Extraction de pleurs de nouveau-nés par segmentation audio-vidéo et Deep Learning

Bertille MET-MONTOT, Sandie CABON, Guy CARRAULT, Fabienne PORÉE

Univ Rennes, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
bertille.met-montot@univ-rennes1.fr, sandie.cabon@univ-rennes1.fr,
guy.carrault@univ-rennes1.fr, fabienne.poree@univ-rennes1.fr

Résumé – L'objectif de cette étude est de proposer une nouvelle méthode d'extraction automatique des pleurs de nouveau-nés prématurés en contexte clinique réel. La première étape, basée sur l'analyse conjointe de l'audio et de la vidéo, permet la segmentation des sons. Puis une étape de classification exploitant des réseaux de neurones convolutifs, alimentés par les spectrogrammes des segments sonores, permet d'identifier les pleurs. Une étude menée sur 19 807 sons annotés manuellement montre des résultats encourageants avec une précision de 93%.

Abstract – The objective of this study is to propose a new method for the automatic extraction of cries from premature newborns in real clinical context. The first step, based on the joint analysis of audio and video, allows the sound segmentation. Then a classification step using convolutional neural networks, with spectrograms of the sound segments as inputs, allows to identify the cries. A study conducted on 19 807 manually annotated sounds shows encouraging results with a 93% precision.

1 Introduction

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'immaturité de leurs organes et des complications qui peuvent en découler.

De nombreuses études ont montré que l'analyse des pleurs de prématurés permettait d'obtenir des informations sur leur état de santé et sur leur maturation [1]. Celles-ci reposent généralement sur une étape de segmentation des pleurs, suivie d'une analyse de leur contenu fréquentiel. Si les premiers travaux se basaient sur une segmentation manuelle de pleurs souvent induits (généralement par la douleur), les travaux actuels s'intéressent désormais aux pleurs spontanés. Ceci nécessite le développement de méthodes d'extraction automatique, qui permettent par ailleurs le traitement de volumes de données plus importants [2]. Malheureusement, les approches classiques, basées le plus souvent sur le calcul de l'énergie du signal sonore, rencontrent des limites lorsqu'elles sont appliquées en USIN. En effet, l'environnement particulièrement bruité ne permet pas d'extraire que les pleurs mais également les autres sons enregistrés (voix d'adultes, alarmes...). Partant de ce constat, des méthodes plus récentes ont abordé l'étape de d'extraction des pleurs comme un problème de classification en utilisant des méthodes de Machine Learning, éventuellement profond [3]. Celles-ci ont l'inconvénient de devoir traiter l'intégralité des bandes sonores (incluant les silences ainsi que les nombreuses sources sonores) et multiplie la complexité de l'analyse (temps de calcul et taux de classification).

Nous proposons ici une nouvelle approche qui favorise l'extraction, l'analyse et la classification des pleurs à partir de la fusion de vidéos et de bandes son. Cette proposition fait suite au projet européen Digi-NewB -qui a eu pour objectif de proposer un système non invasif de décision, pour le diagnostic de l'infection et le suivi de la maturation pour les prématurés hospitalisés-, en s'appuyant entre autres sur l'enregistrement simultané de l'audio et de la vidéo [4]. Jusqu'à présent les analyses des vidéos et des bandes son ont été réalisées indépendamment l'une de l'autre. Le traitement des vidéos a permis d'étudier le mouvement, en se basant en particulier sur la caractérisation des intervalles de mouvement et de non-mouvement [5]. Le traitement de l'audio a eu pour objectif d'extraire et d'analyser les pleurs, mais celui-ci n'a pas encore été déployé sur des grandes quantités de données [6].

Plus précisément, cette nouvelle approche d'extraction des pleurs associe :

- Une étape de segmentation audio-vidéo, qui combine une détection des sons dans l'audio et une détection des intervalles de mouvement dans la vidéo ;

- Une classification des segments en deux classes (pleurs vs non-pleurs), basée sur l'utilisation d'un réseau de neurones convolutif (CNN), appliqué à des images de spectrogrammes.

L'article est organisé comme suit. La section 2 décrit le protocole et les méthodes développées. Dans la section 3, les résultats obtenus sont présentés et comparés à une approche faisant référence dans la littérature [2]. Enfin, la discussion et la conclusion sont données dans la section 4.

2 Méthodes

2.1 Protocole d'acquisition

Les données ont été acquises dans le cadre du projet européen Digi-NewB, dans six hôpitaux du Grand Ouest [4]. Les vidéos ont été enregistrées à partir de caméras noir et blanc infra-rouge, à la fréquence de 25 Hz. Les bandes son ont été obtenues grâce à des microphones omni-directionnels, à un taux d'échantillonnage de 24 kHz. La durée des enregistrements était variable, de 1 à 10 jours, et les données sont stockées dans des fichiers de 30 minutes.

Cette étude porte sur 33 bébés nés prématurément et à terme, d'âge gestationnel compris entre 25 et 41 semaines d'aménorrhée (SA), et d'âge post-mentruel compris entre 27 et 41 SA, ce qui correspond à un total de 158 jours d'enregistrement.

2.2 Segmentation audio-vidéo

2.2.1 Segmentation des sons

La méthode de segmentation des sons est basée sur la méthode proposée par Orlandi et al., initialement développée pour la détection des pleurs de nouveau-nés dans un environnement non bruité [2]. Celle-ci s'appuie sur le calcul de l'énergie à court-terme (« short-term energy », STE) du signal sonore filtré par défaut entre 50 et 1000 Hz, définie par :

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

où n est le nombre d'échantillons dans la fenêtre d'analyse correspondant à 20 ms, $s(i)$ est le signal discret et ε est une constante pour éviter le log(0).

La détection des segments sonores s'effectue en confrontant STE à deux seuils (T_U et T_L) qui sont estimés par la méthode d'Otsu [2]. Le seuil haut T_U est calculé sur l'ensemble des valeurs du STE et permet de détecter les sons. Le seuil T_L, calculé sur les valeurs du STE inférieures à T_U, permet de segmenter le son au niveau des points où la courbe STE descend sous T_L. L'utilisation d'un double seuil permet d'éviter la division incorrecte d'un seul son en plusieurs segments (Figure 1).

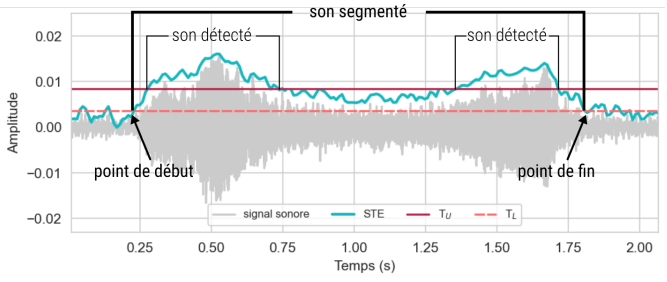

FIG. 1 – Exemple de segmentation d'un pleur.

Cette approche n'ayant pas été initialement conçue pour la segmentation de pleurs dans un environnement bruité de type USIN, elle a été adaptée avec les améliorations suivantes :

- Un premier seuil T, calculé sur des fenêtres de 2 heures, pour s'assurer que la période d'analyse contient bien des sons ; les fichiers où se produisent moins de 10 événements au-dessus de ce seuil sont écartés ;
- Un filtrage passe-bande du signal sonore entre 200 et 1000 Hz, correspondant aux fréquences des pleurs de bébés, appliqué avant l'étape de seuillage ;
- Une étape dite de "re-segmentation" consiste à segmenter une seconde fois les segments de durée supérieure à 5 s.
- Enfin, seuls les sons de durée comprise entre 0.25 et 5 s sont conservés, la limite inférieure étant la durée minimale nécessaire pour faire varier les cordes vocales [8].

2.2.2 Segmentation du mouvement

La segmentation du mouvement s'appuie sur l'analyse des vidéos et repose sur les étapes suivantes (Figure 2) :

- Le calcul du mouvement par une différence inter-images ;
- L'exclusion automatique des périodes où le bébé n'est pas présent dans son lit, ainsi que celles incluant la présence d'adultes (parents ou soignants) dans le champ. Cette étape fondamentale est effectuée grâce à une approche de Deep Learning (voir [7] pour plus de détails) ;
- La segmentation des intervalles de mouvement et de non-mouvement. L'approche est basée sur une classification par Random Forest [5].

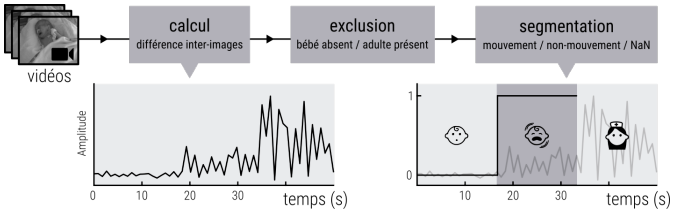

FIG. 2 – Étapes de la segmentation du mouvement.

Finalement, un signal synthétique est construit. Il est égal à 1 pendant les phases de mouvement, à 0 pendant les phases de non-mouvement et « NaN » en cas d'absence du bébé ou de présence d'adultes.

2.2.3 Fusion des deux segmentations

L'extraction automatique des pleurs est une étape de prétraitement nécessaire pour réaliser des analyses fréquentielles sur de grandes bases de données. Donc, à l'issue de la classification, même si nous souhaitons détecter un maximum de pleurs (i.e. une sensibilité élevée), notre priorité est qu'un son extrait soit bien un pleur (i.e. une précision élevée). La segmentation du son permet d'extraire tous les évènements sonores présents dans les enregistrements (pleurs, voix adultes, bip des moniteurs...). La segmentation du mouvement permet d'identifier les intervalles de mouvement, de non-mouvement et ceux avec la présence d'adultes (qui sont souvent bruyants à causes des soins, des voix...). La fusion des deux segmentations a pour

objectif de réduire le nombre de segments à classifier. Au vu de la quantité de données à traiter et sachant qu'un pleur n'apparaît jamais dans un intervalle de non-mouvement, nous avons choisi de ne conserver que les segments apparaissant dans des intervalles de mouvement.

2.3 Classification par CNN

2.3.1 Étapes de la méthode

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. La classification est ainsi réalisée en 4 étapes (Figure 3) :

- Pour chaque son extrait, le spectrogramme est calculé par une transformée de Fourier à court terme à l'aide de fenêtres de Hamming successives de 0.04 ms et d'un recouvrement de 95%. Cette configuration procure une résolution fréquentielle de 23.4 Hz et temporelle de 4.2 ms.
- Les segments étant de longueurs variables, les images de spectrogrammes sont découpées en images de même durée (0.20 ou 0.25 s) avec un recouvrement de 50%. La première durée est la plus communément utilisée dans la littérature, tandis que la seconde correspond à la durée minimale des segments sonores.
- Ces images découpées sont ensuite placées en entrée d'un réseau de neurones convolutif.
- Enfin, pour chaque segment sonore initial, la décision retenue est la prédiction majoritaire sur l'ensemble des images.

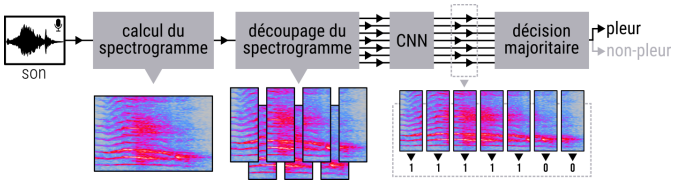

FIG. 3 – Étapes de la classification des sons (ex : pleur).

2.3.2 Stratégie d'entraînement du modèle

L'apprentissage est réalisé par transfert. Pour mémoire, ce principe consiste à réutiliser des réseaux de neurones convolutifs préalablement entraînés sur une large base de données d'images. Les poids de modèles ResNet pré-entraînés avec ImageNet initialisent le modèle de classification, qui sont ensuite optimisés à notre tâche, i.e., la classification pleurs vs non-pleurs, en réalisant un nouvel apprentissage.

Pour adapter le modèle à nos données, certains paramètres (Tableau 1) ont été fixés tandis que les paramètres suivants ont été optimisés :

- la durée des images d'entrées de 0.20 ou 0.25s ;
- la complexité du réseau de neurones : ResNet 18 ou 34 ;
- le taux d'apprentissage (learning rate) : de 10^{-2} à 10^{-5}.

TAB. 1 – Paramètres fixes du modèle.

Paramètre	Choix
Fonction de coût	Entropie croisée
Algorithme d'optimisation	Descente du gradient stochastique associée à un momentum : 0.9
Régularisation par dégradation des pondérations (weight decay)	5.10^{-5}
Régularisation par apprentissage par lots (batch)	16
Nombre d'itérations d'apprentissage (epoch)	10
Gestion du déséquilibre des classes (class weighting, déterminé à partir de la répartition des images)	non-pleurs 0.66, pleurs 0.33

3 Résultats

3.1 Base de données annotées

Pour l'évaluation de la méthode de segmentation, trois fichiers de 30 minutes ont été manuellement annotés en identifiant les points de début et de fin de tous les sons audibles et leur type (pleur ou non-pleur).

Pour l'évaluation de la méthode de classification, 19 807 segments ont été manuellement annotés, conduisant à 5 476 pleurs et 14 331 non-pleurs. Le découpage des spectrogrammes en images de 0.20 s et 0.25 s a fourni 175 899 et 126 583 images respectivement.

3.2 Résultats de segmentation

Fichier 1 : 0 pleur, 23 non-pleurs. Le traitement par la méthode [2] génère 30 segments tandis qu'avec notre approche, aucun son n'est segmenté car le fichier est directement écarté (grâce au seuil T).

Fichier 2 : 155 pleurs, 254 non-pleurs. Le traitement par la méthode [2] génère 75 segments, dont 47 (soit 63%) coïncident avec des pleurs annotés. Avec notre approche, 160 segments sont détectés, dont 112 (soit 70%) coïncident avec des pleurs annotés. De plus, sur les 155 pleurs annotés, la méthode [2] en détecte 36 (soit 23%) et notre méthode en détecte 112 (soit 69%).

Fichier 2 : 776 pleurs, 385 non-pleurs. Le traitement par la méthode [2] génère 656 segments, dont 638 (soit 97%) coïncident avec des pleurs annotés. Avec notre approche, 497 segments sont détectés, dont 466 (soit 94%) coïncident avec des pleurs annotés. De plus, sur les 776 pleurs annotés, la méthode [2] en détecte 611 (soit 79%) et notre méthode en détecte 571 (soit 74%).

L'analyse de ces exemples montre i) l'impact de l'environnement sonore sur les résultats, ii) l'intérêt de notre approche pour le traitement de fichiers hétérogènes et iii) la nécessité d'utiliser un classifieur en sortie de segmentation.

3.3 Résultats de classification

Quatre combinaisons ont été testées et sont rappelées dans le Tableau 2. L'optimisation des paramètres a été faite en deux temps. Premièrement, après entraînement, les taux d'apprentissage donnant la précision la plus élevée sur un ensemble de test ont été identifiés pour chaque combinaison. L'apprentissage a été réalisé sur une base de données constituée de 29 bébés et ont été testés sur un bébé. Puis, en utilisant les meilleurs taux d'apprentissage, la meilleure combinaison, i.e, la meilleure modélisation, a été identifiée par une validation croisée à 5 sets, réalisée sur l'ensemble des 30 bébés. La combinaison avec la meilleure précision moyenne a été retenue car nous souhaitons maximiser le nombre de vrais positifs en sortie du classifieur. Les combinaisons ainsi que les taux d'apprentissage finalement choisis sont présentés (en gras) dans le Tableau 2. Les performances moyennées des cinq sets sont présentées dans le Tableau 3.

TAB. 2 – Combinaisons testées avec les taux d'apprentissage choisis par la recherche par grille en gras.

Comb.	Durée	Réseau	Taux d'apprentissage
1	0.20 s	ResNet18	$10^{-2}, \mathbf{10^{-3}}, 10^{-4}, 10^{-5}$
2	0.20 s	ResNet34	$10^{-2}, 10^{-3}, \mathbf{10^{-4}}, 10^{-5}$
3	0.25 s	ResNet18	$10^{-2}, 10^{-3}, 10^{-4}, \mathbf{10^{-5}}$
4	0.25 s	ResNet34	$10^{-2}, 10^{-3}, \mathbf{10^{-4}}, 10^{-5}$

TAB. 3 – Moyenne et écart-type des cinq sets de la validation croisée pour la base de données de sons.

Comb.	Précision	Sensibilité	F_1-score
1	0.84±0.07	0.81±0.08	0.82±0.05
2	0.85±0.07	0.80±0.08	0.82±0.05
3	0.84±0.08	0.78±0.09	0.80±0.07
4	**0.86**±0.07	**0.81**±0.11	**0.83**±0.06

Le modèle avec la meilleure précision (durée : 0.25 s, réseau : ResNet 34 et taux d'apprentissage : 10^{-4}) a été sélectionné et à nouveau entraîné sur les 30 bébés pour permettre son déploiement.

Pour vérifier la généralisation du modèle retenu, les performances de classification obtenues sur un ensemble de test de trois nouveaux bébés ont été calculées. Elles montrent que 86% des pleurs initialement annotés ont été détectés (sensibilité) et que 93% des sons classés comme pleurs sont effectivement des pleurs (précision).

4 Discussion et Conclusion

Une nouvelle méthode pour l'extraction des pleurs dans un environnement clinique est proposée dans cette communication. Elle combine une étape de i) segmentation exploitant l'audio et la video et de ii) classification par Deep Learning. Pour la segmentation c'est, à notre connaissance, la première étude qui exploite le mouvement pour l'extraction des pleurs. Cette stratégie permet de réduire en moyenne de 60% le nombre de segments détectés (comparé à la segmentation des sons basée sur le signal audio seul) et ainsi écarter de nombreux segments, correspondant essentiellement à des bruits parasites (voix d'adultes, bips des moniteurs...). L'étape de classification qui suit permet de séparer les pleurs des non-pleurs. Appliquée à un ensemble de 19 807 sons, elle fournit une précision de 93%.

L'approche proposée, totalement automatique, dans un environnement très bruité et comparée aux travaux reportés dans la littérature, ouvre de nouvelles perspectives de recherche pour le suivi de maturation des bébés prématurés en USIN ou encore pour l'étude des interactions sociales des prématurés avec son entourage au cours de l'hospitalisation.

Remerciements

Les résultats présentés dans cette publication ont été financés par le programme de Recherche et d'Innovation de l'Union Européenne sous l'accord de financement n^o 689260 (projet Digi-NewB).

Références

[1] S. Cabon, F. Porée, A. Simon, O. Rosec, P. Pladys, and G. Carrault, "Video and audio processing in paediatrics: a review," *Physiological Measurement*, vol. 40, no. 2, pp. 02TR02, 2019.

[2] S. Orlandi, A. Guzzetta, A. Bandini, et al., "AVIM–A contactless system for infant data acquisition and analysis: Software architecture and first results," *Biomedical Signal Processing and Control*, vol. 20, pp. 85–99, 2015.

[3] D. Ferretti, M. Severini, E. Principi, A. Cenci, and S. Squartini, "Infant cry detection in adverse acoustic environments by using deep neural networks," in *EUSIPCO 2018*. IEEE, pp. 992–996, 2018.

[4] "Digi-NewB - GCS HUGO - CHU - monitoring system," `digi-newb.eu`, accessed 14 April 2020.

[5] S. Cabon, "Monitoring of premature newborns by video and audio analyses," Thèse de l'Univ. de Rennes 1, 2019.

[6] B. Met-Montot, S. Cabon, G. Carrault, and F. Porée, "Spectrogram-based fundamental frequency tracking of spontaneous cries in preterm newborns," in *EUSIPCO 2020*. IEEE, pp. 1185–1189, 2021.

[7] R. Weber, S. Cabon, A. Simon, F. Porée, and G. Carrault, "Preterm newborn presence detection in incubator and open bed using deep transfer learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1419–1428, 2021.

[8] B. Lester and L. LaGasse, "Crying," in Social and Emotional Development in Infancy and Early Childhood. Elsevier, pp. 80–90, 2009.