and fully connected layers. These layers are usually arranged in the form of a hierarchy where one can use any number of convolutional layers followed by pooling layers and at the end fully-connected layers. This type of architecture is defined by the number of layers in each component as well as the connections between them. For its part, the ResNet architecture was firstly proposed in 2015 to overcome the issues encountered when using a large number of layers [36]. Indeed, it solved the problem of the vanishing gradient by introducing the concept called Residual Network (ResNet). This technique, which allows skipping connections of a few layers, showed convincing performance in many computer vision applications and is now widely used for image classification.

In light of the ResNet performance, we decided to perform learning by transfer. This principle consists in reusing convolutional neural networks previously trained on a large image database. Hence, the ResNet weights were pre-trained with ImageNet to initialize the classification model [37]. Then weights are optimized to our task (i.e., the crying vs. non-crying classification), by performing a new training through the last fully-connected layer. In our case, this step aims to minimize the cross-entropy loss associated with a class-weighting. In addition, we decided to explore two network depths using the Resnet18 and Resnet34 architectures illustrated in Figure 4.5. This approach may appear unusual since it uses natural images to classify sound spectrogram. However, it should be noted that this particular strategy was used for the identification of crying cause in [15] and that the strategy itself was experimentally verified in [38].

ResNet - 18 Architecture



ResNet - 34 Architecture



Legend:
 — connection
 skip connection (i.e. conv)
 7x7 conv, 64, 2 3x3 conv, 64, 1 3x3 conv, 128, 2 3x3 conv, 128, 1 3x3 conv, 256, 2 3x3 conv, 256, 1 3x3 conv, 512, 2 3x3 conv, 512, 1 fc, 2

Figure 4.5: Residual Network or ResNet architectures used in this work. All layers are described with the convolution kernel, the number of output channels and the stride value except the last one which is a fully connected layer with one predicted class in output: cry or non-cry.

The input data for the CNN are the spectrogram images resized to the shape 224x224 pixels. To adapt the model to our data, some parameters were fixed (see [Table 4.2](#)) while the following parameters have been optimized:

- the spectrogram division with frame durations of 0.20 or 0.25 s ;
- the depth of the neural network through ResNet architectures (i.e., 18 or 34);
- the learning rate : from 10^{-2} to 10^{-5} .

HYPERPARAMETER	CHOICE
Cost function	cross entropy
Optimization algorithm	stochastic gradient descent
Learning rate scheduler	standard decay
Momentum of the optimizer	0.9
Regularization by weight decay	$5 \cdot 10^{-5}$
Regularization by batch learning	16
Number of learning iterations (epoch)	200
Class imbalance management (class weighting according to data distribution, see Section 4.5.1)	non-cry 0.66, cry 0.33

Table 4.2: Fixed hyperparameters used to train the last layer of the CNN.

Moreover, since the inputs of the CNN are spectrogram frames, the final sound prediction (P_{sound}) is computed based on the distribution of the frame predictions ([Figure 4.6](#)) such as:

$$P_{\text{sound}} = \begin{cases} 1 & \text{i.e., cry if the frame majority decision is 1 or balanced,} \\ 0 & \text{i.e., non-cry if the frame majority decision is 0.} \end{cases} \quad (4.6)$$

To limit the number of calculations, we carried out a two-step parameter optimization strategy with defined combinations of parameters which are explained in the next section.

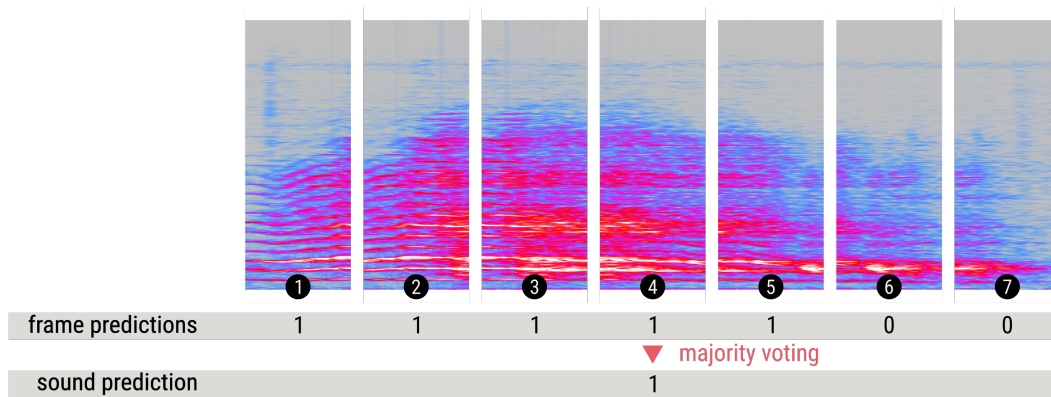


Figure 4.6: Sound prediction reconstruction using majority voting.

4.3.3 Model training

As mentioned above, the parameter optimization was done in two steps. For this purpose, we defined four combinations grouping the frame duration (i.e., 0.20 and 0.25 s) and network depth (i.e., ResNet18 and ResNet34) parameters. The four combinations assessed during this model training steps are given in [Table 4.3](#) with their designations, the step they have been optimized, as well as their parameter values.

The first training is used to identify the learning rates giving the the highest precision for each of the four combinations. Then, using these learning rates, the best combination, i.e. the best model, was identified by 5-folds cross-validation. The best combination is defined with the best average precision because we want to maximize the number of true positives in the classifier output (i.e., to ensure that the sounds predicted as cries are actual cries).

DESIGNATION	BEST CANDIDATE COMBINATIONS		FINAL COMBINATION	
	COMBINATIONS SELECTION		SELECTION	
	LEARNING-RATE		FRAME DUR.	RESNET DEPTH
W020_RESNET18	$10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$		0.20s	18
W020_RESNET34				34
W025_RESNET18			0.25s	18
W025_RESNET34				34

Table 4.3: Definition of the assessed combinations.

All models are trained for a maximum of 200 epochs. However, to reduce the calculation time, three thresholds were set up and the training is automatically stopped when:

- the loss value is not improving for 5 consecutive epochs;
- the differences of the 5 consecutive epochs are less than 10^{-3} .
- the loss value is less than 10^{-5} .

Best candidate combinations selection

This step is used to limit the number of calculations. Therefore, we compare 16 models corresponding to the four defined combination (i.e., frame durations and network depths) trained with the following learning rates: 10^{-2} , 10^{-3} , 10^{-4} , and 10^{-5} . Through a simple train/validation strategy, we want to identify, the learning rates associated with the highest precision achieved on the validation set for each combination (i.e., W020_RESNET18, W020_RESNET34, W025_RESNET18, and W025_RESNET34). For that purpose, the 16 models are trained with the same train/validation datasets which are presented later in this chapter.

Final combination selection

Once the learning rates are identified for each combination, we use 5-folds cross-validation to select the best model. Cross-validation is the most popular method used to detect problems such as under- or over-fitting and to ensure the robustness of the model. It is a resampling method that uses different parts of the data to test and train a model over different iterations.

In our case, the database is re-sampled into 5 folds, which, during iterations, are successively placed in the train or test sets, the process is depicted in [Figure 4.7](#).

Finally, the cross-validation is performed for the four combinations, and the one resulting with the least variation in performance and with the highest averaged precision is considered the best and final model.

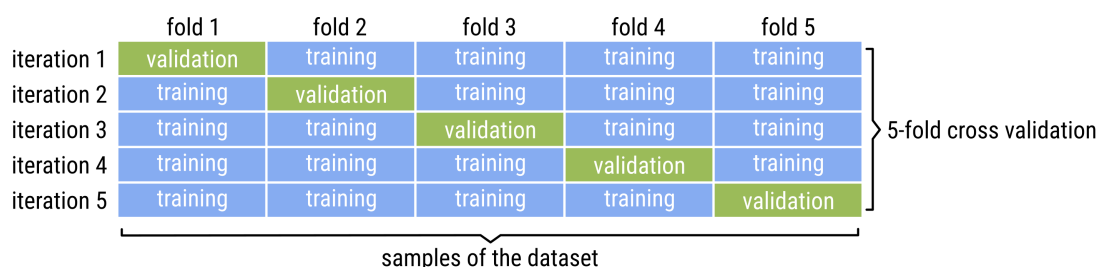


Figure 4.7: Illustration of the 5-folds cross-validation strategy.

4.4 SoundAnnoT: database creation

A convolutional neural network is a supervised type of Deep learning algorithm which means that annotated data are required to train and validate the model. In this section, we present the SoundAnnot software¹ that we specifically designed for annotating sound segments to create a training database for our network.

Thanks to SoundAnnoT it is possible to carry out simple and fast annotation of sound segments derived from the segmentation method with predefined labels. Annotations are performed only one time for each sound segment by a human through hearing and visual inspection of audios. Initially created to simplify the process of annotating sound events, SoundAnnot was then designed to allow non-expert users to handle it. Based on MATLAB software, it requires version R2018a or later ones. The main interface is depicted in [Figure 4.8](#).

1. SoundAnnoT - IDDN.FR.001.020001.000.S.P.2021.000.31230



Figure 4.8: SoundAnnoT user-friendly interface.

4.4.1 Interface

The software is composed of two main panels with the left one related to the audio information while the right one is dedicated to the annotations. In particular, SoundAnnoT is composed of the following components:

1. an audio player for listening to the current sound and navigating within the previous annotations (see Figure 4.9);
2. a spectrogram of the sound with a representation of the spectral components between 0 and 5000 Hz as they vary over time, useful for visual support;
3. an annotation panel for label selection of the sound currently played.

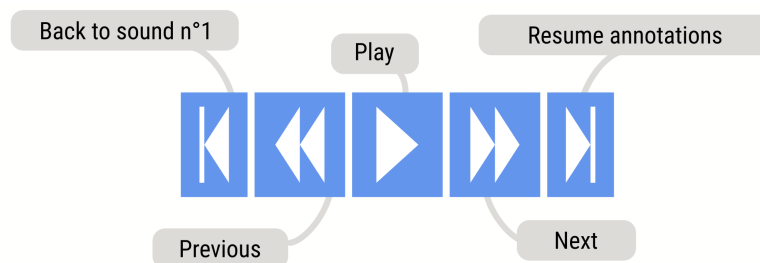


Figure 4.9: Functions of the audio player buttons.

4.4.2 Labels

As we have discussed in the previous chapter, the sound segments resulting from the segmentation step are not just cries but includes also many other sounds which occur in the NICU environment. Thus, seven labels were defined to annotate sounds including cries, cries with other sounds, other baby sounds, alarms, other human sounds, other sounds, and mixtures of non-cry sounds. Detailed descriptions of these categories are given in [Table 4.4](#). Although in this work we focused on infant crying, annotating subcategories can help to know better the sound environment and to analyze and understand what leads to incorrect detection.

Hence, the annotation panel is composed of two sub-panels corresponding to the two classes of the binary classifier:

- upper panel for sounds containing crying (*cry*) ;
- lower panel for sounds unrelated to crying (*non-cry*).

	LABEL	DESCRIPTION
cry	cry	pure cry sound
	cry+	cry sound mixed with another sound source: beep, voice, other ...
non-cry	baby others	sounds produced by the baby: moan, cough, ...
	alarms	short usually high-pitched sounds (from health electronic device) that serves as a signal or warning
	voices	sounds produced by nurses or parents when talking/whispering
	others	sounds produced by none of the previous mentioned categories: care procedures, tv, door, ...
	mixtures	mix of non-cry sounds

Table 4.4: Details of predefined labels in SoundAnnoT software.

4.4.3 Annotations

One annotation consists in assigning a label to a sound segment. When clicking on one of the annotation buttons, the following actions occur:

- the selected label is automatically saved;
- a new sound segment is automatically played and its spectrogram displayed.

In case a mistake has been done in label selection, annotation can be corrected by going back to the previous audio segment using the previous button in the audio player panel. Corrections are automatically saved when clicking on the new label.

In addition, when the sound source is not clear and it is difficult to choose the label, it is possible to click on the "?" button next to the label that seems most relevant. This allows us to take into account the fact that the annotator doubted while still requiring a label choice.

4.4.4 User procedure

To annotate a large number of sound segments, SoundAnnoT was provided to 10 non-expert volunteers who signed a confidentiality agreement. As the users were not familiar with the sounds occurring in the NICU environment, a procedure was set up (Figure 4.10) to guide them in order to obtain homogeneous annotations.

The software is composed of a homepage allowing user identification. Thus, any new user needs to complete a mandatory training phase before starting an official annotation session used to build the database. These two sessions are described below.

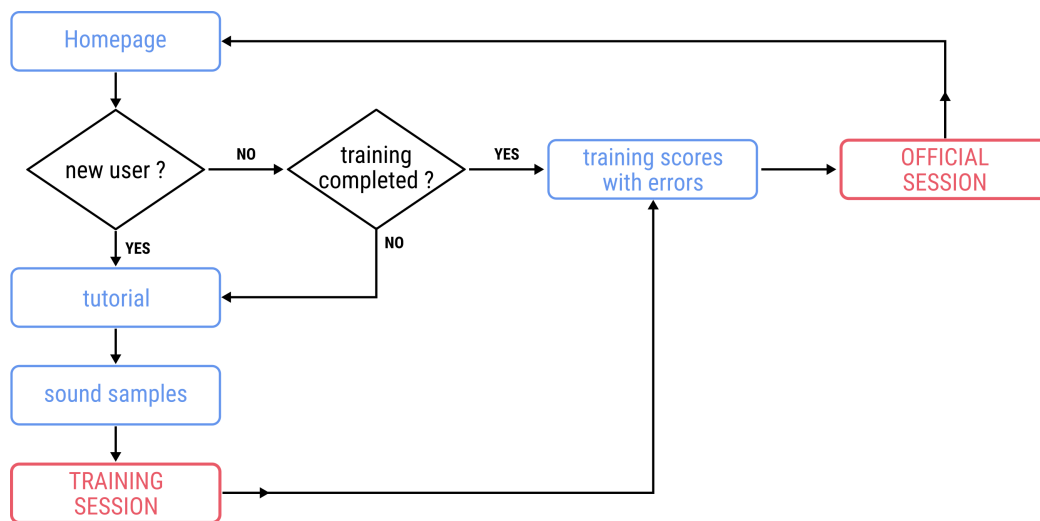


Figure 4.10: SoundAnnoT user procedure diagram.

Training session

The purpose of the training session is to familiarize the user with the software and the sound environment present in the NICU. When registering for the first time, the user goes through the following steps:

- a software tutorial, explaining the different components and how to annotate;
- sound samples to listen to for each of the defined label categories;
- the training session in which 100 sounds have already been annotated by an expert.

Once the user has annotated all the sounds, a score page allowing to listen again to the 100 sounds is displayed. Sounds whose annotations differ from the experts are highlighted in red, yellow, or blue depending on the severity of the error. An example of this page is provided in Figure 4.11.

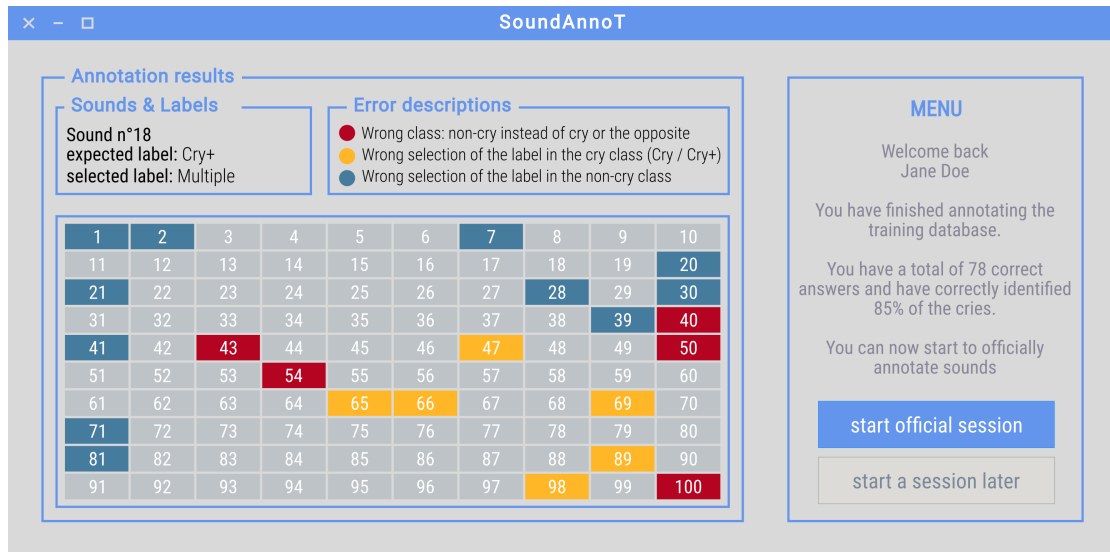


Figure 4.11: Annotation training score page.

Official session

In an official session, the user is invited to annotate consecutively 150 sounds chosen randomly from the database of unannotated sound segments. Each sound segment is annotated only once, however, the user can indicate uncertainty by clicking on the "?" button corresponding to the label that seems the most relevant (see the interface in Figure 4.8).

At any time the user can leave the session which is automatically saved. Once the session is closed, a statistics page is displayed with the duration and amount of annotations made during the current session as well as the overview of all previous sessions.

The database annotated thanks to the SoundAnnot software is described with the number of volunteers, babies, and sounds in the following section.

4.5 Results

This section presents the database annotated with SoundAnnoT software on a population of infants detailed in the first section. Then the learning strategies recently explained are evaluated. The first step aims to identify the best combination candidates while the second one aims to select the best final combination, i.e., the best model according to the averaged sound precision performance. At last, the chosen model is also assessed on a new cohort of infants never seen before.

4.5.1 Annotated data

We selected a dataset representing a large part of the diversity encountered in the framework of the European project Digi-NewB. Hence, 58 recordings were selected, performed in four hospitals: Rennes, Angers, Brest, and Tours, in both types of beds: open or close. They involve 20 boys and 13 girls born between 25+6 and 41+4 GA and recorded between 27+5 and 41+5 PMA. Some babies were recorded up to four different dates at least 48 hours apart. The data distribution for all babies is depicted in [Figure 4.12](#) with the dots representing the dates of the recordings used. In the top part of the figure, all recordings are merged in terms of GA where one can see the lack of recordings of babies born between 30 and 33 GA, and in terms of PMA where the distribution is quite homogeneous.

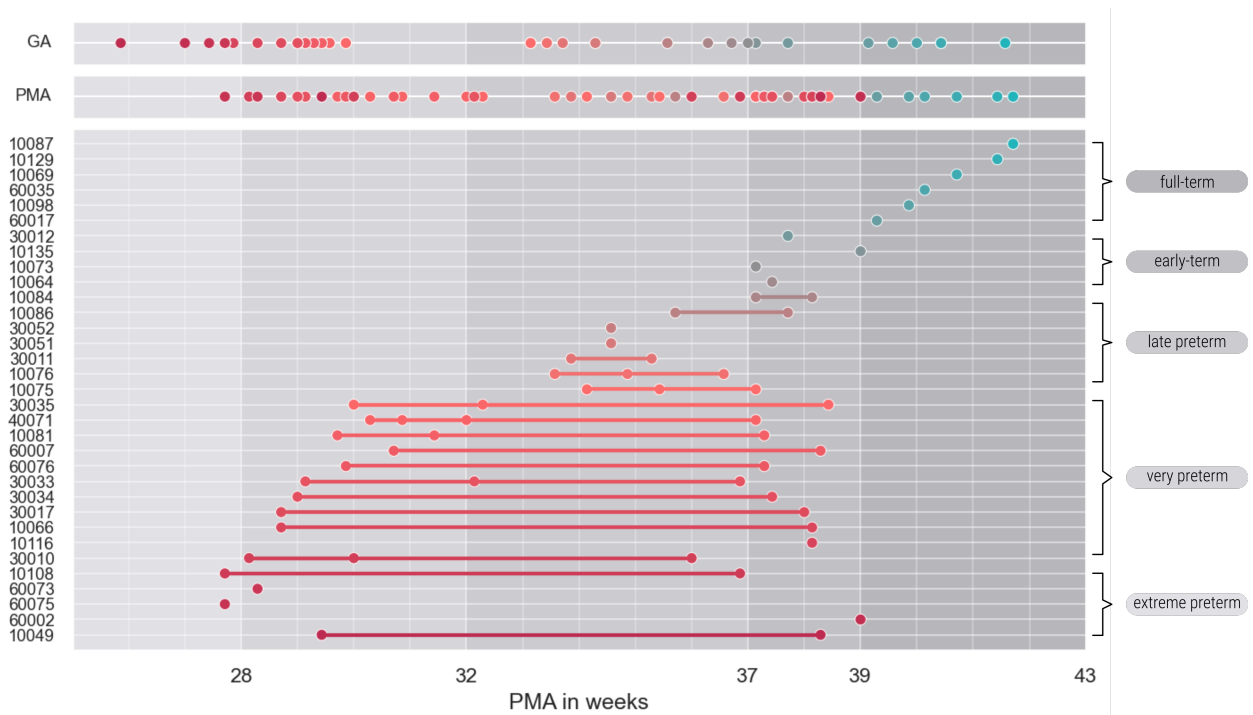


Figure 4.12: Data distribution used for sound annotation.

From each audio recording, sound segments were extracted using the segmentation method (see [Chapter 3](#)). Among the resulting audio segments, a total of 21 340 sounds were annotated by a cohort of 10 volunteers using the SoundAnnoT software. Therefore sounds were classified according to the seven labels defined previously in [Section 4.4.2](#). We mention here that sounds annotated with a doubt "?" were not taken into account in this database to avoid training the model with false labels.

In addition, for each annotated sound, spectrograms were computed and frames of 0.20 and 0.25 s duration were extracted and saved leading to the construction of two datasets including respectively 202 137 and 145 827 frames. Each frame was given the same label as the sound

it was originally extracted from. Thus, the sound and frame databases are detailed in terms of quantities and percentages according to the seven labels in [Table 4.5](#).

	CLASS	SOUNDS		FRAMES 0.20s		FRAMES 0.25s	
		QTY	PERC.	QTY	PERC.	QTY	PERC.
cry	cry	5476	25.66%	58 708	29.04%	42 194	28.93%
	cry+	1533	7.18%	26 238	12.98%	19 244	13.20%
non-cry	baby others	573	2.69%	2 051	1.01%	1 385	0.95%
	voices	119	0.56%	1 293	0.64%	927	0.64%
	alarms	444	2.08%	1 779	0.88%	1 254	0.86%
	others	12 476	58.46%	96 185	47.58%	69 042	47.35%
	mixtures	719	3.37%	15 883	7.86%	11 781	8.08%
TOTAL		21 340	100.00%	202 137	100.00%	145 827	100.00%

Table 4.5: Sound and frame databases annotated according to seven defined labels.

During the two training steps, we chose to ignore sounds with the *cry+* label (i.e., sounds of cries mixed with other sounds) to derive a confident model trained on pure cries. We believe that this strategy helps the model to learn the intrinsic cry characteristics. Thus, the binary classifier is composed of the *cry* class containing the sounds labeled *cry*, while the *non-cry* class merges the remaining sounds, i.e., those with labels: *baby others*, *voices*, *alarms*, *mixtures*, and *others*.

We divided the database into two sets. A training set composed of 30 babies is used to optimize the models through two training steps and a test set is used to ensure the generalization of the model. The data distribution of the two classes and the two sets is reported in [Table 4.6](#). In this table, one can see that there are more sound segments labeled as non-cry than cry. Since the dataset is imbalanced, we chose to use a weighted argument in the calculation of the cross-entropy loss with values corresponding to the data distribution, i.e., 0.66 for the non-cry and 0.33 for the cry class (see [Table 4.2](#)).

At last, the best combination is trained on all the data in the training set. Then, the trained model is deployed on the test set composed of sounds from three babies never seen before. To assess the good generalization of the model the test is performed twice, first excluding *cry+*, then including cries mixed with other sounds. The detailed dataset is also reported in [Table 4.6](#).

4.5.2 Best candidate combinations selection

The selection of the four best combinations is performed using a simple strategy with data divided into two sets: 29 babies used for the training and 1 baby for the validation. While the detailed training and validation sets are reported in [APPENDIX A](#), the metric performance (obtained based on the predictions on sounds resulting from the validation) are detailed for the 16 models in [Table 4.7](#).

		SOUNDS		FRAMES 0.20s		FRAMES 0.25s	
CLASS		QTY	PERC.	QTY	PERC.	QTY	PERC.
TRAIN	TRAINING SET (N=30)						
	cry	4 851	28.46%	53 312	32.68%	38 346	32.61%
	non-cry	12 191	71.54%	109 833	67.32%	79 259	67.39%
	TOTAL	17 042	100.00%	163 145	100.00%	117 605	100.00%
TEST	TEST SET (N=3)						
	cry	625	20.38%	5 396	29.21%	3 848	30.18%
	cry+	302	9.85%	5 119	27.71%	3 774	29.6
	non-cry	2 140	69.78%	7 358	43.08%	5 130	40.23%
	TOTAL	3 067	100.00%	18 473	100.00%	12 752	100.00%
		TOTAL					
		20 109	100.00%	181 618	100.00%	130 357	100.00%

Table 4.6: Subsets of the data used in this study.

	LR	BALANCED	PRECISION	RECALL	F ₁ -SCORE
		ACCURACY			
0.20s	ResNet18	10 ⁻²	0.8506	0.8049	0.7765
		10 ⁻³	0.8681	0.8500	0.8095
		10 ⁻⁴	0.9271	0.8163	0.8602
		10 ⁻⁵	0.9105	0.8444	0.8539
	ResNet34	10 ⁻²	0.8135	0.7692	0.7229
		10 ⁻³	0.8506	0.8049	0.7765
		10 ⁻⁴	0.9157	0.8125	0.8478
		10 ⁻⁵	0.8983	0.7755	0.8172
0.25s	ResNet18	10 ⁻²	0.8392	0.8000	0.7619
		10 ⁻³	0.8991	0.8409	0.8409
		10 ⁻⁴	0.9044	0.8085	0.8352
		10 ⁻⁵	0.9332	0.8511	0.8791
	ResNet34	10 ⁻²	0.8650	0.8293	0.8000
		10 ⁻³	0.9302	0.8333	0.8696
		10 ⁻⁴	0.8908	0.8571	0.8372
		10 ⁻⁵	0.9324	0.7885	0.8542

Table 4.7: Performance of the 16 candidate combinations based on sound predictions obtained on the validation set.

One can see that the four best candidate combinations with the associated learning rates giving the highest precision score have also high recall values ranging from 77% up to 90%. The best combinations retained for the following training step, compared in [Figure 4.13](#), are:

- W020_RESNET18 with a learning rate of 10^{-3} and a precision of 85%,
- W020_RESNET34 with a learning rate of 10^{-4} and a precision of 81%,
- W025_RESNET18 with a learning rate of 10^{-5} and a precision of 85%,
- W025_RESNET34 with a learning rate of 10^{-4} and a precision of 86%.

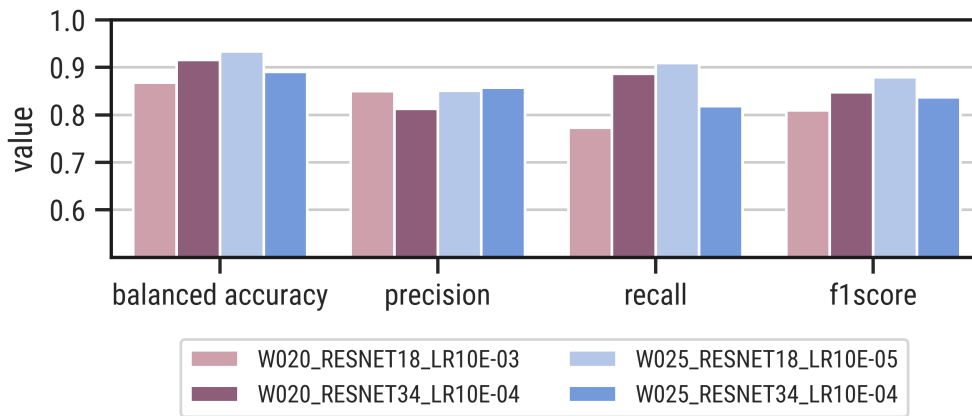


Figure 4.13: Performance of the four best candidate combinations based on sound predictions obtained on the validation set.

4.5.3 Final combination selection

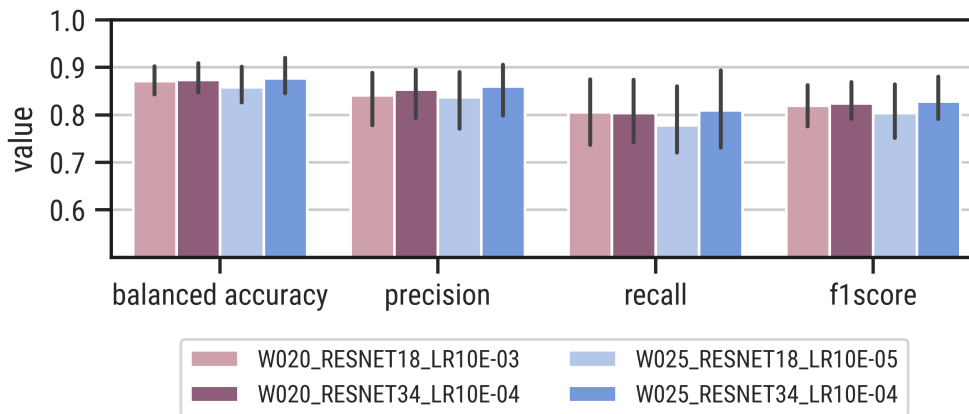
Once the four best candidate combinations are identified, the second step of training is performed on the cohort of 30 babies divided into 5-folds of six babies. The distribution received particular attention so that in each set there is a variety of centers, types of beds, PMA, and GA so that there is a good balance in the amount of available data ([APPENDIX B](#)). Results of the averaged 5-fold cross-validation are presented with numerical results in [Figure 4.14a](#) and illustrated in [Figure 4.14b](#).

The model with the highest averaged precision score is the one using the database with spectrograms framed over 0.25 s, using a ResNet34 architecture and an initial learning rate of 10^{-4} .

From the results, one can see the high prediction score of 86% meaning most of the sounds predicted in the *cry* class were actual cries. Moreover, it can be noticed that the recall value is also quite good (sensitivity of 81%), which proves that only a few cries will not be detected. Hence, this model meets our objectives for sound classification since it gives sufficient prediction of whether a sound belongs to the *cry* or non-*cry* class based on the values of balanced accuracy and f1-score which reach 88% and 83% respectively.

NETWORK	BALANCED			
	ACCURACY	PRECISION	RECALL	F ₁ -SCORE
W020_RESNET18_LR1E-02	0.87 ± 0.04	0.84 ± 0.07	0.81 ± 0.08	0.82 ± 0.05
W020_RESNET34_LR1E-03	0.87 ± 0.04	0.85 ± 0.07	0.80 ± 0.08	0.82 ± 0.05
W025_RESNET18_LR1E-04	0.86 ± 0.05	0.84 ± 0.08	0.78 ± 0.09	0.80 ± 0.07
W025_RESNET34_LR1E-03	0.88 ± 0.05	0.86 ± 0.07	0.81 ± 0.11	0.83 ± 0.06

(a) Results in numerical values.



(b) Results in barplots.

Figure 4.14: Performance of the 5-folds cross-validation for the four best candidate combinations based on sound predictions.

4.5.4 Deployment of the final combination

Finally, the selected model is trained on the 30 babies and evaluated on the test set composed of three new babies. To assess the good generalization of the model, the test is performed twice. Firstly by excluding cry+ (Table 4.8a) and then by including cries mixed with other sounds (Table 4.8b). The detailed confusion matrix with all labels and the resulting binary confusion matrices excluding the cry+ labeled sounds and including them are reported in Section 4.5.4.

Thanks to a training achieved on the full training set (i.e., 30 babies), the results are better than those of the cross-validation, with 92% precision and 88% recall when cry+ is not taken into account. These results demonstrate the good ability of the model to generalize when deployed on new data. As for the deployment with the cries mixed with other sounds, one can see that the precision increased reaching a score of 95% whereas the sensitivity decreased a little (86%). These results indicate that the model is relatively good at detecting crying in sounds containing multiple sound sources.

BALANCED ACCURACY	PRECISION	RECALL	F ₁ -SCORE
0.9287	0.9242	0.8784	0.9007

(a) Excluding the sound labeled *cry+* in the *cry* class.

BALANCED ACCURACY	PRECISION	RECALL	F ₁ -SCORE
0.9194	0.9466	0.8598	0.9011

(b) Including the sound labeled *cry+* in the *cry* class.

Table 4.8: Performance of the final model on the test set including three babies.

		PREDICTIONS		TOTAL
		<i>cry</i>	<i>non-cry</i>	
	REFERENCES			
	<i>cry</i>	549	76	625
	<i>cry+</i>	248	54	302
	baby others	38	97	135
	voices	0	26	26
	alarms	1	79	80
	others	2	1824	1826
	mixtures	4	69	73

(a) Detailed confusion matrix with the seven labels.

		PREDICTIONS		TOTAL
		<i>cry</i>	<i>non-cry</i>	
	REF.			
	<i>cry</i>	549	76	625
	<i>non-cry</i>	45	2095	2140
	TOTAL	594	2171	2765

(b) Confusion matrix without *cry+*.

		PREDICTIONS		TOTAL
		<i>cry</i>	<i>non-cry</i>	
	REF.			
	<i>cry</i>	797	130	927
	<i>non-cry</i>	45	2095	2140
	TOTAL	842	2225	3067

(c) Confusion matrix with *cry+*.

Table 4.9: Confusion matrix results for the test-set.

4.6 Conclusion

This chapter was the natural continuation of the previous one focused on segmentation in which we observed that several sound segments were extracted including cries but also other sounds. Thus, we proposed in this chapter a classification approach based on deep learning.

To fulfill the objective, we presented the SoundAnnoT software that we created to annotate sound segments derived from the segmentation step. Thanks to this program, a database was designed to gather a total of 21 340 sounds annotated according to seven labels. This database was very useful for the training of the CNN which requires a lot of data to achieve good performance.

This database constitutes the first and quite important result of this chapter as it contains a large variety of sound events recorded in the NICU in four different hospitals. It offers, to our knowledge, probably the first large annotated database of sounds and cries acquired in a real environment.

Moreover, thanks to this database, we assessed the designed cry classifier with different parameter combinations. Based on the ResNet34 architecture, known to have good performances in image classification, the final model was trained on spectrograms divided into 0.25 s frame duration and with an initial learning rate of 10^{-4} . This model gave good validation results. In addition, its robustness was evaluated through a deployment on a test set composed of three new babies. Once again, the results showed good performance either when excluding or including cries mixed with other sounds by reaching up to 94.6% of precision and 85.9% recall, with a balanced accuracy of 91.9%.

Therefore, our model accuracy achieved equivalent performance to those reported in the literature (91.1% accuracy in [23], 82.8% F1-score in [24], and 86.6% accuracy-precision score in [22]) and it is worthwhile to mention that we overcome the given limitations regarding the representativeness of the training and evaluation datasets thanks to our annotated database. Then, although the deployments are different, we can notice that we also obtain a higher accuracy (i.e., 95%) than the previous study conducted by our team which achieved a score of 92.2% [27]. In addition, by selecting spectrograms with a full frequency band as input to the CNN, we overcame the problem encountered in the classification of some sounds whose spectral energy was not in the frequency band considered.

Furthermore, even if the two-step parameter optimization strategy is not very common, it has proven to be relevant since it avoided the long computation times associated with classical strategies (e.g., grid search) while obtaining a final model with very good performance. Of course, it is also thanks to the transfer learning approach that this strategy could be applied. Indeed, as the ResNet models were already finely optimized, a less greedy optimization could be undertaken. Having identified that this architecture gives good results, we can consider going further in the optimization of hyper-parameters to obtain an even more efficient model. For now, ours meets the

goal to design a robust classifier for deployment in clinics with a good precision score. Indeed, by ensuring that the sounds predicted as cries are actual cries, we ensure the reliability of the further cry analyses.

APPENDIX A - BEST CANDIDATE COMBINATIONS SELECTION

CLASS	SOUNDS		FRAMES 0.20s		FRAMES 0.25s	
	QTY	PERC.	QTY	PERC.	QTY	PERC.
TRAIN SET (N=29)						
cry	4 807	28.56%	52 681	32.77%	37 892	32.69%
non-cry	12 027	71.44%	108 095	67.23%	78 012	67.31%
TOTAL	16 834	100.00%	160 776	100.00%	115 904	100.00%
VALIDATION SET (N=1)						
cry	44	21.15%	631	26.64%	454	26.69%
non-cry	164	78.85%	1 738	73.36%	1 247	73.31%
TOTAL	208	100.00%	2 369	100.00%	1 701	100.00%
TOTAL						
	17 042	100.00%	163 145	100.00%	117 605	100.00%

Table 4.10: Detailed database used during the best candidate combinations selection using a simple-validation with a train set including 29 babies and a validation set of 1 baby.

		BALANCED				
		LR	ACCURACY	SPECIFICITY	RECALL	F ₁ -SCORE
0.20s	Resnet18	10 ⁻²	0.8213	0.8385	0.6910	0.7576
		10 ⁻³	0.8310	0.8558	0.7052	0.7732
		10 ⁻⁴	0.8649	0.8390	0.7845	0.8108
		10 ⁻⁵	0.8375	0.8460	0.7227	0.7795
	Resnet34	10 ⁻²	0.8103	0.8221	0.6735	0.7404
		10 ⁻³	0.8094	0.8510	0.6609	0.7440
		10 ⁻⁴	0.8655	0.8612	0.7765	0.8167
		10 ⁻⁵	0.8460	0.8363	0.7448	0.7879
0.25s	Resnet18	10 ⁻²	0.8255	0.8431	0.6982	0.7639
		10 ⁻³	0.8578	0.8564	0.7621	0.8065
		10 ⁻⁴	0.8510	0.8645	0.7445	0.8000
		10 ⁻⁵	0.8485	0.8454	0.7467	0.7930
	Resnet34	10 ⁻²	0.8326	0.8541	0.7093	0.7750
		10 ⁻³	0.8879	0.8472	0.8304	0.8387
		10 ⁻⁴	0.8628	0.8722	0.7665	0.8159
		10 ⁻⁵	0.8648	0.8396	0.7841	0.8109

Table 4.11: Performance of the classifier for the frame database during the best candidate combinations selection using a simple-validation on one baby.

APPENDIX B - FINAL COMBINATION SELECTION

		GA		PMA	SET	CRY	NON-CRY	TOTAL
SET 1	<i>mean</i>	32+6	<i>mean</i>	36+6	<i>sounds</i>	885	2 142	3 027
	<i>s.d</i>	5+1	<i>s.d</i>	3+4	<i>frames (0.20s)</i>	11 774	12 428	24 202
	<i>range</i>	27+0-40+3	<i>range</i>	29+6-41+3	<i>frames (0.25s)</i>	8 514	8 512	17 026
SET 2	<i>mean</i>	32+5	<i>mean</i>	32+6	<i>sounds</i>	830	1 978	2 808
	<i>s.d</i>	4+3	<i>s.d</i>	4+2	<i>frames (0.20s)</i>	7768	18 208	25 976
	<i>range</i>	27+5-39+1	<i>range</i>	28+1-39+2	<i>frames (0.25s)</i>	5 525	13 236	18 761
SET 3	<i>mean</i>	35+3	<i>mean</i>	35+7	<i>sounds</i>	1 094	2 917	4 011
	<i>s.d</i>	4+4	<i>s.d</i>	4+3	<i>frames (0.20s)</i>	15 940	38 795	54 735
	<i>range</i>	27+5-41+4	<i>range</i>	28+2-41+5	<i>frames (0.25s)</i>	11 633	28 363	39 996
SET 4	<i>mean</i>	30+5	<i>mean</i>	31+3	<i>sounds</i>	993	2 933	3 926
	<i>s.d</i>	4+4	<i>s.d</i>	4+1	<i>frames (0.20s)</i>	8 701	15 233	23 934
	<i>range</i>	25+6-39+4	<i>range</i>	27+5-39+6	<i>frames (0.25s)</i>	6 179	10 774	16 953
SET 5	<i>mean</i>	33+5	<i>mean</i>	34+3	<i>sounds</i>	1 049	2 221	3 270
	<i>s.d</i>	4+4	<i>s.d</i>	4+5	<i>frames (0.20s)</i>	9 129	25 169	34 298
	<i>range</i>	29+0-40+3	<i>range</i>	29+0-40+5	<i>frames (0.25s)</i>	6 495	18 374	24 869

Table 4.12: Detailed database used during the final combination selection using a cross-validation with 5 folds including six babies each.

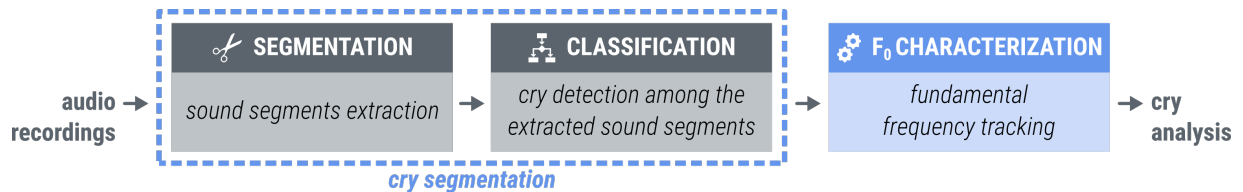
BIBLIOGRAPHY

- [1] 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*, 358–365. Springer (2004).
- [2] HARIHARAN M., YAACOB S., AND AWANG S.A. Pathological infant cry analysis using wavelet packet transform and probabilistic neural network. *Expert Systems with Applications*, vol. 38, 15 377–15 382 (2011).
- [3] 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*, 342–349. Springer (2003).
- [4] 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*, vol. 17, 38–46 (2015).
- [5] ONU C.C., LEBENSOLD J., HAMILTON W.L., AND PRECUP D. Neural transfer learning for cry-based diagnosis of perinatal asphyxia. *arXiv preprint arXiv:1906.10199* (2019).
- [6] 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*, 408–415. Springer (2004).
- [7] REYES-GALAVIZ O.F. AND REYES-GARCÍA C.A. Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system. In *Mexican International Conference on Artificial Intelligence*, 949–958. Springer (2005).
- [8] ZABIDI A., MANSOR W., KHUAN L.Y., SAHAK R., AND RAHMAN F. Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In *2009 5th International Colloquium on Signal Processing & Its Applications*, 204–208. IEEE (2009).
- [9] ZABIDI A., KHUAN L.Y., MANSOR W., YASSIN I.M., AND SAHAK R. Detection of infant hypothyroidism with mel frequency cepstrum analysis and multi-layer perceptron classification. In *2010 6th International Colloquium on Signal Processing its Applications*, 1–5 (2010).
- [10] 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*, vol. 46, 965–975 (2008).
- [11] BHAGATPATIL M.V. AND SARDAR V. An automatic infant's cry detection using linear frequency cepstrum coefficients (lfcc). *International Journal of Scientific & Engineering Research*, vol. 5, 1379–1383 (2014).
- [12] CHANG C.Y. AND LI J.J. Application of deep learning for recognizing infant cries. In *Consumer Electronics-Taiwan (ICCE-TW), 2016 IEEE International Conference on*, 1–2. IEEE (2016).
- [13] FRANTI E., ISPAS I., AND DASCALU M. Testing the universal baby language hypothesis - automatic infant speech recognition with CNNs. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, 1–4 (2018).
- [14] MAGHFIRA T.N., BASARUDDIN T., AND KRISNADHI A. Infant cry classification using CNN - RNN. *Journal of Physics: Conference Series*, vol. 1528, page 012019 (2020).

- [15] LE L., KABIR A.N.M., JI C., BASODI S., AND PAN Y. Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, 106–110 (2019).
- [16] 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*, 1–5. IEEE (2016).
- [17] TORRES R., BATTAGLINO D., AND LEPAULOUX L. Baby cry sound detection: A comparison of hand crafted features and deep learning approach. In G. Boracchi, L. Iliadis, C. Jayne, and A. Likas, editors, *Engineering Applications of Neural Networks*, 168–179. Springer International Publishing (2017).
- [18] XIE J., LONG X., OTTE R., AND SHAN C. Convolutional neural networks for audio-based continuous infant cry monitoring at home. *IEEE Sensors Journal*, vol. 21, 27 710–27 717 (2021).
- [19] COHEN R., RUINSKIY D., ZICKFELD J., IJZERMAN H., LAVNER YIZHAR" E.W., AND CHEN S.M. *Baby Cry Detection: Deep Learning and Classical Approaches*, In *Development and Analysis of Deep Learning Architectures*, 171–196. Springer International Publishing, Cham (2020).
- [20] CHANG C.Y. AND TSAI L.Y. A CNN-based method for infant cry detection and recognition. In L. Barolli, M. Takizawa, F. Xhafa, and T. Enokido, editors, *Web, Artificial Intelligence and Network Applications*, 786–792. Springer International Publishing (2019).
- [21] 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*, vol. 39, 390–395 (2018).
- [22] FERRETTI D., SEVERINI M., PRINCIPI E., CENCI A., AND SQUARTINI S. Infant cry detection in adverse acoustic environments by using deep neural networks. In *26th European Signal Processing Conference, EUSIPCO 2018*. European Signal Processing Conference, EUSIPCO (2018).
- [23] ABOU-ABBAS L., TADJ C., GARGOUR C., AND MONTAZERI L. Expiratory and inspiratory cries detection using different signals' decomposition techniques. *Journal of Voice*, vol. 31, 259–e13 (2017).
- [24] 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*, vol. 2018, 1–14 (2018).
- [25] SEVERINI M., FERRETTI D., PRINCIPI E., AND SQUARTINI S. Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation. *IEEE Access*, vol. 7, 51 982–51 993 (2019).
- [26] SEVERINI M., PRINCIPI E., CORNELL S., GABRIELLI L., AND SQUARTINI S. Who cried when: Infant cry diarization with dilated fully-convolutional neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (2020).
- [27] CABON S., MET-MONTOT B., PORÉE F., ROSEC O., SIMON A., AND CARRAULT G. Extraction of premature newborns' spontaneous cries in the real context of neonatal intensive care units. *Sensors*, vol. 22, page 1823 (2022).
- [28] CABON S., PORÉE F., SIMON A., ROSEC O., PLADYS P., AND CARRAULT G. Video and audio processing in paediatrics: A review. *Physiological Measurement*, vol. 40 (2019).
- [29] SPRENGEL E JAGGI M K.Y.H.T. Audio based bird species identification using deep learning techniques. In *Working notes of CLEF 2016* (2016).
- [30] LASSECK M. Audio-based bird species identification with deep convolutional neural networks. *CLEF (working notes)*, vol. 2125 (2018).
- [31] BADSHAH A.M., AHMAD J., RAHIM N., AND BAIK S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. *2017 International Conference on Platform Technology and Service (PlatCon)*, 1–5 (2017).

- [32] BADSHAH A., RAHIM N., ULLAH N., AHMAD J., MUHAMMAD K., LEE M., KWON S., AND BAIK S. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, vol. 78 (2019).
- [33] SATT A., ROZENBERG S., AND HOORY R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In *Proc. Interspeech 2017*, 1089–1093 (2017).
- [34] LI Y., ZHAO T., AND KAWAHARA T. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech*, 2803–2807 (2019).
- [35] 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*, vol. 6, 1–10 (2018).
- [36] arXiv, HE K., ZHANG X., REN S., AND SUN J. Deep residual learning for image recognition (2015).
- [37] RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATY A., KHOSLA A., BERNSTEIN M., BERG A.C., AND FEI-FEI L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, vol. 115, 211–252 (2015).
- [38] PALANISAMY K., SINGHANIA D., AND YAO A. Rethinking CNN models for audio classification. *CoRR*.

Fundamental frequency characterization



5.1 Introduction

In **Chapter 1**, we saw that crying is produced by a complex biological phenomenon that is a combination of neural and physiological mechanisms. Thus, the vocal cords variation or the fundamental frequency analysis is particularly interesting since it is intimately related to the infant's neurological development. Hence, it is of great interest to develop methods to automatically track this acoustic parameter to perform infant cry analyses.

This chapter is therefore part of the continuity of our processing chain whose objective is the automatic analysis of infant crying. Indeed, we present a new fundamental frequency characterization method to track the main spectral component of infant cries which were automatically extracted thanks to the methods proposed previously.

After a review of the state of the art, we introduce the automated method based on contour detection in spectrograms for the tracking of the fundamental frequency. Furthermore, a validation of the method is carried out by comparing our results with those obtained with the BioVoice software whose performances were validated on synthetic basic melodic shapes of the newborn cry [1].

5.2 State of the art

A cry signal is known to be a periodic signal, which is a signal that repeats itself at a specific time interval called the period. The fundamental frequency is defined as the inverse of this period, while the harmonic frequencies are defined as integer multiples of F_0 .

When working with continuous, stable, periodic signals, acoustic analysis can be quite simple. However, real-world signals, such as speech or infant cries, are not perfectly periodic, which

makes their analysis more complex. Indeed, due to the intrinsic periodic variations in time, it is not relevant to characterize these signals with a single frequency parameter. However, according to the quasi-periodicity assumption, it is possible to assume that these signals are periodic in very small time frames. Therefore the most common cry characterization consists to track the frequency components along the cry unit by computing features for each of these frames.

5.2.1 Methods

There are several techniques to solve the problem of fundamental frequency estimation, such as temporal, spectral, wavelet, and image domain approaches. Below is an overview of the most common methods used in the field of infant crying analysis.

Time domain

Regarding the time domain, the auto-correlation function is the most largely used method to estimate the fundamental frequency in cries. It is a measure of self-similarity of a signal in the time domain when compared to a delayed version of itself. Designed to determine the periodicity of the signal, it was firstly used in [2] and was then implemented in 2002 in the famous PRAAT software [3] widely adopted nowadays for the F_0 characterization of crying [1, 4–8]. However, described as hard to manipulate [1], it can require manual correction of the F_0 estimation tracking errors [6]. For their part, Naithani et al. proposed to use the YIN algorithm which is known to be more robust on quasi periodic signals [9].

Spectral domain

In early studies, frequency feature characterization used to be based on spectrographic analysis through visual inspection of sound spectrograms [10–15]. Then, through the emergence of computer audio signal processing methods, it has become possible to use automatized estimation methods. The analysis of a signal in terms of frequency is done thanks to the conversion of the signal from the time to the frequency domain by using Fourier Transform.

Most of the time, energy features are directly computed from spectrum and peak-picking procedures were implemented to extract F_0 or resonance frequencies [16]. Although simple to implement, these methods are not suitable for complex cases such as when there is background noise in the recording or when the harmonics of the cry have a much higher intensity than the fundamental frequency.

In speech analysis, Long Time Average Spectrum (LTAS) is used for the identification of pathological speech. The calculation of average spectra allows LTAS to eliminate short-term variations present in the human voice due to the filtering properties of the vocal tract. In particular, it has been proven that this method provides a good representation of the acoustic signal with minimal

influence of the vocal tract in order to better distinguish between different types of vocal behavior in infants, as well as between healthy and unhealthy infants [17–19].

For their part, Varallyay et al. used the smoothed spectrum method which is a very accurate algorithm for detecting the most probable value of the fundamental frequency. It is based on the spectral analysis and is usually combined with noise filtering and statistical processing [20, 21].

Wavelet domain

The wavelet transform is another way to transform the audio signal from the time domain into a time-frequency (more precisely time-scale) representation. It calculates the inner product of the signal with a wavelet family. There are two types of wavelets: the continuous and the discrete wavelet transform. Both have the ability to extract information from non-stationary signals such as audio. In addition, thanks to their variable time-frequency resolutions these transforms can overcome the shortcomings of STFT which has a uniform resolution.

Although this approach has proven to be useful in the estimation of the fundamental frequency in adult voices [22, 23], it is still very little used in infant cry analysis. Only the Italian team of Manfredi et al. proposes continuous wavelet transform approaches, known for their robustness to noise [24, 25].

Image domain

Eventually, some methods are based directly on the extraction of parameters in the spectrogram image. This is for example the case of the Scale Invariant Feature Transform (SIFT) which is initially a feature extraction algorithm in computer vision used to detect local information in images. Manfredi et al. also proposed a tuned method of the algorithm [26] which gave better F_0 estimation than the original method [27] also used in [28].

5.2.2 Softwares

Nowadays, the most popular software in acoustic analysis is PRAAT [3]. Initially designed for adult voice by Boesrma in 2002, it was then used in [1, 4–8] for fundamental frequency estimation, in [29–32] for frequency features (such as MFCCs) and in [28] for noise filtering and segmentation of the recordings into useful and non-useful categories. Next, the openSMILE tool also allows the extraction of acoustic parameters [33–35]. Both of these programs perform the automatic calculation of a wide variety of features (e.g., F_0 , formants, MFCC, LPCC, jitter, shimmer) but must be initialized manually to give a meaningful analysis, particularly when analyzing infant cries.

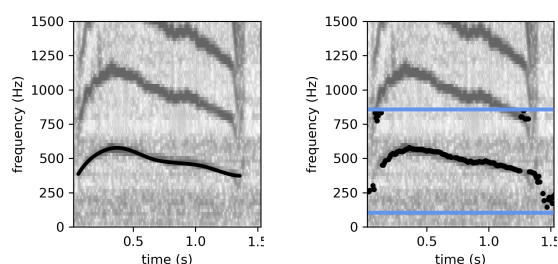
For their part, Manfredi et al. developed BioVoice [1, 26, 36] and WInCA [37], two programs developed for infant cry analysis, where different estimation methods of F_0 (respectively, SIFT

and wavelet) and resonance frequencies (respectively, peak picking in the power spectral density and wavelet) were implemented.

5.2.3 Limitations on the fundamental frequency estimations

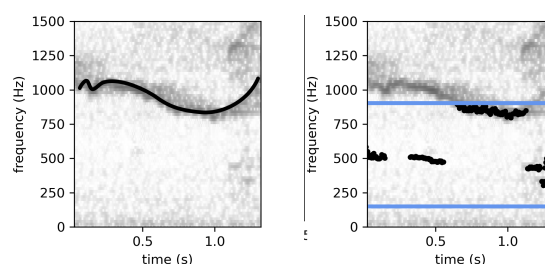
Cries are quasi-periodic signals with high-energy harmonic components. Therefore it can be difficult to estimate the fundamental frequency with the reported methods which are based on signal energy computation. Indeed, methods such as auto-correlation sometimes detect peaks that may correspond to harmonic components. To prevent such jumps in the tracking, authors usually limit the F_0 estimation within a fixed frequency band which is mostly set between 150 and 1000 Hz [5, 25, 26] corresponding to phonation cries [38]. As a consequence, this frequency band in which the tracking is performed is of utmost importance and has a great influence on the results [1]. The only method proposed to our knowledge without frequency limit is the YIN algorithm which has been developed for speech or musical sounds. It was used to extract F_0 features for the purpose of automatic segmentation of infant cry signals [9].

To illustrate this issue, we give two specific cry examples for which the fundamental frequency tracking was performed using the BioVoice software. This program is a user-friendly software tool for the acoustical analysis of the human voice and is described later in this manuscript (see [Section 5.4.1](#)). In both cases, the F_0 was computed within the fixed frequency band ranging from 150 to 900 Hz [1]. In the first sample, one can see that around the edges the high energy harmonic components affect the tracking with a shift of the estimation up to these components ([Figure 5.1](#)). While, in the second sample, the F_0 is impossible to track since the fundamental frequency of the hyperphonation cry is outside the analyzed frequency band ([Figure 5.2](#)).



(a) Expected estimation. (b) Resulting estimation.

Figure 5.1: Sample 1 - phonation cry with high energy harmonic components.



(a) Expected estimation. (b) Resulting estimation.

Figure 5.2: Sample 2 - hyperphonation cry with F_0 outside the analysed band.

In addition, although some fundamental frequency tracking methods have been adapted to the infant cry analysis, none has been performed for the purpose of continuous monitoring in a clinical context.

5.3 Proposed method

Therefore, in this section, we propose a new method to extract the fundamental frequency component of the cries automatically extracted with the automated methods presented in the previous chapters. Furthermore, we focused on the development of an initial step for the automatic selection of a relevant frequency band in which to perform the F_0 tracking. The fundamental frequency estimation is achieved through a contour detection in the spectrogram. A similar procedure has already been carried out on underwater audio recordings with a two-steps algorithm including a peak-picking and a contour detection steps [39]. The whole framework of the proposed method is illustrated in **Figure 5.3** and described hereafter.

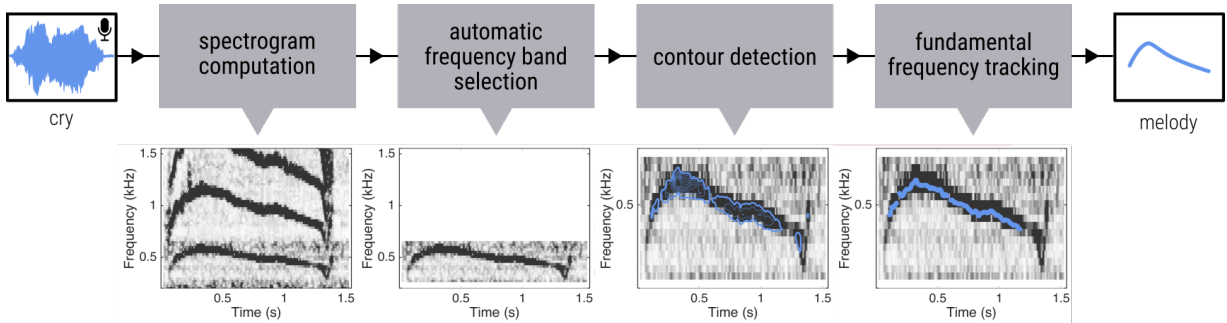


Figure 5.3: Fundamental frequency tracking flowchart.

5.3.1 Spectrogram computation

First of all, the cry unit is filtered by a Kaiser Window finite impulse response filter with cut-off frequencies between 250 and 1500 Hz. These bounds correspond to a wide variability range in which to expect an infant cry fundamental frequency considering hyperphonation cries (F_0 higher than 1000 Hz). Then, the spectrogram calculation is the same as the one done in **Chapter 4** (i.e., STFT of successive 1000 samples Hamming-windowed frames with a 90% overlap). Since the audio data are recorded at a 24 kHz sampling rate, the given configuration provides a spectrogram with a frequency resolution of 23.4 Hz ($n = 53$ frequency bins per frame) and a time resolution of 4.2 ms (**Figure 5.4**).

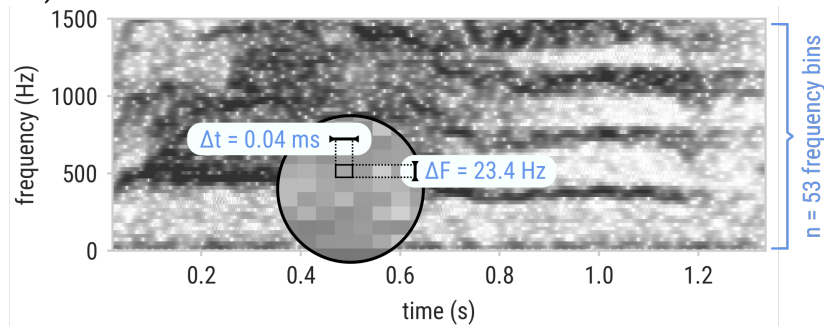


Figure 5.4: Spectrogram computation with frame resolution for signal sample at 24kHz.

5.3.2 Automatic frequency band selection

Once the spectrogram is calculated, the frequency bounds are automatically found by detecting the fundamental frequency location and surroundings in the maximum value distribution. This strategy is based on the four steps ([Figure 5.5](#)) described below.

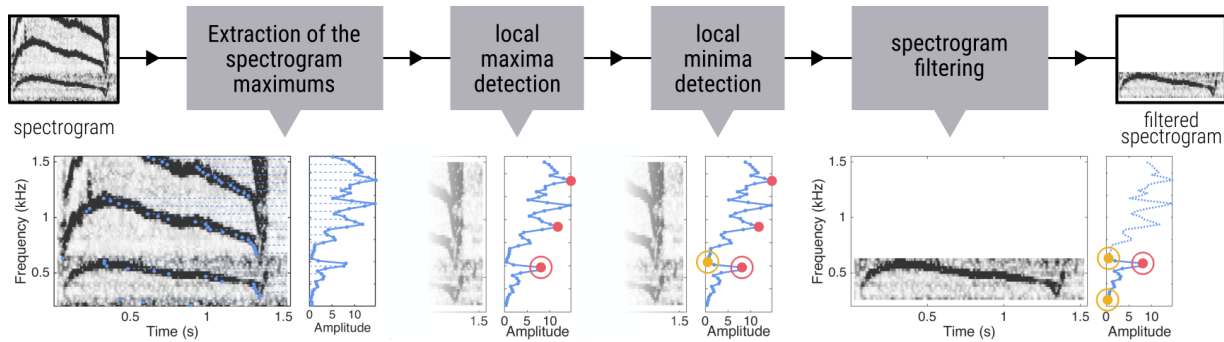


Figure 5.5: Automatic frequency band selection framework.

- Extraction of the spectrogram maximums** For each frequency row of the spectrogram, the point with the maximum amplitude is extracted. Result is called the maximum curve (c_{\max}) which is, therefore, of the same size $n=53$ than the frequency vector.
 - Local maxima detection** Local maxima are sought in the c_{\max} curve. In order to avoid detection of peaks that are not main frequency components, only peaks whose amplitude is greater than 5% of the maximum amplitude are detected. Furthermore only peaks separated by more than 300 Hz are retained since infant cries are harmonic signals. The fundamental frequency is detected at the lowest peak based on the frequency scale.
 - Local minima detection** Local minima are also sought in the c_{\max} curve. In order to avoid detection of irrelevant minima, a threshold is chosen as 20% of the amplitude of the selected maximum peak. Local minima surrounding this peak are chosen to be the new limits for the chosen frequency band.
- When the first maximum has been detected close to the initial bound (250 or 1500 Hz), one surrounding minimum might be missing. In this case, the missing bound is set to the initial corresponding one (such as illustrated in [Figure 5.5](#) where the lower bound is set to 250 Hz).
- Spectrogram filtering** Spectrogram values with corresponding frequency above or under the new limits are removed (i.e., set to NaN).

5.3.3 Contour detection

Contours are detected using a low-level contour matrix computation where isolines are calculated over cross-sections of the spectrogram magnitude with respect to the time-frequency plane [Figure 5.6a](#). Main spectral components contours are obtained by selecting the low-height isolines. Moreover, as the melody tends to be continuous over the cry unit width, contours of duration less than 0.05 s are disregarded. At last, contours included inside another contour are neglected and when two or more contours are temporally overlapping, only the wider one is kept [Figure 5.6b](#).

5.3.4 Fundamental frequency tracking

Fundamental frequency tracking is performed by computing the average of the contours with a time step of 4 ms. Finally, a vector (\mathbf{v}_0) of the size of the cry unit is created and filled with :

$$\mathbf{v}_0 = \begin{cases} \text{averaged contours} & \text{as they appear over time,} \\ \text{NaN} & \text{otherwise.} \end{cases} \quad (5.1)$$

An example of the resulting fundamental frequency vector is provided in [Figure 5.6c](#).

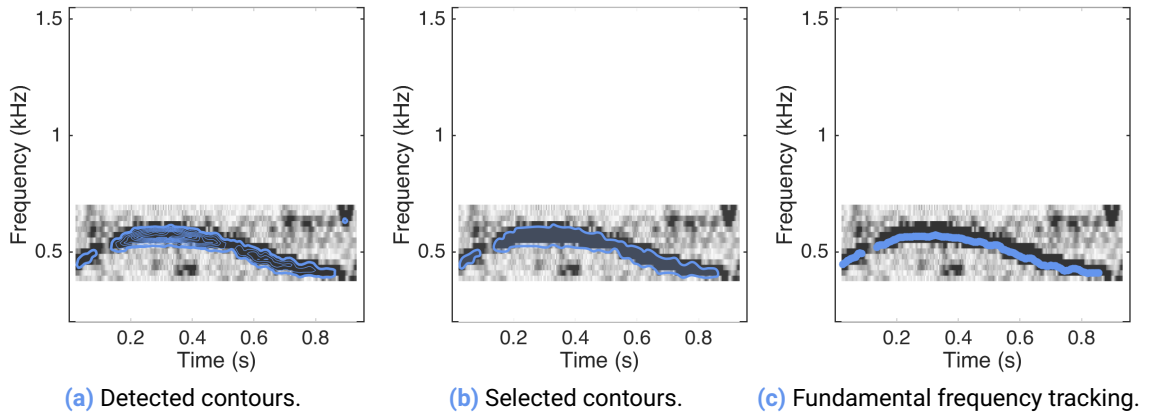


Figure 5.6: Illustration of the contour detection step with the detected contours (left) and the selected ones (middle). The small right contour is discarded since it is shorter than 0.05 s otherwise only the wider contours are kept (i.e., those including other contours). Then, fundamental frequency tracking (right) is computed within the selected contours. Thus, in this example, the resulting vector \mathbf{v}_0 is filled with two averaged contours.

5.4 Evaluation strategy

To validate the proposed method, we chose to compare our results with those obtained with a method from the literature. While PRAAT is today the most common software used in speech analysis, we computed the frequency estimation with BioVoice. Indeed, this software has proven

to be relevant for infant cries analysis since their performances were validated on synthetic basic melodic shapes of the newborn cry [1].

First we present the BioVoice software tool for acoustic analysis. Then, we propose to perform a qualitative comparison based on visual annotations as well as a statistical comparison of the usual parameters evaluated in such a study, i.e., min, max, mean and median values of the F_0 estimation.

5.4.1 BioVoice software

BioVoice is a user-friendly software tool for the acoustic analysis of various vocal emissions, from newborns to adults and singers [1]. It is designed in MATLAB and distributed free of charge on GitHub. First, a sound file must be loaded. It can either be:

- a long recording in which case BioVoice applies a segmentation and then characterizes all the extracted sounds;
- sound segments which are directly characterized.

In both cases, an initialization step is necessary to inform the software of the type of signals to process (i.e., newborn, child, adult).

Software execution

In case of a long recording, the software automatically process a segmentation step to detect the voiced and unvoiced segments in the input signal [40].

Then pressing "Start" button launch the estimation of more than 20 acoustic temporal and frequency parameters based on advanced and robust analysis techniques. We can mention the following ones: detecting the number, length, and percentage of voiced and unvoiced segments and calculating the fundamental frequency, formant frequencies (F_1 - F_3) [37], noise level, and jitter [41]. Specifically, for newborn cry and child voice, it computes the melodic shape of F_0 , automatically identifying up to 12 melodic shapes [37, 42, 43].

As for the F_0 tracking it is performed by means of a two-step algorithm [26]. After applying the SIFT to time windows of short and fixed length, the fundamental frequency is then adaptively estimated on signal frames of variable length through the average magnitude difference function within the range provided by the SIFT. Therefore, the resulting estimation vectors all have different time steps varying, according to our observation on the processed cries, from 4 to 500 ms.

The computed acoustic parameters are saved in 13 Excel files for each sound segment. The one containing the F_0 estimation is saved in a file named `{audioname}_F0.xls`.

Cry characterization for comparison

To perform the method comparison, cries are directly loaded in the platform using with parameters adapted to infant crying analysis *Age - Range*: Newborn/Infant and *Voice Emission*: Cry (see [Figure 5.7](#)). Then cries are characterized by the 20 acoustic parameters.

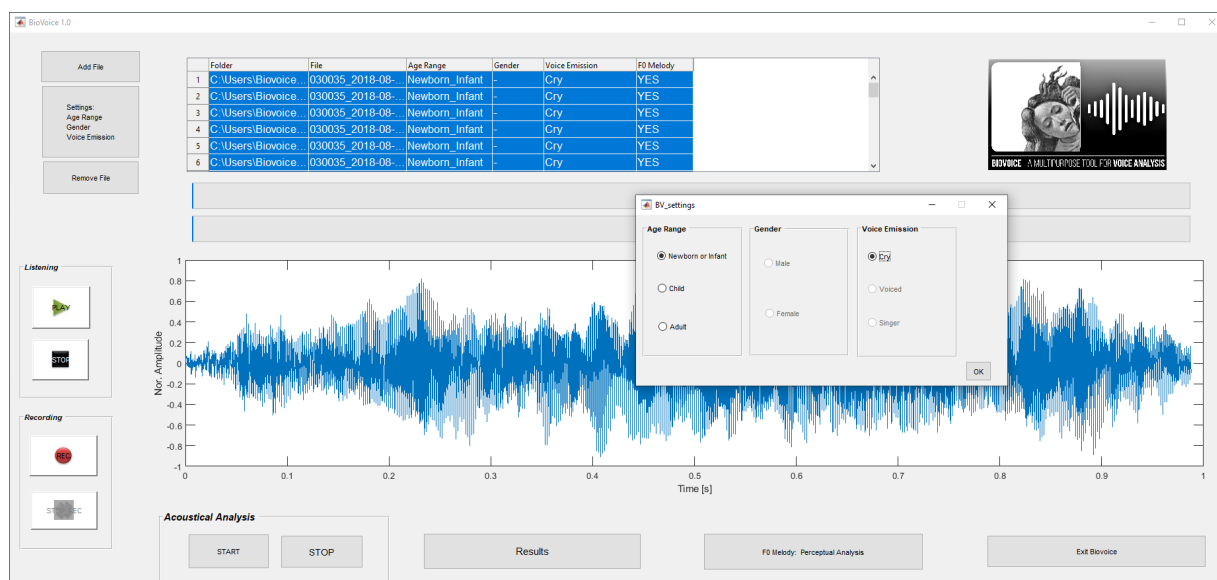


Figure 5.7: BioVoice setting interface - Newborn cry analysis selection.

Limitations

Despite the real interest of the BioVoice system, this platform is not adapted for real-time processing in the NICU environment since it requires manual interactions. In addition, the calculation of the numerous acoustic parameters proposed can be very long.

5.4.2 Qualitative comparison with BioVoice

The proposed method is compared to the BioVoice through a visual annotation of both fundamental frequency trackings superimposed on the spectrogram.

Thanks to a graphical interface developed under Python, we annotated the F_0 estimations considering two aspects:

- first, the accuracy of the estimates. In practice, it means that through a visualization of the signals, we judge if one, both or none of the resulting estimation are correctly tracking the fundamental frequency;
- in a second step we assess whether one method is better than the other or if both estimations are equivalent.

To remain as objective as possible, the signals were anonymized in the interface. During the annotation step the signals were displayed randomly in blue or black without any distinction criteria.

The annotation procedure is illustrated in [Figure 5.8](#) with both method estimations anonymized and superimposed a spectrogram. In this example we considered that both estimations correctly tracked the fundamental frequency and that they are equivalent, i.e., there is not one better than the other.

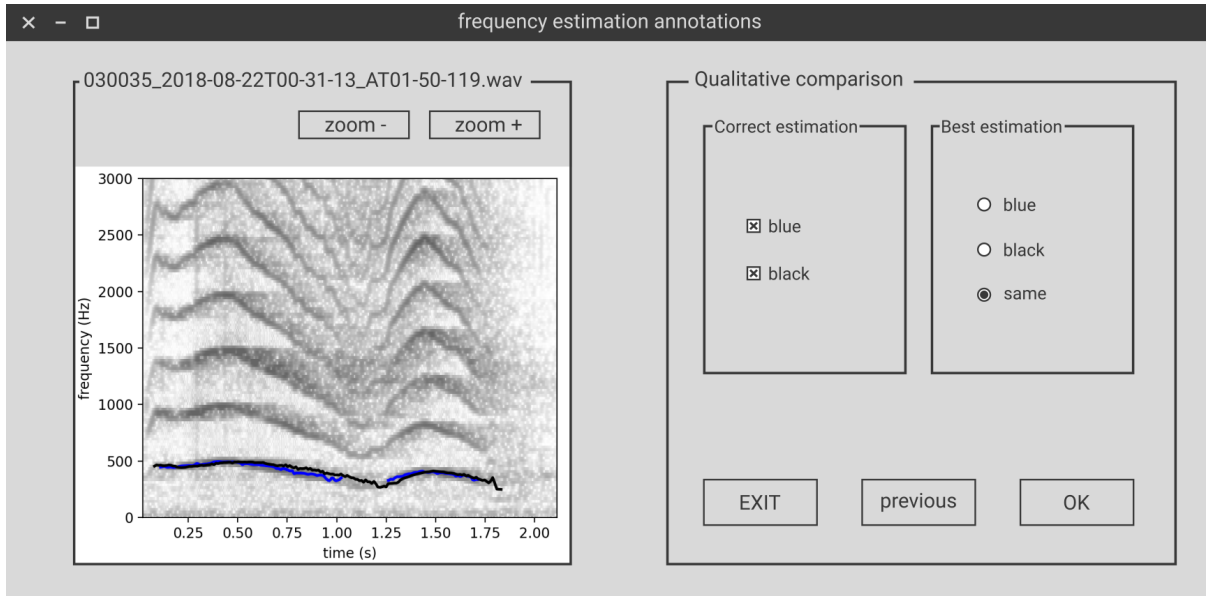


Figure 5.8: Python interface designed for the annotation of anonymized signals (randomly affected to blue or black color). In this example, both methods achieved similar good F_0 tracking.

5.4.3 Statistic parameters comparison

Usually in cry analysis, fundamental frequency is described in terms of statistical parameters. Therefore, for each cry, the F_0 mean, median, standard deviation, maximum, and minimum values are calculated for both methods. Then the Pearson correlation coefficient is computed to measure the linear correlation between the two datasets.

5.5 Results

This section presents the cry database annotated with the Python interface. Then the qualitative and statistic comparisons are performed.

5.5.1 Annotated database

The SoundAnnoT software (see [Section 4.4](#)) was used to identify cries to evaluate the fundamental frequency tracking method. A total of 806 cries recorded in a preterm female infant¹ (GA: 29 weeks + 6 days) were collected from three different recordings in open and closed bed with characteristics described in [Table 5.1](#).

	PMA (w+d)	BED TYPE	CRIS
RECORDING 1	30+0	closed	281
RECORDING 2	32+2	closed	309
RECORDING 3	38+3	open	216
TOTAL			806

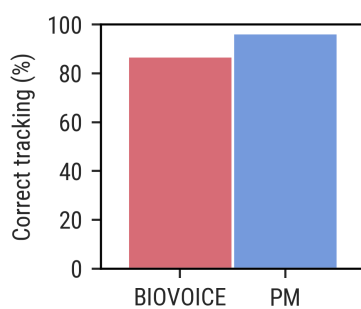
Table 5.1: Annotated cries database for the fundamental frequency estimation evaluation.

5.5.2 Qualitative comparison with BioVoice

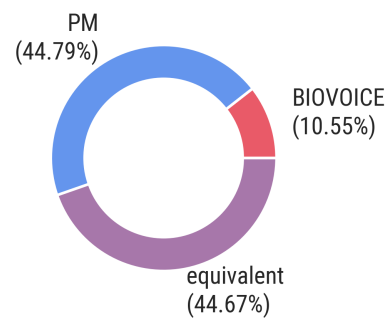
The qualitative comparison of the 806 cries, based on the visual annotation of the anonymized results, shows that both fundamental frequency tracking methods have good accuracy with 702 (87%) cries correctly estimated by BioVoice and 779 (97%) by the proposed method ([Figure 5.9a](#)).

When comparing the two methods with each other ([Figure 5.9b](#)), 360 cry estimations were considered equivalent (44.67%). While BioVoice appeared to give better results for 85 cries (10.55%), our method was reported with a better estimation in 361 cases (44.79%).

Since fundamental frequency tracking estimations vary greatly according to the acoustical cries characteristics (energy components, type, melody), the following section presents examples for the three comparison types: when i) both estimations are equivalent, ii) BioVoice is better and, iii) the proposed method is better.



(a) Accuracy of both methods.



(b) Comparison of the two methods.

Figures 5.9: Qualitative F_0 methods annotation results.

1. 030035 recorded between 9:00 PM and 7:00 AM the 2018-06-23 (1), 2018-07-09 (2), and 2018-08-21 (3).

Both estimations are equivalent

In many cases, both estimations correctly track the fundamental frequency, some examples are given in [Figures 5.10](#) for several types of cries with different melodies. Since both methods are correct, it can be hard to see both estimations as they are fitting. BioVoice is depicted in red dotted line whereas our estimation method is in blue dashed line.

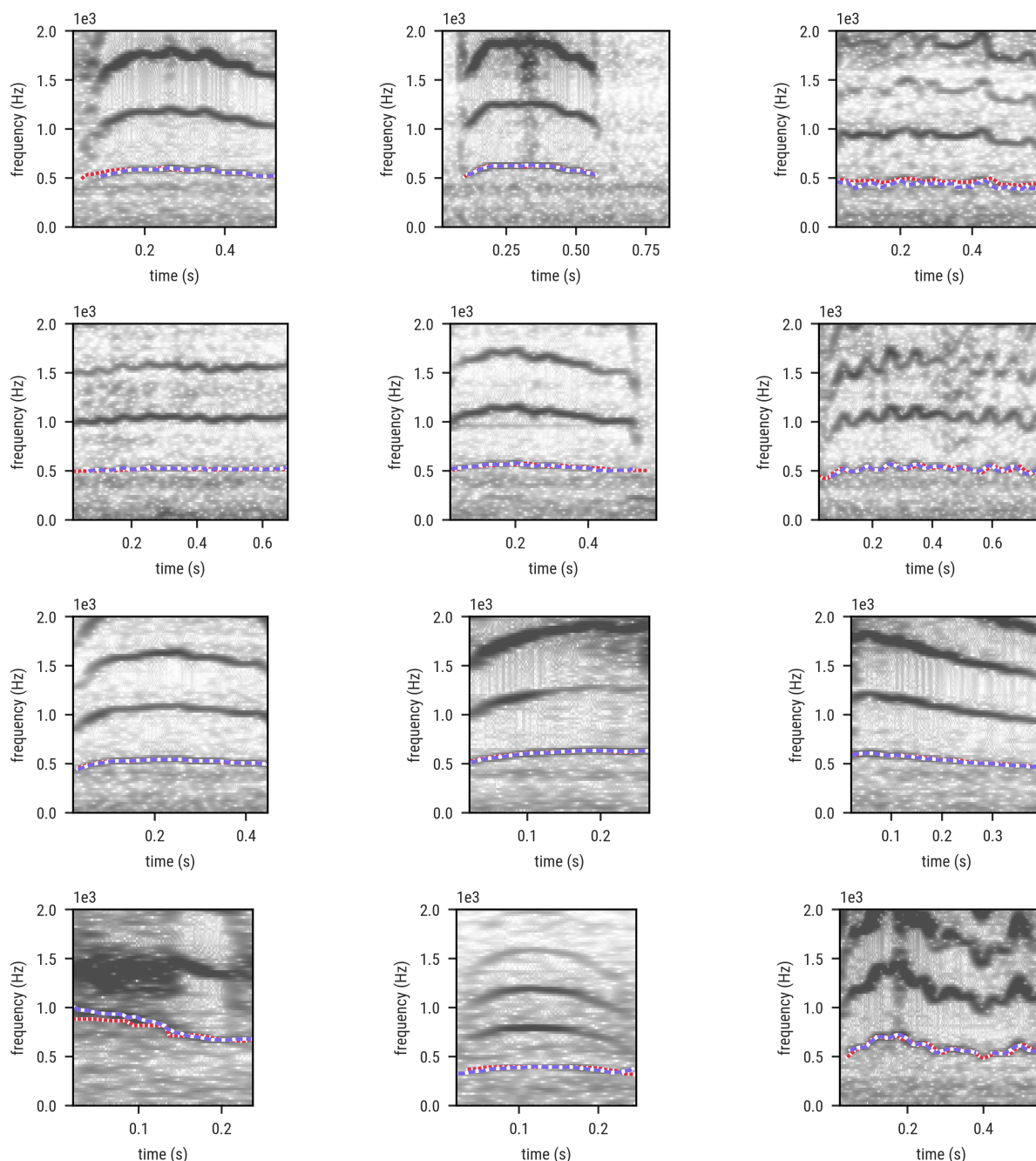
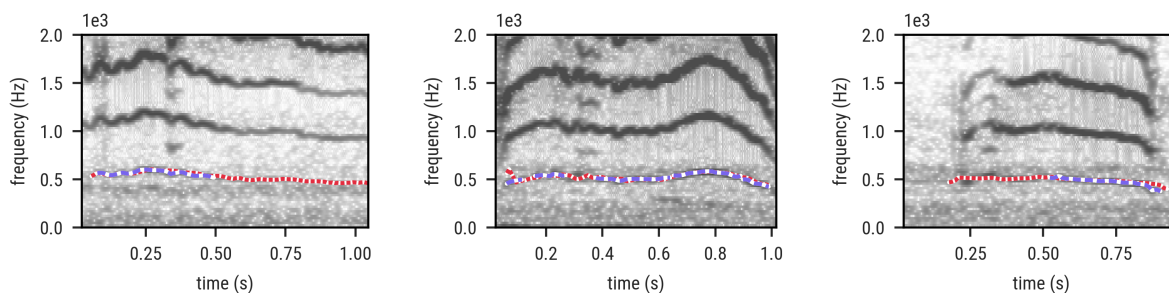


Figure 5.10: Correct estimation examples for both F_0 tracking methods.

--- proposed method BioVoice

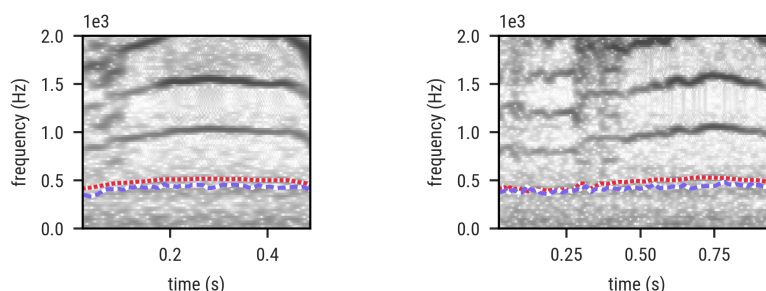
BioVoice method is better

PARTIAL ESTIMATION - The BioVoice method happens to be better in some cases where the proposed method failed to track the fundamental frequency along with the whole cry unit. The F_0 tracking can be missing at the end (left), in the middle (center) or at the start (right) of the cry such as in illustrated in [Figures 5.11](#).



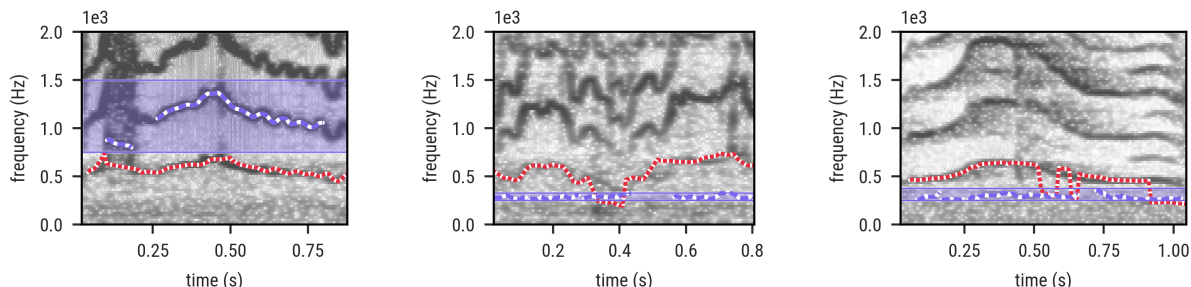
[Figures 5.11](#): Partial estimation with the proposed method.

SMALL FREQUENCY SHIFT - In some cases, the proposed method estimates the fundamental frequency with a small shift in favor of low frequencies [Figures 5.12](#).



[Figures 5.12](#): Small frequency shift with the proposed method.

BAD AUTOMATIC FREQUENCY BAND SELECTION - In limited cases, the frequency band automatic selection has failed. Therefore it is not possible to follow the fundamental frequency which is then outside the analysis band. Three examples are given in [Figures 5.13](#) with the frequency bounds and frequency ranges illustrated by horizontal lines and a colored patch respectively.

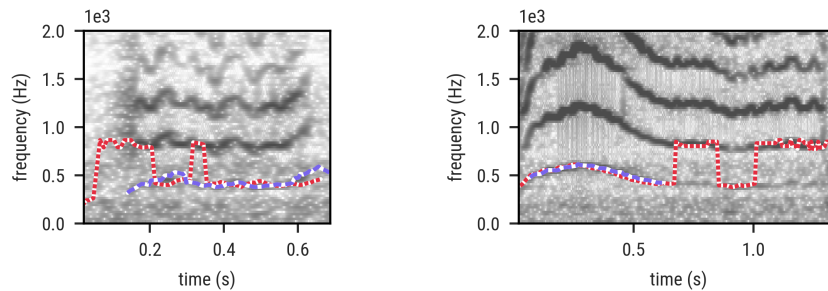


[Figures 5.13](#): Bad frequency band selections in the first step of the proposed method.

--- proposed method BioVoice

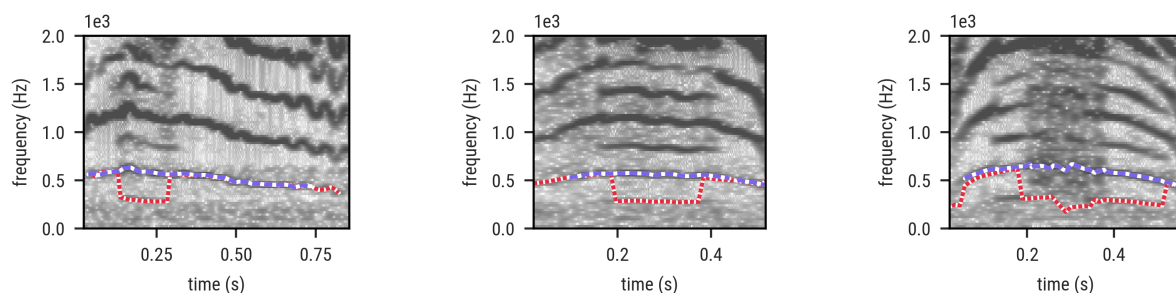
The proposed method is better

HIGH ENERGY HARMONICS - Harmonics are defined as multiples of the fundamental frequency, they occur in phonation cries and can have high energy. Therefore, the estimation resulting from BioVoice, which is based on energy, tends to jump to these upper high energy frequencies whereas the proposed method provides continuity in the estimation. (Figures 5.14).



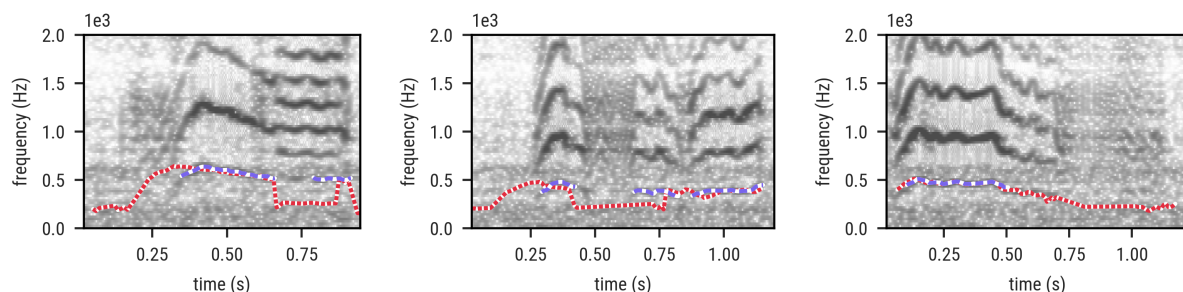
Figures 5.14: Cries with high energy harmonics.

HIGH ENERGY SUBHARMONICS - Subharmonics or double harmonic breaks are defined as a simultaneous parallel series of harmonics in-between the harmonics of the fundamental frequency [14, 44, 45]. In the case of such cries, the proposed method provides continuity in the estimation whereas BioVoice tends to jump to these in-between frequencies (Figures 5.15).



Figures 5.15: Cries with subharmonics.

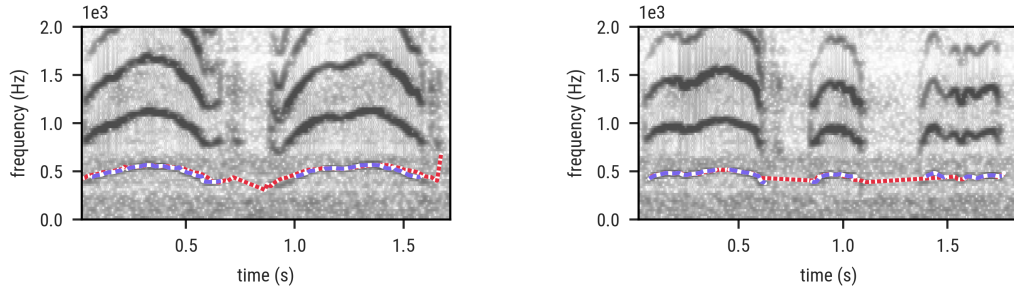
VIBRATIONS - It is common for babies to produce vibrations when crying. Three examples are given in Figures 5.16 with vibrations occurring at the start (left), middle (center), or end (right) of the cries. In such cases, the BioVoice method tends to track something within the vibration region while the proposed method correctly estimates the phonation parts only.



Figures 5.16: Cries with breaks are estimated in the middle by BioVoice.

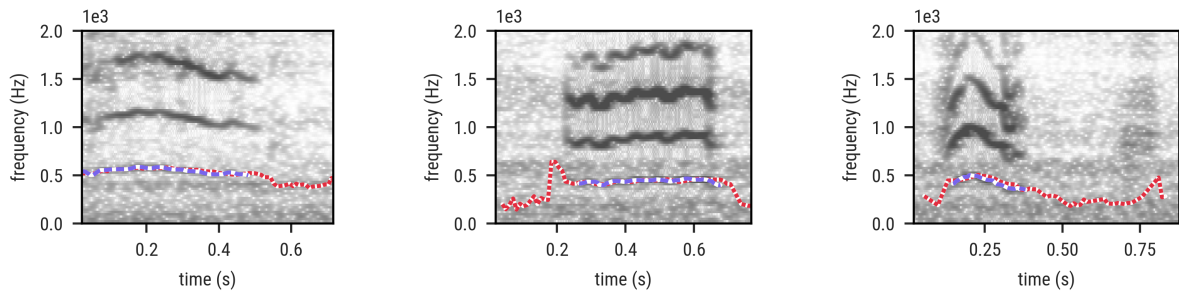
--- proposed method BioVoice

MULTIPLE CRY UNITS - Due to the segmentation method, some audio segments may contain several cries. In such case, the proposed method correctly detects and estimates the cry units whereas BioVoice tends to track the whole segment as a single cry unit (Figures 5.17).



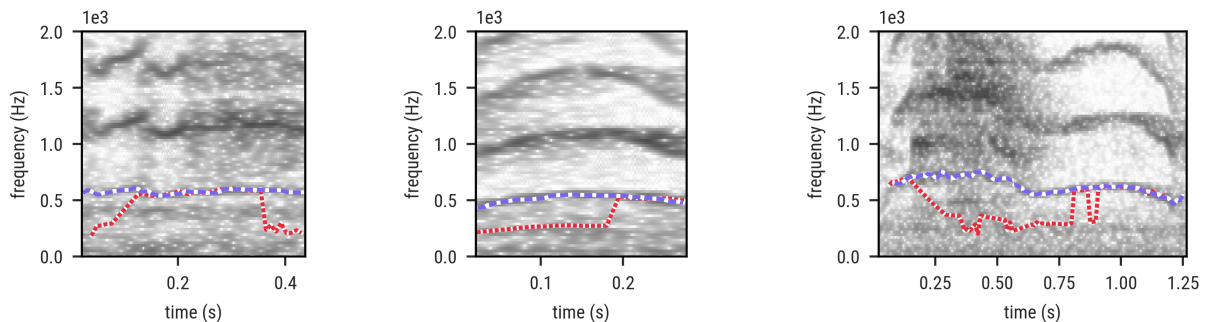
Figures 5.17: Segments with multiple cry units estimated as a single one by BioVoice.

EXTENDED SEGMENTATION - In addition, some of the audio segments resulting from the segmentation method are longer than the original cries. Due to this extended segmentation, only a part of the audio segment corresponds to the cry unit with the fundamental frequency to track. Here again, the proposed method seems to be better at detecting the fundamental frequency start and stop points whereas BioVoice estimation tracks F_0 along the whole segment (Figures 5.18).



Figures 5.18: Cries with partial F_0 are estimated over the whole cry unit by BioVoice.

LOW-FREQUENCY NOISE - The proposed method seems to work better in the case of cries with low-frequency noise where BioVoice tends to jump in unclear lower frequencies. (Figures 5.19).



Figures 5.19: Cries with jumps in the BioVoice frequency estimations due to low-frequency noise.

--- proposed method BioVoice

5.5.3 Performance and parameters comparison

The qualitative comparison showed good results, yet it remains a visual assessment, annotator dependent. To investigate further, we also propose a statistical parameters comparison of the tracking obtained with the two methods. Results of the parameters computed on each cry for all the 806 cries, defined as CASE 1, are illustrated in [Figure 5.20a](#) with the proposed method designated as *PM*.

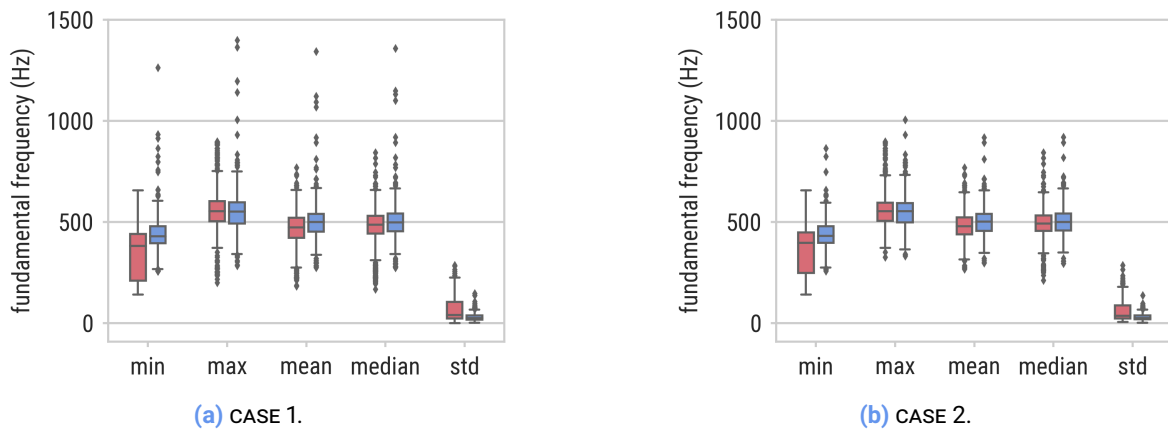


Figure 5.20: Fundamental frequency tracking parameters comparison between BioVoice (left) and the proposed method (right) with F_0 parameter values computed for each cry for the 806 cries.

At this stage we computed Pearson's coefficient ([Table 5.2](#)) which did not reveal a linear correlation between the different parameters of the two methods. Therefore we investigated the same comparison by taking into account only the cries that were correctly estimated by both methods (CASE 2). The new results computed on 695 cries are illustrated in [Figure 5.20b](#) and the new Pearson's coefficients are also given in [Table 5.2](#).

	CRIS	MIN	MAX	MEAN	MEDIAN
CASE 1	806	0.21	0.56	0.50	0.49
CASE 2	695	0.31	0.84	0.78	0.83

Table 5.2: Person's coefficient for both cases.

This time, Pearson's coefficient revealed a high positive linear correlation between the maximum, mean, and median values. Indeed, the overall comparison shows similar results for these three parameters in both methods. One can note that values resulting from the BioVoice method tend to be lower than ours and this result is exacerbated for the minimum parameter. According to the examples given when our method is better [Section 5.5.2](#) this tendency can be explained by the many cases where the BioVoice F_0 estimation drop to low-frequency noise.

5.6 Conclusion

In this chapter, we introduced a new method for fundamental frequency tracking of infant cries in the context of real-time monitoring in the NICU. The particularity of the proposed method consists of an initial step performed to automatically find the relevant frequency band in which to perform the F_0 tracking. Indeed, this band has been proven to be very important to achieve good estimations. Then, once this band is computed, the fundamental frequency tracking is performed using a contour detection in the spectrogram.

For validation of the proposed method, we compared our F_0 estimation results to those computed by the BioVoice software which we identified as the reference program for cry analysis. Indeed, the method developed by Manfredi et al. achieved good performances on synthetic basic melodic forms of newborn cries [37, 42].

With a selection of 806 cries recorded in a preterm infant, we evaluated results, both, in terms of a quantitative visual comparison as well as a common parameters comparison. First, our visual inspections of the fundamental frequency tracking superimposed on the cry spectrograms showed good results with correct estimation rates of 87% with BioVoice and 97% with the proposed method. Then, comparing both methods, we reported that F_0 estimations were equivalent for 44.67% of the cries evaluated and that one of the two methods was better than the other in 10.55% of the cases for BioVoice and in 44.79% for our method.

In addition, the visual comparison of the common parameter distribution showed similar results for both methods except for the minimum which was generally lower in BioVoice. We explain this difference by the fact that despite some jumps in the F_0 estimations, cries were considered as correctly detected during the visual annotation. Indeed, for some of the BioVoice mis-detections, we considered the whole signal without taking into account the small jumps occurring in the cases of high energy harmonics, vibrations, or low-frequency noises for example. Nevertheless, through a calculation of the Pearson coefficient, we found a linear correlation for the maximum, mean, and median parameters when comparing the cries which were judged with a correct visual estimation.

Therefore, in this chapter, we presented a new automatic fundamental frequency tracking method for the purpose of long-time monitoring in the NICU. Thanks to the initial step, we ensure that we perform an estimation of the F_0 in a relatively relevant frequency band selection, which has been proven by results consistent with the literature. Yet, if this step seems relevant for future calculations of harmonics, we also believe that the parameters defined here through experimentation will require optimization later.

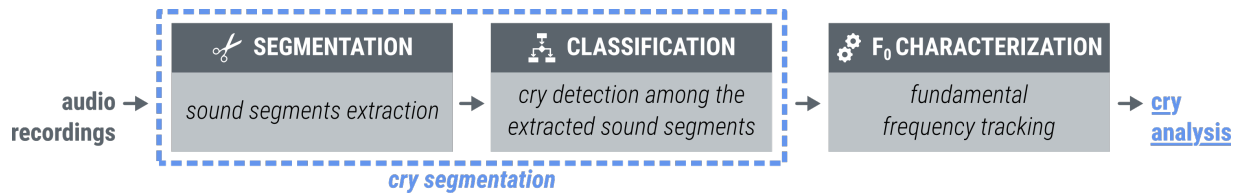
BIBLIOGRAPHY

- [1] MORELLI M.S., ORLANDI S., AND MANFREDI C. Biovoice: A multipurpose tool for voice analysis. *Biomedical Signal Processing and Control*, vol. 64, page 102302 (2021).
- [2] GRUNAU R.V. AND CRAIG K.D. Pain expression in neonates: Facial action and cry. *Pain*, vol. 28, 395–410 (1987).
- [3] BOERSMA P. PRAAT, a system for doing phonetics by computer. *Glott international*, vol. 5, 341–345 (2002).
- [4] 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*, vol. 7, 43–49 (2012).
- [5] 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*, vol. 10 (2014).
- [6] 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* (2016).
- [7] 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*, vol. 142, 1318–1331 (2017).
- [8] SALEHIAN MATIKOLAIE F. AND TADJ C. On the use of long-term features in a newborn cry diagnostic system. *Biomedical Signal Processing and Control*, vol. 59, page 101889 (2020).
- [9] 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*, vol. 2018, 1–14 (2018).
- [10] MICHELSSON K. Cry analyses of symptomless low birth weight neonates and of asphyxiated newborn infants. *Acta Pædiatrica*, vol. 60, 9–45 (1971).
- [11] MICHELSSON K., JÄRVENPÄÄ A., AND RINNE A. Sound spectrographic analysis of pain cry in preterm infants. *Early Human Development*, vol. 8, 141–149 (1983).
- [12] THODÉN C.J., JÄRVENPÄÄ A.L., AND MICHELSSON K. Sound spectrographic cry analysis of pain cry in prematures. In *Infant Crying*, 105–117. Springer (1985).
- [13] GRUNAU R.V., JOHNSTON C.C., AND CRAIG K.D. Neonatal facial and cry responses to invasive and non-invasive procedures. *Pain*, vol. 42, 295–305 (1990).
- [14] MICHELSSON K. AND MICHELSSON O. Phonation in the newborn, infant cry. *International Journal of Pediatric Otorhinolaryngology*, vol. 49 Suppl 1, S297–301 (1999).
- [15] RUNEFORS P., ARNBJÖRNSSON E., ELANDER G., AND MICHELSSON K. Newborn infants' cry after heel-prick: Analysis with sound spectrogram. *Acta Paediatrica*, vol. 89, 68–72 (2000).
- [16] FULLER B.F. Acoustic discrimination of three types of infant cries. *Nursing Research*, vol. 40, 156–160 (1991).

- [17] 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*, vol. 83, 204–11 (1994).
- [18] FULLER B.F. AND HORII Y. Spectral energy distribution in four types of infant vocalizations. *Journal of Communication Disorders*, vol. 21, 251–61 (1988).
- [19] 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*, vol. 42, 850–61 (1999).
- [20] VÁRALLYAY G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, vol. 71, 1699–1708 (2007).
- [21] 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*, vol. 1, 313–316. IEEE (2004).
- [22] CNOCKAERT L., SCHOENTGEN J., AUZOU P., OZSANCAK C., DEFEBVRE L., AND GRENEZ F. Low-frequency vocal modulations in vowels produced by parkinsonian subjects. *Speech Communication*, vol. 50, 288–300 (2008).
- [23] FALEK L., AMROUCHE A., FERGANI L., TEFFAHI H., AND DJERADI A. Formantic analysis of speech signal by wavelet transform. *Proceedings of the World Congress on Engineering 2011, WCE 2011*, vol. 2, 1572–1576 (07 2011).
- [24] MANFREDI C., D'ANIELLO M., BRUSCAGLIONI P., AND ISMAELLI A. A comparative analysis of fundamental frequency estimation methods with application to pathological voices. *Medical engineering & physics*, vol. 22, 135–147 (2000).
- [25] 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*, vol. 20, 85–99 (2015).
- [26] MANFREDI C., BOCCHI L., ORLANDI S., SPACCATERRA L., AND DONZELLI G.P. High-resolution cry analysis in preterm newborn infants. *Medical Engineering & Physics*, vol. 31, 528–32 (2009).
- [27] MARKEL J. The SIFT algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics*, vol. 20, 367–377 (1972).
- [28] KHEDDACHE Y. AND TADJ C. Identification of diseases in newborns using advanced acoustic features of cry signals. *Biomedical Signal Processing and Control*, vol. 50, 35–44 (2019).
- [29] 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*, 358–365. Springer (2004).
- [30] REYES-GALAVIZ O.F. AND REYES-GARCÍA C.A. Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system. In *Mexican International Conference on Artificial Intelligence*, 949–958. Springer (2005).
- [31] 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*, 876–881. Springer (2006).
- [32] 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*, vol. 17, 38–46 (2015).
- [33] 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*, MM '13, 835–838. ACM, New York, NY, USA (2013).

- [34] 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*, 2997–3001 (2016).
- [35] POKORNY F.B., BARTL-POKORNY K.D., EINSPIELER C., ZHANG D., VOLLMANN R., BÖLTE S., GUGATSKHA 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*, vol. 82, 109–119 (2018).
- [36] ORLANDI S., REYES GARCIA C.A., 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*, vol. 30, 656–663 (2016).
- [37] ORLANDI S., BANDINI A., FIASCHI F., AND MANFREDI C. Testing software tools for newborn cry analysis using synthetic signals. *Biomedical Signal Processing and Control*, vol. 37, 16–22 (2017).
- [38] 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*, vol. 11, 83–93 (2005).
- [39] MADHUSUDHANA S.K., ERBE C., AND GAVRILOV A. Spectrogram-based detection of signal contour tracks in underwater audio recordings (2013).
- [40] 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*, vol. 8, 799–810 (2013).
- [41] MANFREDI C., BOCCHI L., AND CANTARELLA G. A multipurpose user-friendly tool for voice analysis: Application to pathological adult voices. *Biomedical Signal Processing and Control*, vol. 4, 212–220 (2009). *New Trends in Voice Pathology Detection and Classification*.
- [42] 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*, vol. 45, 174–181 (2018).
- [43] MANFREDI C., VIELLEVOYE R., ORLANDI S., TORRES-GARCÍA A., PIERACCINI G., AND REYES-GARCÍA C. Automated analysis of newborn cry: relationships between melodic shapes and native language. *Biomedical Signal Processing and Control*, vol. 53, page 101561 (2019).
- [44] GOLUB H.L. AND CORWIN M.J. *A Physioacoustic Model of the Infant Cry*, In B.M. Lester and C.F. Zachariah Boukydis, editors, *Infant Crying: Theoretical and Research Perspectives*, 59–82. Springer US, Boston, MA (1985).
- [45] FUAMENYA N.A., ROBB M.P., AND WERMKE K. Noisy but effective: Crying across the first 3 months of life. *Journal of Voice*, vol. 29, 281–286 (2015).

Automatic processing for cry analysis: deployment



6.1 Introduction

In the last three chapters, we have presented the methods developed to build an automatic processing chain for the automatic analysis of infant crying. As we have seen, the automation of such a process is a real challenge considering the complexity of the sound environment where the recordings are made as well as the quantity of data to be processed. Therefore, we proposed a two-step crying segmentation method composed of a sound event extraction step followed by a classification of these events to detect crying. Once the cries are extracted, it remains to characterize them, thanks to the fundamental frequency tracking which is performed by a contour detection in the spectrogram.

The objective of this last chapter is therefore to present results from the deployment of the automatic processing chain for the purpose of crying analysis in a routine care environment. After a brief review of the literature on previous studies on the topic, we present the database used and the work that was done to manage the data. Then, in a first step, we propose to compare our results with those presented in the literature and performed by semi-automatic methods on preterm infants. These assessments are important because they compare the results of well-defined trials with our results obtained in the NICU environment. Thus, they show a real interest in using our strategy in a clinical context. Finally, we present new insights into the duration and fundamental frequency trends in time for the whole studied database. These results have also a very valuable clinical impact because this is the first time that such longitudinal trends of normal evolution cohorts are drawn.

6.2 State of the art

In this review, we focus on the studies related to our objective which is the evaluation of the maturation through cry analysis in preterm newborns. The characterization of crying in preterm infants has been extensively explored for *i)* the assessment of the evolution, *ii)* the early detection of pathologies, and *iii)* the comparison between full-term newborns. For the latter case, studies attempted to explain the differences observed in their neurophysiological maturity and the subsequent impact on their language development. As mentioned in **Chapter 1**, while studies investigated pain-induced crying, studies in recent decades concern the analysis of spontaneous crying and both topics are described in the following sections.

6.2.1 Pain induced cries

The first studies were not automatic and consisted of audio recordings performed at the induction of pain followed mostly by spectrographic analysis [1–3] or dedicated methods [4–7]. The pain cries were induced by a pinch at the infant arm [1], rubber band hit [4], pinch in the infant ear [2, 3], or during health check-ups such as heel-stick procedure [5, 6] and auditory brainstem response hearing screening test [7].

When comparing sick and healthy infants, Michelsson et al. showed in 1971 that the sick infant's cries were higher-pitched than those of symptomless premature babies, which were themselves higher pitched than those of healthy full-term newborns [1]. Later, Stevens et al. included two variables, severity of illness and behavioral state (sleep or awake) in the analysis [6]. The behavioral state was found to influence the facial action variables, and the severity of illness modified the acoustic cry variables.

Regarding the comparison between preterm and full-term infant crying, Tenold et al. could not show significant differences in fundamental frequency between the two groups. Yet, they did show greater variability in the preterm infant cry spectra which was interpreted as likely reflecting differences in neurophysiological maturity [4]. Later, Michelsson et al. showed that the cries of the smallest premature newborns were shorter, with higher fundamental frequency, and included bi-phonation and glide more often compared to control newborns [3]. Such results were also reported in [2], which shows that pain cries in preterm infants, observed between 31 and 33 weeks PMA, are higher-pitched than those of full-term infants. In addition, with increasing post-menstrual age comes an increase in the pain-cry duration and a decrease in the pain-cry fundamental frequency, which can represent a maturation of the central nervous system. Both last studies indicated that the cry characteristics changed with increasing post-menstrual age and the older the infant the more the crying pattern resembled that of the full-term [2, 3]. For their part, Goberman et al. identified clear differences in first spectral peak, mean spectral energy, and spectral tilt between full-term and preterm infants [7]. According to them, the observation of higher F_0 in the preterm infant's cries may either be related to smaller vocal folds, resulting from physical size differences

at birth or because preterm infants display a more stressful response to pain stimuli. In addition, through the evaluation of pain from facial expressions and crying performed in premature infants, but also in full-term and 2- and 4-month-old infants, Johnston et al. showed that: *i)* premature infants were different from older infants, *ii)* full-term newborns were different from others, but *iii)* 2- and 4-month-old were similar [5].

6.2.2 Spontaneous cries

The analysis of spontaneous crying is much more recent and only a few studies have investigated the subject [8–13]. In 2002, Wermke et al. compared the spontaneous crying of six preterm infants (three pairs of twins) recorded at different PMA (8-9 weeks, 15-17 weeks, and 23-24 weeks) [8]. Essential changes in the cries were observed from the 8th-9th week of life up to the 23rd-24th week of life, where they showed that the melody increased in complexity, from simple rising-falling patterns to composed patterns. This development was interpreted as an intentional articulatory activity related to neurophysiological maturation. In 2012, Orlandi et al. investigated if the distress occurring during crying in preterm newborns was related to central blood oxygenation. The results indicate that a similar decrease in oxygenation level occurs in both groups of patients, but that the recovery time after the crying episode is more stable and rapid in term infants than in preterm newborns. For their part, Shinya et al. inspected the effects of gestational age, body size at recording, and intrauterine growth retardation [11]. The acoustic analysis of spontaneous cries before feeding in both healthy preterm infants at term-equivalent ages and full-term newborns showed that shorter gestational age was significantly associated with a higher fundamental frequency, although no relation was found with smaller body size at recording or IUGR. Regarding the fundamental frequency and formants of preterm newborns, Manfredi et al. showed a decrease in frequency with increasing age that can be explained by increasing length and structural changes of both vocal folds and vocal tract [9] and Orlandi et al. showed that preterm newborn cries were generally higher than those of full-term infants [12]. Finally, in 2020, André et al. proposed the first vocal repertoire of preterm newborns in non-painful resting contexts [13]. They observed a broad range of vocalizations that they separated into nine vocal types distinguishable acoustically and non-randomly associated with behaviors.

6.3 Deployment of the proposed methods

In the following sections, we present the deployment of the overall processing chain for the purpose of crying analysis in a routine care environment. After presenting the data processed by the three methods, namely *i)* audio-video segmentation, *ii)* classification for cry detection, and *iii)* fundamental frequency characterization, we present our results. First, we propose to replicate existing studies, and then we give new perspectives for longitudinal tracking.

6.3.1 Database

A part of the Digi-NewB database was selected to drive a maturation study. To perform this selection, a rigorous examination of clinical records was made by clinicians in order to identify a subset of newborns without pathological development. Such a medical inspection is time-consuming as it requires observing the entire journey of the infants during their hospitalization. As a result, 57 newborns were selected and divided into five groups depending on the prematurity severity:

- EXTREME PRETERM (EP) - 7 newborns with GA between 24 weeks and 27 weeks + 6 days;
- VERY PRETERM (VP) - 12 newborns with GA between 28 weeks and 31 weeks + 6 days;
- LATE PRETERM (LP) - 16 newborns with GA between 32 weeks and 36 weeks + 6 days;
- EARLY TERM (ET) - 8 newborns with GA between 37 weeks and 38 weeks + 6 days;
- HEALTHY FULL-TERM (FT) - 14 newborns with GA greater than 39 weeks.

For each newborn, we processed all available recordings ranging in duration from a few hours to 10 consecutive days (related to the birth date, see the maturation recording protocol described in [Section 2.3](#)). It corresponds to 235 recordings with a total duration of 232 days, 13 hours, and 30 minutes or 11 163 WAV files of 30-minute duration. The database is illustrated in [Figure 6.1](#) with the detailed available recordings for each baby.

Since we work on a very large database, special attention was paid to data management. In the following sections, we give details of the data used through the whole processing chain.

6.3.2 Audio-Video segmentation

The sound segmentation step applied to the 235 recordings belonging to the 57 babies returned 3 548 006 sound segments corresponding to 27 days, 19 hours, 14 minutes, and 17 seconds. Thus, this step considerably reduces the amount of data to be further processed since the total duration of the sound segments corresponds to almost 12% of the total recording duration.

Then, when collecting only the sound segments occurring within motion periods, the number was reduced to 1 142 148, corresponding to a total duration of 9 days, 19 hours, 30 minutes, and 28 seconds. This step also reduces the data since only 32% of the sound segments occurred within the periods detected with babies' motion.

6.3.3 Classification for cry detection

To perform the automatic classification step, the selected sound segments are transformed into spectrograms and split in windows of 0.25 s with a 50% overlap, which gives a total of 5 432 892 images. After the classification of these spectrograms and the reconstruction of the predictions

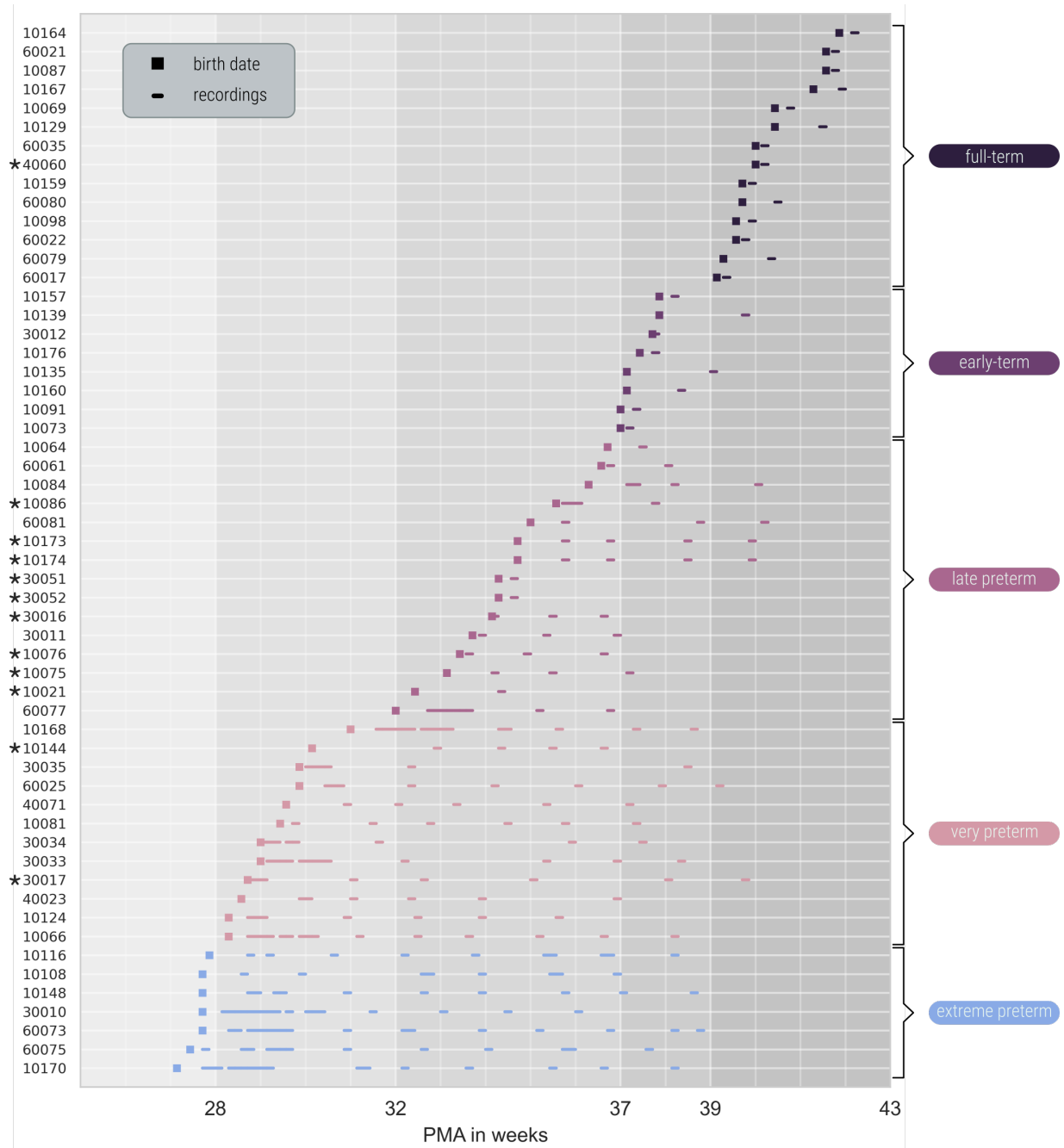


Figure 6.1: Illustration of the maturation database detailed for each of the 57 newborns. Babies with stars are recorded in shared-bedroom or co-bedding settings.

for the sounds, 117 947 audio segments were detected as crying among the 1 142 148 processed corresponding to a total duration of 1 day, 5 hours, 3 minutes, and 16 seconds. It has to be noticed that no crying was detected for one baby¹.

6.3.4 Data management

As the recordings vary in duration (from a few hours to 10 consecutive days), we decided to arrange the data in, what we decided to call, *periods* of up to 24 hours. Thus, the records whose duration exceeds this limit are split into several *periods*. Therefore, we now consider 278 periods recorded in 56 babies. For each period, the post-menstrual and postnatal ages are computed at the starting date of the corresponding periods.

From there, the detected cry distribution in a period varies from one cry to 2 687 cries. To ensure consistency in the cry analysis, we chose to remove periods in which fewer than 10 cries were detected. In addition, after listening to some recordings with overlapping cries, we realized that some rooms had several infants. With the help of the clinicians, these recordings (usually in the case of twins) were identified and removed from this study. Removing the recordings guarantees that the cries studied correspond to the selected babies, which is very important for the analyses performed later.

Therefore, the resulting database is composed of 221 periods recorded in 43 babies with cry distribution in a period varying from 10 to 2 687 cries with a total of 93 691 cries corresponding to a duration of 21 hours 31, minutes and 51 seconds.

6.3.5 Fundamental frequency characterization

The proposed fundamental frequency tracking method is applied to the 93 691 detected cries, then, for each cry, the minimum, maximum, mean, median, and standard deviation values are computed. Next, for each period statistical values are combined through mean and median values. Finally, all further studies are based on these extracted values considering either periods or infants, in that case, period values are averaged for each infant. This process is illustrated in [Figure 6.2](#).

1. No crying was detected for baby 010038.

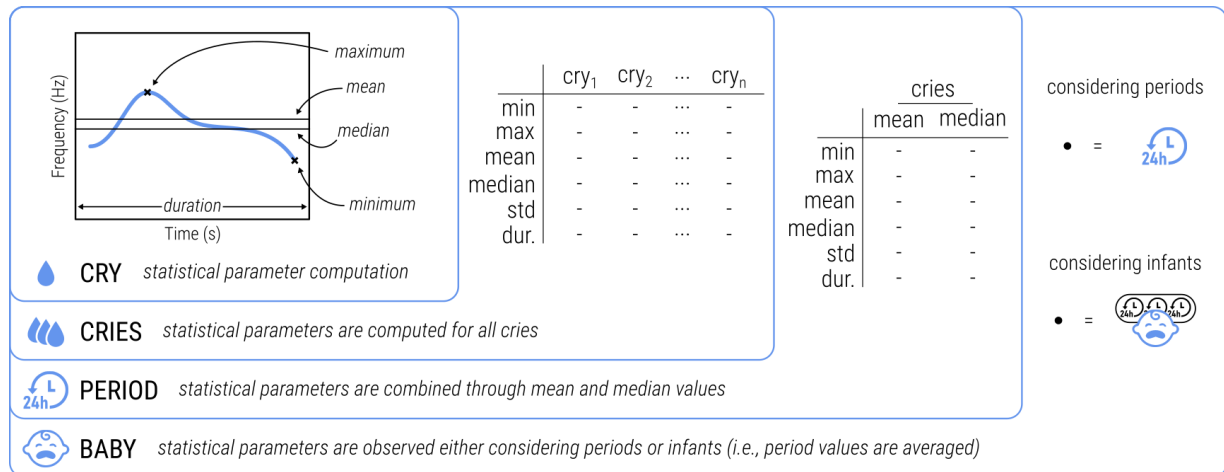


Figure 6.2: Fundamental frequency statistical parameters computation workflow.

6.4 Reproduction of existing studies

The proposed work is the first one, to our knowledge, to propose a fully automatic processing chain for crying analysis in the NICU. Nevertheless, as mentioned in the state of the art, very few studies have focused on the crying analysis in preterm infants. In this section, we would like to reproduce some studies from the literature in order to validate the proposed cry analysis processing chain from recordings performed in a noisy context of a routine hospital care environment. Two important studies for further clinical uses are reported. The first evaluates the fundamental frequency for newborns by comparing two groups with distinct ages and birth weights [9] while the second proposes to observe the minimum, maximum, and mean parameters of the F_0 at term equivalent age [11].

6.4.1 Fundamental frequency according to GA and birth weight

In their study, Manfredi et al. tried to understand if gestational age (g.a.) and/or weight at birth (w.a.b.) can influence newborn cries [9]. They analyzed a group of 18 preterm newborns without relevant pathology, with gestational age ranging from 23 to 38 weeks and birth weight between 590 and 3 020 g and a small control group composed of 2 full-term healthy infants (GA greater than 37 weeks, weight at birth greater than 3 000 g). Through several audio recordings for each infant (with PNA less than one month and in an open bed) they manually selected about 60 cry episodes, defined as a sequence of high energy utterances of approximately 5–6 s each. The analysis of all available data provided consistent indication for a cut-off point of 34 weeks GA and 2500 g for weight at birth. They showed a decrease in frequency with increasing age or weight for all the parameters (i.e., F_0 , F_1 , F_2) and explained this result with increasing length and structural changes of both vocal folds and vocal tract. Their results are presented in Figure 6.3 for a F_0 comparison between the groups.

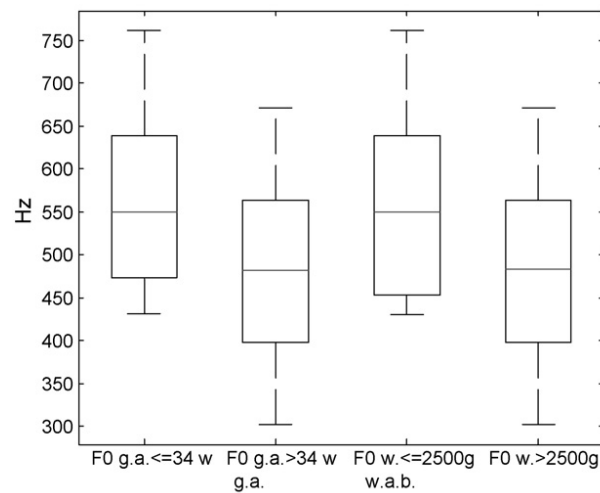


Figure 6.3: From Manfredi et al. [9], boxplots comparing the fundamental frequency for the newborn cry data, divided according to gestational age (g.a.) and weight at birth (w.a.b.). Results consistently show a decrease in frequency with increasing age or weight.

To replicate this study, we extracted our data into two subsets that matched at best the conditions of the Manfredi et al. dataset. Considering the gestational age, we selected all the periods corresponding to preterm infants while for birth weight, we selected periods corresponding to newborns for whom weight information was available. In both cases, periods were selected with recordings performed during the first month of life (i.e., PNA less than 30 days), the databases used are described in [Table 6.1](#).

	≤ 34 w.	$34 < GA < 37$ w.	TOTAL		≤ 2.5 kg	> 2.5 kg	TOTAL
BABIES	19	3	22	BABIES	25	16	41
PERIODS	131	5	136	PERIODS	139	16	155
CRIS	35 239	2 371	37 610	CRIS	39 369	6 763	46 132

(a) Database considering gestational age.

(b) Database considering weight at birth.

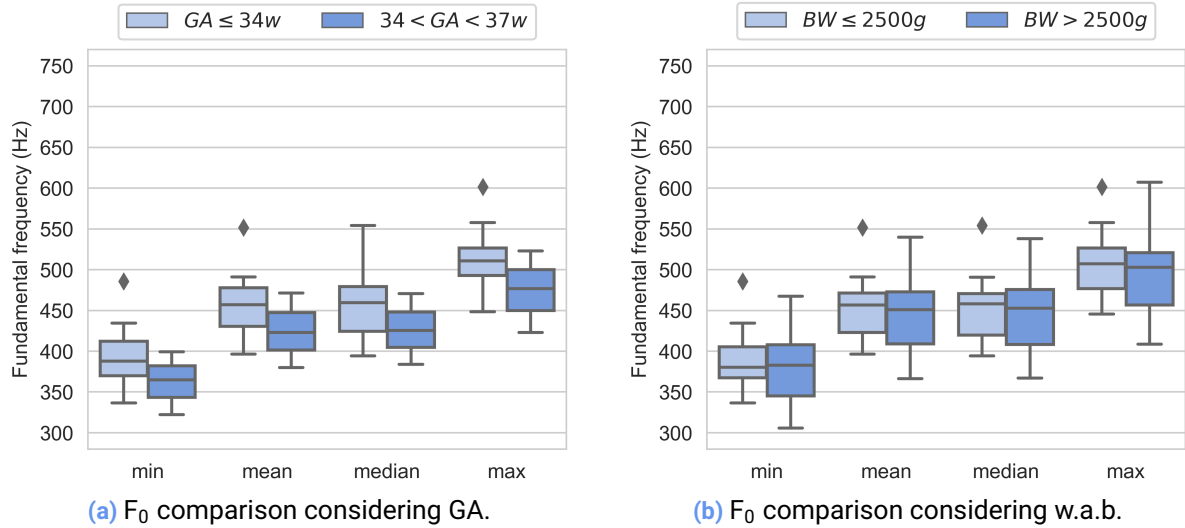
Table 6.1: Databases for the comparison of fundamental frequencies with gestational age and weight at birth cut-offs.

Fundamental frequency results of our database are presented in [Figure 6.4a](#) considering the gestational age and in [Figure 6.4b](#) considering the weight at birth. We decided to show all the statistical parameters since we don't know which one is presented in [Figure 6.3](#). In our case, statistical parameters are averaged for each infant, and cry values are combined per period through median values. These results show a decrease in frequency with increasing age or weight for all statistical parameters, which is consistent with what Manfredi et al. reported in their study [9]. However, the results when comparing infants with a birth weight of less or greater than 2500 g are less significant.

Moreover, it has to be mentioned that the F_0 value distribution is located in much lower fundamental frequency values in our case, which is surprising since our frequency range of analysis is

much larger and higher (250-1500 Hz) than the one used by Manfredi et al. (150-900 Hz).

It is worthwhile to mention that the results proposed in this section are important and introduce a range of possible values for some babies presenting a normal evolution. Further studies would be to assess the fundamental frequency of abnormal babies.



Figures 6.4: Boxplots comparing the fundamental frequency for the newborn cry data, divided according to gestational age and weight at birth.

6.4.2 Fundamental frequency at term-equivalent age

In their study, Shinya et al. proposed to observe the crying fundamental frequency minimum, mean and maximum values in preterm and full-term newborns at an equivalent age [11]. The analysis performed on 2 321 manually extracted cries recorded in 64 babies with a PMA greater than 37 weeks and lower than 42 weeks, showed that shorter gestational age was significantly associated with higher F_0 . These results are illustrated in [Figure 6.5a](#), where the very preterm newborns (white circles) have higher fundamental frequency values than the full-term infants (black circles).

To reproduce this study, we selected all the *periods* recorded at the same equivalent age (i.e., PMA between 37 and 42 weeks). In addition, since Shinya et al. used a frequency range between 150 and 900 Hz, we decided to remove all cries with a fundamental frequency maximum above 900 Hz. The database used is divided according to the GA into three groups (i.e., two preterm groups and one full-term) and is described in [Table 6.2](#).

Results of the fundamental frequency are presented in terms of minimum, mean, and maximum for Shinya et al. in [Figure 6.5a](#), and with statistical parameters averaged for each infant with cries values combined per period for our database through the mean in [Figure 6.5b](#) and the median values in [Figure 6.5c](#).

	GA < 32 w.	32 w. ≤ GA < 37 w.	GA ≥ 32 w.	TOTAL
BABIES	13	3	20	36
PERIODS	20	4	21	45
CRIS	15 614	1 625	8 870	26 109

Table 6.2: Crying duration evolution database according to GA divided in two preterm and one full-term infant groups.

Unfortunately, by removing the babies in double rooms, we have reduced the number of moderate-to-late preterm infants to be analyzed. In addition, in the chosen baby subset, no preterm infant was born before 27 weeks. Therefore, the data distribution is quite sparse and does not cover the gestational age range with the same efficiency as in the original study [11]. However, where Shinya et al. offer analysis for 64 babies with 2 321 manually selected cries, our analysis is performed for 26 109 cries automatically extracted from 36 babies.

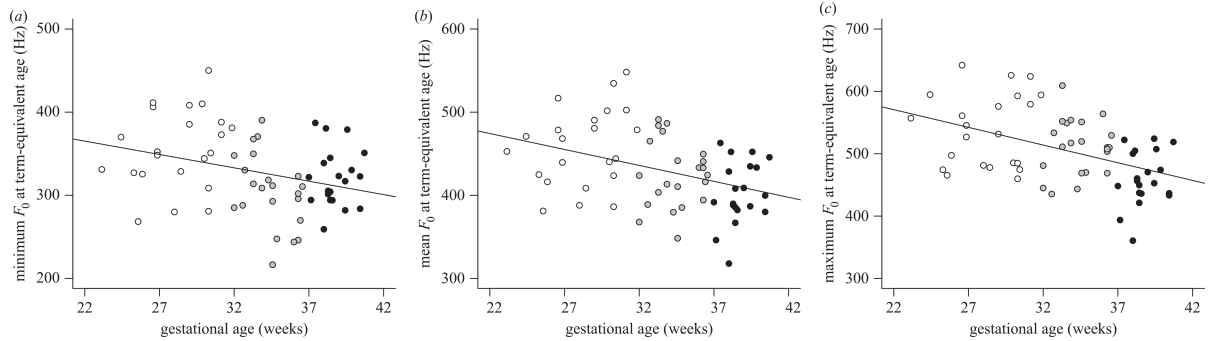
When comparing the results, the trends of the three parameters are the same for both cases (mean and median). However, we qualitatively observe that the median case offers results closer to those proposed by Shinya et al. (Figure 6.5a). This interpretation is, of course, subjective since it is inconsistent to compare mean and median values.

One can also observe a very similar decay in the regression line for the median of the minimum F_0 but with a significant shift in high frequencies for our method. Indeed, based on the numerical results presented in Table 6.3 most of our parameters have higher values than those of the original method, especially for the minimum and the mean. This might be due to the frequency range where the fundamental frequency analysis is performed which is lower for Shinya et al. (150–900 Hz) than us (250–1500 Hz).

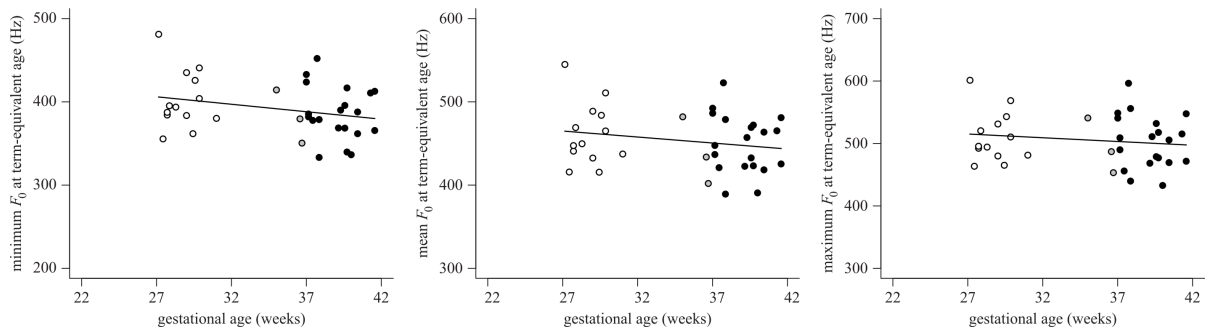
While Shinya et al. showed that shorter gestational age was significantly associated with higher F_0 we were unable to retrieve such a conclusion. Nevertheless, despite a total automatic analysis performed in the routine care environment, we were able to reproduce the same evolution.

This opens the door to future comparison to better understand why preterm birth is associated with an increase in the fundamental frequency of spontaneous cries at term-equivalent age. From a clinical perspective, Shinya et al. suggested several explanations. First, it might be due to a longer postnatal period. Second, it might reflect the reduced vagal activity in preterm infants. They reported that the vagal input has an inhibitory effect on laryngeal contraction and results in vocal fold tightening. Thus, the decreased vagal activity is assumed to cause increased vocal fold tension and higher F_0 .

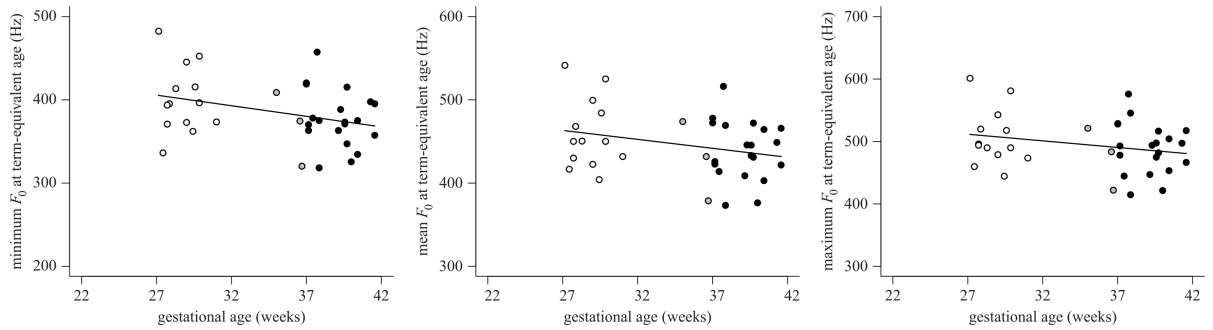
The Digi-NewB database, composed of ECG, respiratory signals, and cries simultaneously recorded, opens the door to exciting studies that could be performed with our proposed method to assess the influence of vagal activity.



(a) Averaged for each infant of the mean values (n=64), from Shinya et al. [11].



(b) Averaged for each infant of the mean values (n=36), Digi-NewB database.



(c) Averaged for each infant of the median values (n=36), Digi-NewB database.

Figures 6.5: Scatter plots showing the relationships between gestational age and minimum (left), mean (center) and maximum (right) fundamental frequency F_0 of spontaneous cries at term-equivalent age. The groups of infants were VP (white circles), MLP (grey circles) and FT (black circles).

	PRETERM						FULL-TERM		
	GA < 32 weeks (n = 22)			32 ≤ GA < 37 weeks (n = 22)			GA ≥ 37 weeks (n = 20)		
	mean	s.d.	range	mean	s.d.	range	mean	s.d.	range
minimum F ₀ (Hz)	356	48	268-450	306	44	217-390	321	35	259-387
mean F ₀ (Hz)	458	47	381-548	425	40	348-491	403	38	318-463
maximum F ₀ (Hz)	539	59	460-642	511	44	435-609	460	44	361-524

(a) Results from Shinya et al. [11].

	PRETERM						FULL-TERM		
	GA < 32 weeks (n = 14)			32 ≤ GA < 37 weeks (n = 8)			GA ≥ 37 weeks (n=21)		
	mean	s.d.	range	mean	s.d.	range	mean	s.d.	range
minimum F ₀ (Hz)	406	34	356-482	383	32	339-422	382	34	333-452
mean F ₀ (Hz)	466	37	415-544	448	41	398-501	445	37	385-522
maximum F ₀ (Hz)	517	41	464-602	506	46	448-566	499	44	431-597

(b) Our results.

Table 6.3: Comparison of the fundamental frequency of spontaneous crying according to GA divided into two preterm and one full-term infant groups for both methods.

6.5 New cry characterization insights

In this section, we propose new visualizations of the fundamental frequency and duration evolution of the spontaneous cries from our database. First, in their study, Shinya et al. suggested that the increased F₀ of spontaneous cries is not related to the body size, but rather might be owed to their different intrauterine and extrauterine experiences [11]. Thus, we propose to investigate the difference between F₀ values of preterm newborns recorded at birth and after some time living an extrauterine to a group of infants newly born at term. Then, we investigate the general evolution of cry duration and fundamental frequency with increasing post-menstrual for all populations and with increasing postnatal age for the extreme preterm group. Finally, we propose to observe a longitudinal evolution of the fundamental frequency with increasing post-menstrual age, a representation that has never been done before to our knowledge.

6.5.1 Fundamental frequency comparison at two postnatal ages

The objective of this study is to compare the fundamental frequency of preterm newborns at birth and with a certain experience of extra-uterine life to infants newly born at term. To perform this study, we separate the database into three subsets corresponding to:

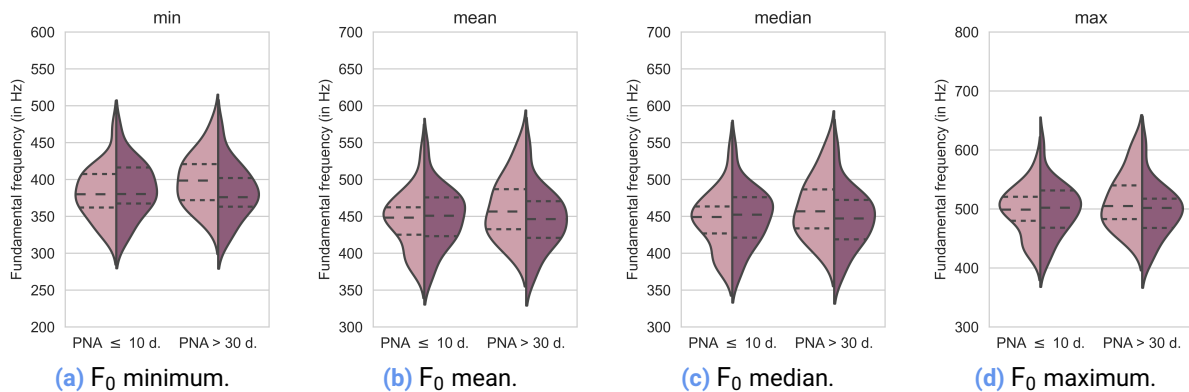
- preterm newborns recorded at birth with a PNA less than 10 days;
- preterm newborns recorded with a PNA greater than 30 days;
- full-term newborns recorded at birth with a PNA less than 10 days.

The data used are described in detail in [Table 6.4](#).

	PRETERM		FULL-TERM
	PNA \leq 10 days	PNA $>$ 30 days	PNA \leq 10 days
BABIES	24	19	19
PERIODS	81	64	19
CRIES	22 881	46 604	7 775

Table 6.4: Database used to assess F_0 at two postnatal ages.

Results of the fundamental frequency are computed with statistical parameters (i.e., minimum, maximum, mean, and median) averaged for each infant and cries values combined per period through median values. The results are presented in **Figures 6.6** as two pairs of violin plots, with preterm newborns represented by the left-hand sides and full-term infants by the right-hand sides. A violin plot is an attractive way to represent the data distribution since it draws a combination of a boxplot and a kernel density estimate. In the left pair, preterm newborns registered at birth are compared to full-term newborns while, in the right pair, preterm newborns with a postnatal age greater than 30 days are compared to full-term newborns. In both pairs, we use the same full-term infant distributions (i.e., PNA less than 10 days), which are duplicated for better visualization.



Figures 6.6: Violinplots comparing the fundamental frequency for the newborn cry data, divided according to post-natal age and prematurity status. For each pair, left-hand side distributions (light) represent preterm newborns while right-hand side distributions (dark) represent infants newly born at term. In the left pair, preterm newborns are observed at birth and in the right pair preterm newborns are observed with a postnatal age greater than 30 days.

The results are in agreement with Shinya et al. who suggested that intrauterine and extrauterine experiences might have an impact on the fundamental frequency of spontaneous cries [11]. Indeed, the results show that the fundamental frequency is quite similar when comparing preterm to full-term infants newly born. However F_0 is higher in preterm infants who have already experienced an extra-uterine life.

For their part, Orlandi et al. showed higher parameter values for preterm than those of full-term infants even when the preterm reaches a post-menstrual age similar or equal to that of the term infant (between 35 weeks and 43 weeks) [12]. They suggested that prematurity might create a delay in the neuromotor control development in the preterm infant, who would therefore need more than the expected birth age to fully recover.

6.5.2 Crying evolution with age

This study aims to observe the fundamental frequency and duration of spontaneous crying evolution for all infants, preterm and full-term newborns, and for all recordings included in the maturation database. Data used in this study are described in [Table 6.5](#).

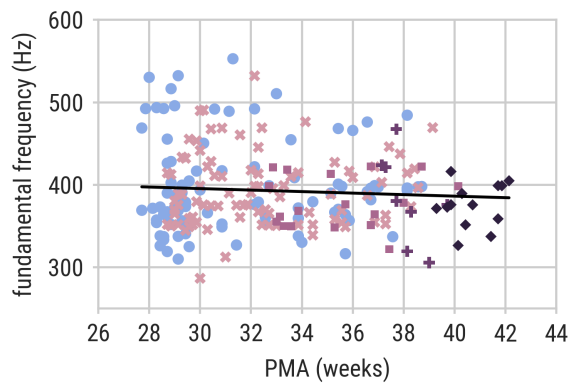
	EP	VP	LP	ET	FT	TOTAL
BABIES	7	10	5	8	13	43
PERIODS	92	90	18	8	13	221
CRIS	41 338	34 244	8 632	3 917	5 560	93 691

Table 6.5: Database used to assess crying evolution.

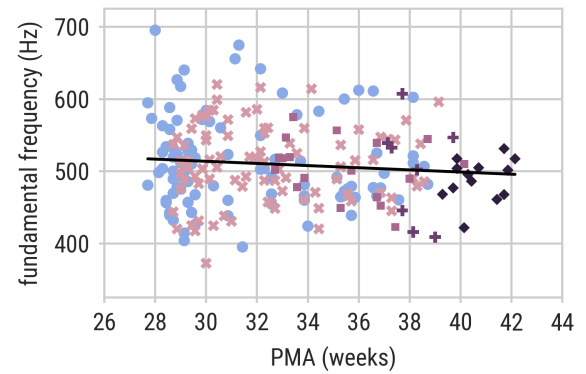
Fundamental frequency evolution with PMA

Here we present values of the fundamental frequency in preterm and full-term newborns with increasing post-menstrual age. Results of the F_0 are computed with statistical parameters (i.e., minimum, maximum, mean, and median) combined per period through median values. The results are presented in [Figures 6.7](#).

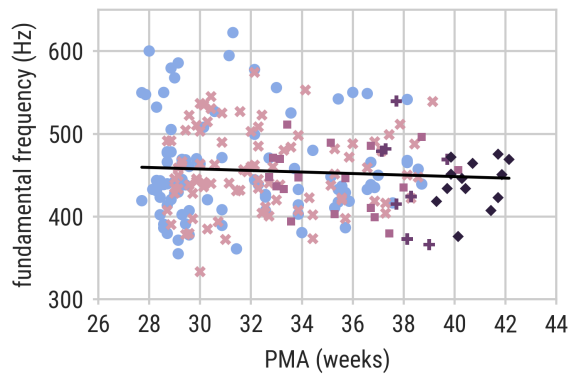
We can observe that mean F_0 tends to decrease with increasing PMA and that the values are more dispersed for low PMA. This is in agreement with the literature. Actually, Tenold et al. suggested that the differences observed in spectral variability between cries of premature and full-term infants probably reflect neurophysiological maturity [4]. In addition, in their study, Thoden et al. showed, in the case of pain cries, that the more premature a newborn is, the higher the fundamental frequency and that it decreases with increasing post-menstrual age [2]. A statement in accordance with the previous work proposed by Michelsson et al. who also mentioned that even if the cries of the smallest premature babies were generally high-pitched, they were also sometimes lower-pitched and thus look like the cries of infants born at term [3].



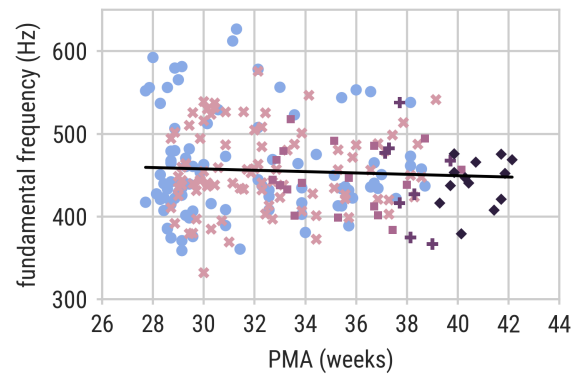
(a) Minimum fundamental frequency.



(b) Maximum fundamental frequency.

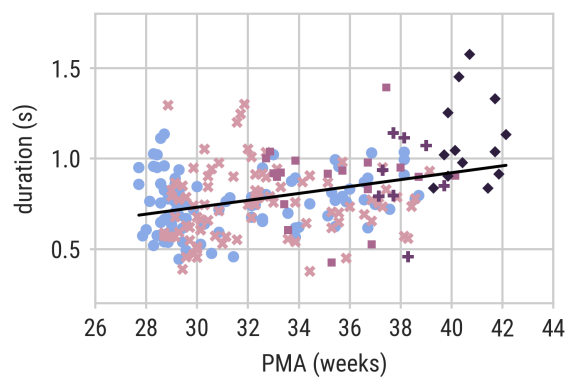


(c) Mean fundamental frequency.

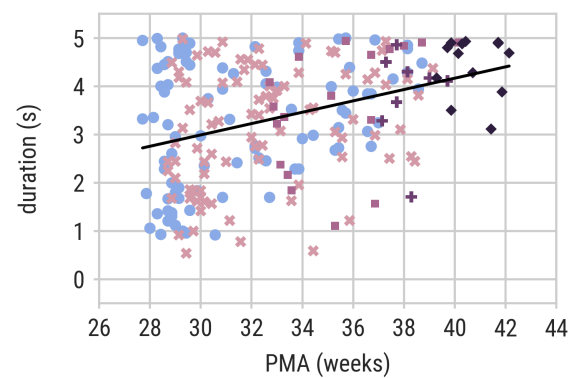


(d) Median fundamental frequency.

Figures 6.7: Cry F_0 evolution considering all newborns according to the PMA.



(a) Mean duration.



(b) Maximum duration.

Figures 6.8: Cry duration evolution considering all newborns according to the PMA.



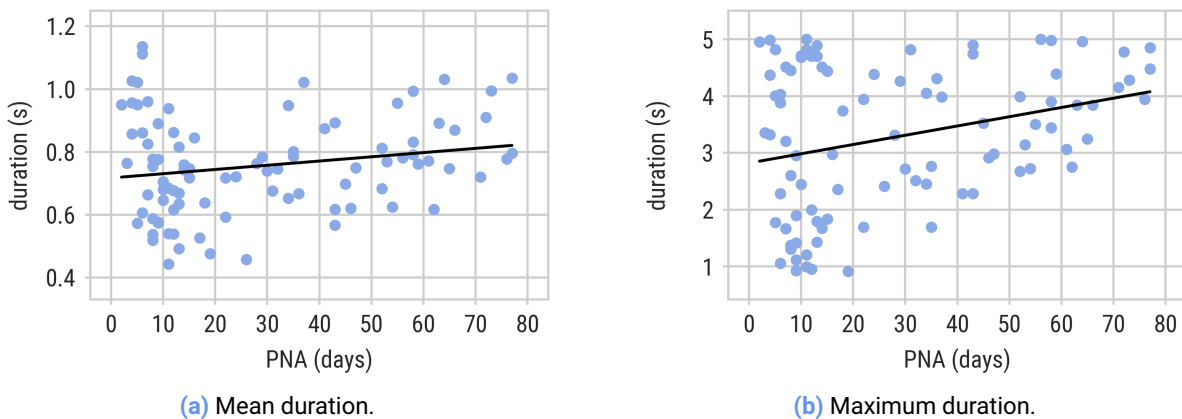
Duration evolution with PMA

To our knowledge, the duration of evolution in preterm infants has not been studied on spontaneous crying, however, studies performed on induced crying showed that crying duration increases with age [2, 3, 6]. In addition, preterm infant cries were considered almost identical to those of the full-term when reaching the age of 38 conceptual weeks [3] or when comparing the crying duration between preterm with PMA greater than 37 weeks and full-term newly born [7].

Cry duration results are presented with the mean (left) and maximum (right) duration values averaged for each 24-hour period as the function of the PMA in **Figures 6.8**. One can see that the mean cry duration is increasing with the post-menstrual age, in other words, the older the infant is the longest cry she/he can produce. When considering the maximum of the crying, one can see that the value is increasing with post-menstrual age. As a reminder, during the segmentation stage, only the sound segments whose duration is between 0.25 and 5 seconds are kept, which is why the maximum duration values are limited to 5 seconds. Although only a few periods of preterm newborns were recorded after 37 weeks of PMA, we can see, both for the mean and maximum durations, that the values tend to be closer to those of infants born at term.

Duration evolution with PNA in extreme preterm newborns

This study aims to focus on the spontaneous crying duration evolution during the hospitalization of the 7 extreme preterm infants (GA between 24 weeks and 27+6 weeks+days) included in the maturation database (i.e., 7 babies, 92 periods, and 41 338 cries). Results are presented with the mean (left) and maximum (right) duration values averaged for each 24-hour period as a function of the postnatal age in **Figures 6.9**.



Figures 6.9: Cry duration evolution focusing on the EP newborns according to the PNA.

From [Figure 6.9](#), we can see that the average crying time increases after birth and throughout the hospitalization period. It can also be seen that the maximum crying duration also increases with age, which is probably due to an increase in vocal power acquired during the infant's development.

From a clinical point of view, these results are interesting and could be useful to detect in real time certain states of the infant since sequences of crying of long duration are associated with states of stress [13] while crying with high fundamental frequencies are associated with pain [14] or pathologies [15]. In future works, this information could also be coupled with melody analyses that also allow a good detection of these different states.

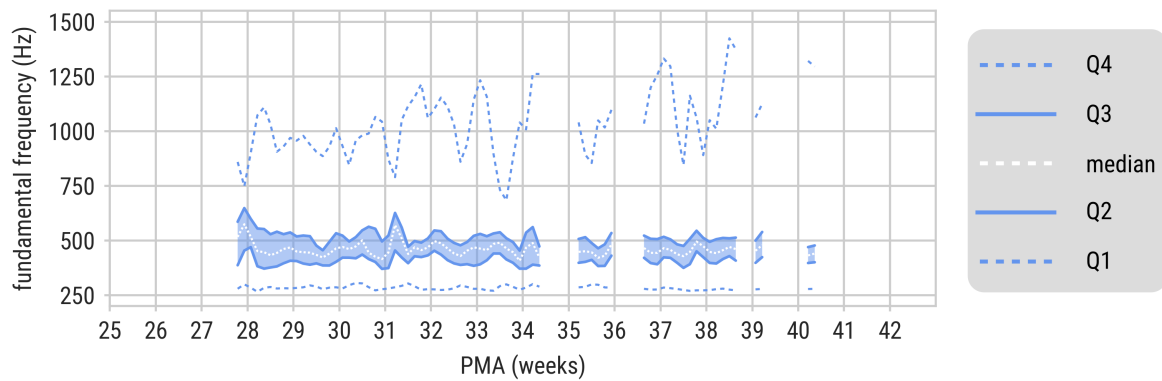
6.5.3 Fundamental frequency longitudinal evolution

The comparison with studies from the literature as well as the previous sections clearly showed evolution with GA and PMA. Therefore, in this section, we wonder if a longitudinal visualization of the fundamental frequency with increasing post-menstrual age could be interesting. As the normalized weight and height curves are used to monitor the growth of children, we propose to observe the general evolution of the fundamental frequency when combining all the available data. Therefore, the database used is based on the one described in [Section 6.3](#) including 43 babies, 221 periods, and 93 691 cries.

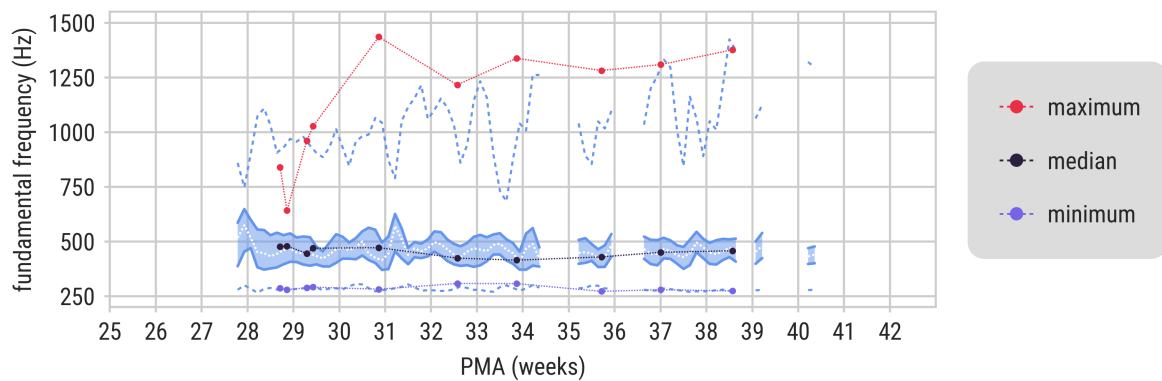
To carry out this study, additional statistical values are calculated for each cry, namely, the four quartiles Q1, Q2, Q3, and Q4. Based on the same process as before, the cry parameters are combined by period through the median values and all the fundamental frequency statistical values are averaged over the PMA through a two-day rolling average. The resulting process presenting the pseudo-normal evolution of the fundamental frequency observed in infants with no complication during their hospitalization in NICU is depicted in [Figure 6.10a](#).

Once these evolutionary trends are defined, we propose to superimpose the minimum, maximum, and median parameters for an extreme preterm ([Figure 6.10b](#)) and a very preterm newborn ([Figure 6.10c](#)) from the cohort.

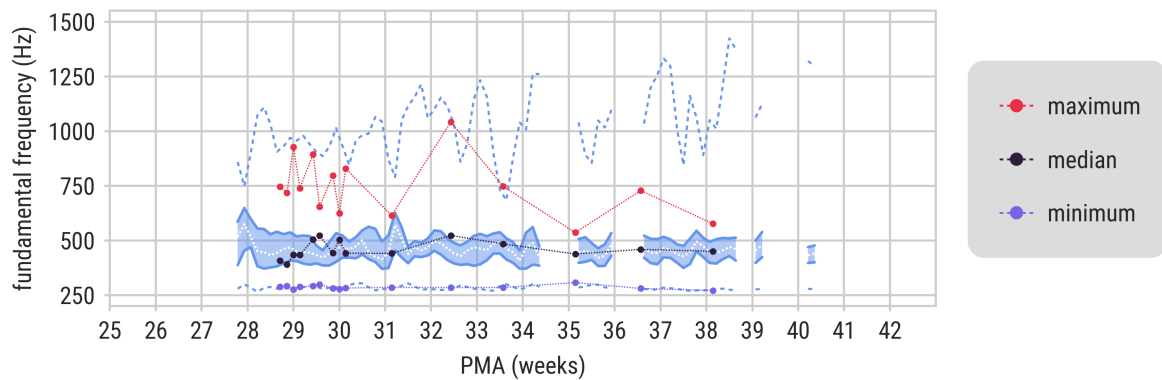
These results are the first to observe the longitudinal evolution of spontaneous crying in preterm infants. Although we have no conclusion to draw on the proposed examples, we think that, with such visualization, it would be interesting to follow infants and check whether abnormal courses have any impact on cry production.



(a) Averaged fundamental frequency evolutionary trends.



(b) Extrem preterm - 010148.



(c) Very preterm - 010066.

Figures 6.10: Averaged fundamental frequency evolutionary trends of the 43 healthy babies without reported complication during their hospitalization in NICU (a). Additional minimum, median and maximum values are superimposed for two babies: (b) an extreme preterm and (c) a very preterm.

6.6 Conclusion

Thanks to the deployment of the complete automatic processing chain, we reached in this chapter some relevant conclusions. First, we showed a comparison with the literature through two interesting and well-designed studies. While the first one compares preterm newborns for different post-menstrual ages and weights at birth, the second one compares preterm and full-term infants at term equivalent age. In both cases, our results are consistent with the literature which seems to demonstrate that the proposed signal processing chain is robust even in a noisy environment. Considering this powerfulness, we can encourage further clinical applications as well as the exploration of new issues.

Then, through new visualization of the duration and fundamental frequency evolution, we showed that the cry duration is increasing with increasing PNA and PMA while the fundamental frequency tends to decrease with PMA. Last but not least, we proposed to assess the pseudo-normal evolution of the fundamental frequency observed in infants without complications during their hospitalization in the NICU. This work has never been done before and gives new issues for the evaluation of sepsis and pathology during monitoring in the NICU.

It is worthwhile to remind that this is the first automatic processing chain created and deployed on such a large scale. In fact, while previous studies were based on the analysis of a few hundred cries or a few thousand (see [Table 1.1](#)), we presented results obtained on more than 90 000 cries.

BIBLIOGRAPHY

- [1] MICHELSSON K. Cry analyses of symptomless low birth weight neonates and of asphyxiated newborn infants. *Acta Pædiatrica*, vol. 60, 9–45 (1971).
- [2] THODÉN C.J., JÄRVENPÄÄ A.L., AND MICHELSSON K. Sound spectrographic cry analysis of pain cry in prematures. In *Infant Crying*, 105–117. Springer (1985).
- [3] MICHELSSON K., JÄRVENPÄÄ A., AND RINNE A. Sound spectrographic analysis of pain cry in preterm infants. *Early Human Development*, vol. 8, 141–149 (1983).
- [4] 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*, vol. 56, 975–80 (1974).
- [5] 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*, vol. 52, 201–208 (1993).
- [6] STEVENS B.J., JOHNSTON C.C., AND HORTON L. Factors that influence the behavioral pain responses of premature infants. *Pain*, vol. 59, 101–9 (1994).
- [7] 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*, vol. 42, 850–61 (1999).
- [8] WERMKE K., MENDE W., MANFREDI C., AND BRUSCAGLIONI P. Developmental aspects of infant's cry melody and formants. *Medical Engineering & Physics*, vol. 24, 501–14 (2002).
- [9] MANFREDI C., BOCCHI L., ORLANDI S., SPACCATERRA L., AND DONZELLI G.P. High-resolution cry analysis in preterm newborn infants. *Medical Engineering & Physics*, vol. 31, 528–32 (2009).
- [10] ORLANDI S., BOCCHI L., DONZELLI G., AND MANFREDI C. Central blood oxygen saturation vs crying in preterm newborns. *Biomedical Signal Processing and Control*, vol. 7, 88–92 (2012).
- [11] 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*, vol. 10 (2014).
- [12] ORLANDI S., REYES GARCIA C.A., 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*, vol. 30, 656–663 (2016).
- [13] ANDRÉ V., DURIER V., HENRY S., NASSUR F., SIZUN J., HAUSBERGER M., AND LEMASSON A. The vocal repertoire of preterm infants: Characteristics and possible applications. *Infant Behavior and Development*, vol. 60, page 101463 (2020).
- [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*, vol. 55, 142–146 (2004).
- [15] 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*, vol. 11, 83–93 (2005).

Conclusions & perspectives

In this manuscript, we focused on the presentation of a complete processing chain for the automatic characterization of cries in preterm newborns. This objective is in line with those of the European project Digi-NewB which aimed at combining clinical signs, physiological signals, and video and sound recordings in a decision support system for the monitoring of newborns. As a reminder, its two clinical targets were the early diagnosis of late sepsis and the objective assessment of maturation in premature babies cared for in neonatal intensive care units. If physiological signals (cardiac, respiration, ...) have already been widely studied to evaluate the risks and evolutions, the analysis of audio and video signals is more recent and tries to bring new clinical indicators. Thus, the work done during this thesis is the first one, to our knowledge, on the analysis of vocalizations in premature infants in a routine care setting in the NICU. This explains why it focused on the development of relevant methods for the automation of audio processing of the data collected during the Digi-NewB project.

The analysis of the literature showed that many studies underlined the interest of analyzing crying in infants to evaluate their neurobehavioral development and thus analyze their maturation stage. Although teams have already been interested in the spontaneous crying of premature infants, this is the first time that such an audio recording device has been set up in a NICU and that such a database has been created. The acquisition of data in a routine hospital care environment led us to the design of a new automatic processing chain composed of three steps. This chain gathers *i*) an audio segmentation step using video signal processing to extract only the sounds occurring within the infant's movement; *ii*) a classification step using a deep-learning approach for the detection of crying among the extracted sounds (adult voices, alarms, etc. and *iii*) a fundamental frequency characterization step using a contour detection in the spectrogram to track the cry F_0 .

The segmentation method developed was inspired by the one proposed by Orlandi et al. which is based on a calculation of the short-term energy followed by the Otsu method thresholding [1]. After removing the 30-minute audio files that do not contain sound, two steps are added to the method to improve it. The first step is a double frequency filter and the second step is a re-segmentation. The evaluation of the segmentation method in comparison with manual annotations, performed on three 30-minute files, gave good results. Indeed, we showed that it allows a reliable extraction of events containing cries while reducing the number of extracted audio segments. To go further, we also proposed to use the newborns' motion information computed by another team of our laboratory during the project [2, 3]. By focusing only on the sounds appearing in the periods detected as motion, we showed that it is possible to reduce considerably the amount of data to be

processed while keeping the vocalizations. Babies also produce a lot of sounds in the presence of adults. However, we have chosen to ignore these periods because of the data quantity and complexity (superimposition of voices and cries, lots of care-related noise, ...). The evaluation of this strategy on 303 hours of audio recordings performed on 22 newborns showed that they are very little in motion (12% of the time). We showed that collecting the sounds within these motion periods helped to remove up to 87% of the segments initially extracted.

Then, the classification method, after the segmentation step, is necessary to identify the cries among the extracted sound segments. We chose to use a time-frequency representation of the cries (spectrograms) as input to a Resnet convolutional neural network algorithm. The classification is thus performed in four steps: i) calculation of the spectrogram by Fast Fourier Transform (FFT) using successive Hamming windows of 0.04 ms and an overlap of 95 %, ii) slicing the spectrograms into images of the same duration with an overlap of 50 %, iii) using the convolutional neural network for the prediction of cries in the images, and iv) reconstruction of the sound predictions by retaining the majority prediction on the whole set of images. Thanks to transfer learning, the initial weights of the ResNet model were pre-trained with ImageNet and then optimized to our task (i.e., the crying vs. non-crying classification) by performing new learning. To adapt the model to our data, the parameters of input image duration, neural network complexity, and learning rate were optimized. In a two-step strategy, we first set the learning rate, then the evaluation of several combinations using cross-validation allowed us to identify the model with the best precision. This model corresponds to input images of 0.25 s duration, a ResNet34 architecture, and an initial learning rate of 10^{-4} . After being trained again on 30 babies (17 042 sounds), the classification performance obtained on three new babies (2 765 sounds) showed that 85.9% of the initially annotated cries were detected (sensitivity) and that 94.6% of the sounds classified as cries were indeed cries (precision).

The particularly noisy hospital sound environment (beep, machine, voice, etc.) complicates the task of automating crying detection. Even if by proposing a two-step strategy through the sound segmentation and classification for cry detection, we succeeded in extracting segments containing crying, there are some interesting issues to explore.

First, it is worth remembering that infants can be recorded in an open bed or in an incubator. In the latter case, very fragile premature newborns may have respiratory difficulties and will not be able to produce the same cries as older infants. It could therefore be possible to train a deep-learning model for both bed configurations. It would also be necessary to optimize more parameters of the neural network such as the optimization algorithm, the cost function, or the regularization by degradation of the weights which, in our case, have been fixed a priori.

Secondly, we are interested in sounds that contain several superimposed sound sources. Indeed, in an environment as noisy as the neonatal intensive care unit, it is normal that several sounds are mixed. In this thesis, we focused only on the periods when the baby was moving to limit this type

of data to be processed. It is surely appropriate to consider methods of source separation that will allow broadening the periods studied and to study the vocalizations of babies in the presence of adults. Once the different sources are separated, it may be easier to detect and characterize the crying segments.

For the estimation of the fundamental frequency characterization, we proposed a new method for tracking the infant cry F_0 in the context of real-time monitoring in the NICU. While methods in the literature typically set the frequency band in which to perform the tracking F_0 Orlandi13, Manfredi09, Orlandi15a, orlandi2017testing, we proposed an initial step to automatically identify this band. Once computed, the fundamental frequency tracking is performed using contour detection in the spectrogram.

To validate the method, we compared our estimation results to those computed by the software BioVoice which we identified as the reference program for the analysis of newborn cries. In fact, the method developed by Manfredi et al. obtained good performances on synthetic melodic forms of newborn cries [4, 5]. A qualitative comparison of the fundamental frequency tracks performed on 806 cries showed correct estimations in 87% of cases with BioVoice and 97% of cases with our method.

Although the proposed method offers good results in a large part of the cases, it could nevertheless be improved. The relevance of this method lies mainly in the automatic detection of the frequency band. This step is very little implemented in the literature and yet allows to improve considerably the estimation performances. Indeed, by concentrating the analysis band on a reduced frequency range, it avoids jumping to high energy frequency components. The currently proposed method is based on empirical parameters optimized experimentally which should be refined in order to allow the estimation of hyperphonations whose F_0 can be well beyond 1 000 Hz.

In addition, many studies sought to identify crying either to know the cause [6–10] or to identify pathologies [11–19]. If these analyses necessarily involve characterization of the fundamental frequency, it is especially based on the characterization of the cry melodies produced by the infants [5, 20–22]. Melody detection and evaluation could not be studied in this thesis but is a necessary step for anyone wishing to continue on the subject.

Finally, the automatic processing chain was deployed on a database of 57 babies born prematurely and at term for a total of 232 days of recording. Thanks to the successive treatments by the three proposed methods, we were able to automatically detect and characterize 117 947 cries. In a comparison with the literature, we showed that our results are consistent with two studies that inspect the fundamental frequency of crying in i) preterm infants according to their gestational age and birth weight [23] and ii) preterm and term infants at an equivalent post-menstrual age [24]. Through the analysis of longitudinal recordings from infants during their hospitalization, we presented changes in the duration and fundamental frequency of crying as a function of post-menstrual and postnatal ages. Finally, for the first time, the evolution of the fundamental

frequency for a population of preterm infants of normal evolution has been described and traced. These results are a major advance for the evaluation of the maturation of preterm infants during their hospitalization.

In conclusion, if this thesis brings the tools for the evaluation of the maturation and the tendencies of the cry parameters evolution according to the age in neonatal intensive care units, there is still much to be done. Some of these technical improvements have already been listed in this conclusion, but more clinical perspectives are also to be drawn. They are naturally part of the dynamics already studied and will aim to process as much data as possible in order to confirm and reinforce the trends observed and to cover the widest possible period of hospitalization with the aim of assessing possible deviations linked to infections or pathologies. This work will then be the basis for future developments in order to develop a fully automatic solution for a new generation of non-invasive monitoring systems for premature newborns through audio analysis.

BIBLIOGRAPHY

- [1] 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*, vol. 8, 799–810 (2013).
- [2] 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*, vol. 38, 207–213 (2017).
- [3] WEBER R., SIMON A., PORÉE F., AND CARRAULT G. Deep transfer learning for video-based detection of newborn presence in incubator. In *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2020, Montreal, QC, Canada, July 20-24, 2020*, 2147–2150. IEEE (2020).
- [4] ORLANDI S., BANDINI A., FIASCHI F., AND MANFREDI C. Testing software tools for newborn cry analysis using synthetic signals. *Biomedical Signal Processing and Control*, vol. 37, 16–22 (2017).
- [5] 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*, vol. 45, 174–181 (2018).
- [6] BHAGATPATIL M.V. AND SARDAR V. An automatic infant's cry detection using linear frequency cepstrum coefficients (lfcc). *International Journal of Scientific & Engineering Research*, vol. 5, 1379–1383 (2014).
- [7] CHANG C.Y. AND LI J.J. Application of deep learning for recognizing infant cries. In *Consumer Electronics-Taiwan (ICCE-TW), 2016 IEEE International Conference on*, 1–2. IEEE (2016).
- [8] FRANTI E., ISPAS I., AND DASCALU M. Testing the universal baby language hypothesis - automatic infant speech recognition with CNNs. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, 1–4 (2018).
- [9] MAGHFIRA T.N., BASARUDDIN T., AND KRISNADHI A. Infant cry classification using CNN - RNN. *Journal of Physics: Conference Series*, vol. 1528, page 012019 (2020).
- [10] LE L., KABIR A.N.M., JI C., BASODI S., AND PAN Y. Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, 106–110 (2019).
- [11] 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*, 342–349. Springer (2003).
- [12] 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*, 358–365. Springer (2004).
- [13] 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*, vol. 17, 38–46 (2015).
- [14] ONU C.C., LEBENSOLD J., HAMILTON W.L., AND PRECUP D. Neural transfer learning for cry-based diagnosis of perinatal asphyxia. *arXiv preprint arXiv:1906.10199* (2019).

- [15] 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*, 408–415. Springer (2004).
- [16] REYES-GALAVIZ O.F. AND REYES-GARCÍA C.A. Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system. In *Mexican International Conference on Artificial Intelligence*, 949–958. Springer (2005).
- [17] ZABIDI A., MANSOR W., KHUAN L.Y., SAHAK R., AND RAHMAN F. Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In *2009 5th International Colloquium on Signal Processing & Its Applications*, 204–208. IEEE (2009).
- [18] ZABIDI A., KHUAN L.Y., MANSOR W., YASSIN I.M., AND SAHAK R. Detection of infant hypothyroidism with mel frequency cepstrum analysis and multi-layer perceptron classification. In *2010 6th International Colloquium on Signal Processing its Applications*, 1–5 (2010).
- [19] 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*, vol. 46, 965–975 (2008).
- [20] WERMKE K., MENDE W., MANFREDI C., AND BRUSCAGLIONI P. Developmental aspects of infant's cry melody and formants. *Medical Engineering & Physics*, vol. 24, 501–14 (2002).
- [21] VÁRALLYAY G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, vol. 71, 1699–1708 (2007).
- [22] VÁRALLYAY JR G. Z.B. AND ILLÉNY. A. Automatic infant cry detection. page 11–14 (2009).
- [23] MANFREDI C., BOCCHI L., ORLANDI S., SPACCATERRA L., AND DONZELLI G.P. High-resolution cry analysis in preterm newborn infants. *Medical Engineering & Physics*, vol. 31, 528–32 (2009).
- [24] 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*, vol. 10 (2014).

List of publications

International Journals

- [1] CABON S., **MET - - MONTOT B.**, PORÉE F., ROSEC O., SIMON A. and CARRAULT G. Extraction of Premature Newborns' Spontaneous Cries in the Real Context of Neonatal Intensive Care Units. *Sensors*, vol. 22(5), page 1823 (2022).
- [2] 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*, vol. 52, 362-370 (2019).

International Conferences - paper

- [1] **MET - - MONTOT B.**, CABON S., CARRAULT G., PORÉE F. Spectrogram-based fundamental frequency tracking of spontaneous cries in preterm newborns. *EUSIPCO 2020*, Amsterdam, Netherlands, 1185-89 (2020).
- [2] CABON S., **MET - - MONTOT B.**, PORÉE F., ROSEC O., SIMON A., CARRAULT G. Automatic extraction of spontaneous cries of preterm newborns in Neonatal Intensive Care Units. *EUSIPCO 2020*, Amsterdam, Netherlands, 1200-04 (2020).

International Conference - abstract

- [1] **MET - - MONTOT B.**, CABON S., PLADYS P., CARRAULT G. and PORÉE F. Automatic fundamental frequency characterization of premature newborns' cries in Neonatal Intensive Care Unit. *ICVPB'20*, Grenoble, France, résumé (2020).

National Conference

- [1] **MET - - MONTOT B.**, CABON S., CARRAULT G., PORÉE F. Extraction de pleurs de nouveau-nés par segmentation audio-vidéo et Deep Learning. *GRETSI'22*, Nancy, France (2022).

Software

- [1] **MET - - MONTOT B.**, CABON S., PORÉE F., CARRAULT G. SoundAnnoT: Sound Annotation Tool. Dépôt APP no: IDDN.FR.001. 020001.000.S.P.2021.000.31230. (2021).

Titre : Détection et caractérisation des vocalisations chez des nouveau-nés prématurés.

Mot clés : nouveau-nés prématurés, développement neuro-comportemental, surveillance, unités de soins néonatales, analyse audio, pleurs spontanés

Résumé : Le nombre de naissances prématurées est estimé à 15 millions par an dans le monde et représente 7% des naissances en France. Ces bébés sont pris en charge en Unités de Soins Intensifs Néonatales (USIN) et font l'objet d'une surveillance particulière du fait de l'immaturation de leurs organes et des complications qui peuvent en découler.

De nombreuses études ont montré que l'analyse des pleurs de nourrissons permettait d'obtenir des informations sur leur état de santé et dans le cas des prématurés sur leur maturation. Si les premiers travaux se basaient sur une segmentation manuelle de pleurs souvent induits (généralement par la douleur), les travaux actuels s'intéressent aux pleurs spontanés, ce qui nécessite le développement de méthodes d'extraction automatiques. Cette approche non-invasive de monitoring apparaît comme extrêmement pertinente au vu de la fragilité des sujets étudiés. Cependant, l'environnement hospitalier particulièrement bruyant où se déroulent les enregistrements complexifie grandement l'automatisation des méthodes.

Dans ce contexte, et dans le cadre du projet européen Digi-NewB, l'objectif de ces travaux est de présenter une chaîne complète de traitements automatiques pour l'analyse des pleurs des prématurés enregistrés en USIN. Cette chaîne regroupe : i) une nouvelle approche de détection des pleurs composée d'une segmentation, réalisée à partir de la fusion de vidéos et de bandes son ; ii) une classification par deep-learning pour l'identification des pleurs parmi tous les sons segmentés (voix d'adultes, alarmes...); iii) l'estimation de la fréquence fondamentale des pleurs détectés par une nouvelle approche basée sur la détection de contours dans le spectrogramme.

Le déploiement de la chaîne de traitements sur une base de données de pleurs enregistrés en USIN montre des résultats en accord avec ceux publiés dans la littérature. Cette validation est encourageante et annonce la possibilité d'observer automatiquement sur des grandes cohortes l'évolution des pleurs des prématurés, notamment en vue de caractériser leur développement.

Title: Detection and characterization of vocalizations in preterm newborns.

Keywords: preterm newborn, neuro-behavioral development, monitoring, neonatal intensive care units, audio analysis, spontaneous cries

Abstract: The number of premature births is estimated at 15 million per year worldwide and represents 7% of births in France. These babies are cared for in Neonatal Intensive Care Units (NICU) and are subject to special surveillance because of the immaturity of their organs and the complications that may arise.

Numerous studies have shown that the analysis of infant crying provides information on their health status and, in the case of premature infants, on their maturation. While early work was based on manual segmentation of often induced crying (usually by pain), current work focuses on spontaneous crying, which requires the development of automatic extraction methods. This non-invasive monitoring approach appears to be extremely relevant given the fragility of the subjects studied. However, the particularly noisy hospital environment where the recordings are made makes the automation of the methods very complex.

In this context, and within the framework of the

European project Digi-NewB, the objective of this work is to present a complete chain of automatic treatments for the analysis of the cries of premature babies recorded in NICU. This chain gathers: i) a new approach of crying detection composed of a segmentation, realized from the fusion of videos and soundtracks; ii) a classification by deep-learning for the identification of crying among all the segmented sounds (adult voices, alarms...); iii) the estimation of the fundamental frequency of the detected crying by a new approach based on the detection of contours in the spectrogram.

The deployment of the processing chain on a database of cries recorded in NICU shows results in agreement with those published in the literature. This validation is encouraging and announces the possibility of automatically observing the evolution of crying in premature babies on large cohorts, in particular in order to characterize their development.