ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

SEGMENTATION AUTOMATIQUE DES CRIS DES NOUVEAU-NÉS EN VUE DU DÉPISTAGE PRÉCOCE DES PROBLÈMES NEUROPHYSIOLOGIQUES

PAR
Lina ABOU-ABBAS

THÈSE PAR ARTICLES PRÉSENTÉE À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DU
DOCTORAT EN GÉNIE
Ph. D.

MONTRÉAL, LE 21 OCTOBRE 2016

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

M. Chakib Tadj, directeur de thèse
Département de génie électrique à l'École de technologie supérieure

M. Jérémie Voix, président du jury
Département de génie mécanique à l'École de technologie supérieure

M. Patrick Cardinal, membre du jury
Département de génie électrique à l'École de technologie supérieure

M. Mohand Said Allili, jury externe indépendant
Département d'informatique et d'ingénierie à l'université du Québec à l'Outaouais

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 27 SEPTEMBRE 2016

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

First and foremost I would like to thank Allah for giving me the strength to carry on this work and the perseverance to see it to completion.

To the memory of my Mom, Mariam Eltawil (1957-2002), who was and still is my constant source of inspiration and the determination of my academic abilities: “I miss you more than words can say. You are not here physically to help celebrate this achievement, but I know you are here in spirit. I hope I have made you proud.”

I express my heart-felt gratitude to my wonderful professor Chakib Tadj, my mentor, for his support, guidance, patience, and encouragement throughout my doctoral studies. His door was always widely opened to me. He has been eager to spend his precious time to encourage me and to provide comments that acted as motivation in my work. Many thanks for your unyielding moral and financial support and for your professionalism during my academic effort.

Special thanks go to the committee members who took time to review my work and offer valuable suggestions.

To my dearly loved husband, Hussein Eltawil, my rock, and soul mate: “You gave me the strength to move forward effortlessly. Without you, none of my success would be possible. Your confidence was an inspiration for me to gain focus and complete what I started. Your love will always be engraved in my heart and deeply embedded in my soul forever. I am truly thankful for having you in my life.”

To my precious treasure, my children Mariam and Hadi:”You are truly an amazing gift from Allah to me. Thank you for being the sun of my life. Thank you for bringing a sense of reality to my life that seemed consumed with the research journey. You inspired me in different ways to be more creative and thoughtful. You made me a better mother. I love you so much.”

I offer my humble salutation to my wonderful family: my father Kamel Abou-Abbas, my sisters, Linda, Racha, Amal, my loving brother Mohamad, my mother-in-law Nada, my brothers-in-law Hazar Mohanna and Ali Kachmar:« It would have been difficult for me to surpass this research studies without your love and your prayers. Thank you for loving me unconditionally, for always believing in me, for your kind support and for feeling so close despite being so far ».

I am very grateful to everyone who helped me to embark this journey of my doctoral studies. I especially thank my friends and colleagues in the MMS laboratory, for supporting me all the way, for their suggestions, advice, assistance, friendly working atmosphere and much needed support. I want to express my deepest thanks to Hesam, Prabin and Leila for sharing with me ideas and opinions and for the detailed review of my English writing.

Many thanks to all of you for being part of this journey.

SEGMENTATION AUTOMATIQUE DES CRIS DES NOUVEAU-NÉS EN VUE DU DÉPISTAGE PRÉCOCE DES PROBLÈMES NEUROPHYSIOLOGIQUES

Lina ABOU-ABBAS

RÉSUMÉ

Plusieurs études ont établi l'existence d'un grand nombre d'informations présentes dans un signal du cri des nouveau-nés. En se basant sur cette hypothèse, de nombreuses recherches se sont consacrées à l'analyse de ce signal dans le but de classifier d'une part, le type du cri (cri de naissance, douleur, faim, inconfort, etc.) et d'autre part, l'état pathologique du nouveau-né. Cette thèse décrit le développement et la validation d'un outil de segmentation automatisé pour la détection des phases vocalisées d'expiration et d'inspiration des cris des nouveau-nés collectés dans un environnement médical bruité. Cet outil fera partie de la phase du prétraitement des signaux audio des cris des nouveau-nés, et ce en amont du système automatique de classification des pathologies chez les nouveau-nés. Dans un premier axe, nous avons contribué à la mise en place d'une base de données de cris des nouveau-nés sains et pathologiques destinée à être publique, accessible en permanence aux fins de recherches multiples relatives spécifiquement à la santé des nourrissons. La réalisation de la base de données a satisfait les attentes, entre autres : l'archivage sécurisé des données, la consultation facile et accélérée des informations à l'aide d'une interface adéquate et efficace et le téléchargement rapide vers d'autres emplacements pour différentes utilisations. Un corpus de 2073 signaux de cris a été recueilli. 769 bébés ont participé à cette collecte dont 372 sont atteints de diverses maladies telles que maladies respiratoires, cardiaques et neurologiques. Dans un deuxième axe, nous avons proposé des méthodes d'apprentissage supervisées GMM et HMM pour la conception de l'outil de segmentation automatique des signaux de cris. Compte tenu de la grande variabilité rencontrée dans une base de données réelle des signaux de cris, cet outil est capable de détecter les parties importantes de cris parmi d'autres activités acoustiques présentes dans le corpus comme les sons provenant des machines médicales, la parole, les bruits à des niveaux variés et enfin le silence. Plusieurs outils de traitement et de reconnaissance des signaux ont été exploités dans ce travail afin de proposer un module de segmentation des signaux de cris complètement automatique robuste vis-à-vis du bruit et applicable dans un environnement clinique réel et qui ne nécessite pas un réglage manuel des seuils. Afin d'exploiter l'information contenue dans les signaux de cris de différentes manières, nous avons appliqué et comparé les méthodes de décomposition de signaux les plus utilisées: Transformée de Fourier rapide, transformée en ondelettes et enfin décomposition modale empirique. Nous avons procédé à l'extraction des divers descripteurs afin de caractériser et modéliser séparément et d'une manière efficace chacun des types d'expiration et d'inspiration avec vocalisation. Dans un troisième axe, et pour améliorer les résultats obtenus des approches supervisées en réduisant les erreurs de localisation des points de début et de fin des segments utiles, nous avons intégré une phase de post-traitement afin d'exploiter l'information temporelle du signal. L'architecture complète réalisée est basée sur deux modules successifs. Le premier vise à utiliser les approches statistiques supervisées et obtenir une première classification et le second consiste à se servir des paramètres temps-fréquences pour corriger les erreurs de la première classification et améliorer ainsi les résultats globaux. Les différentes approches proposées ont été testées sur une base de données différente de celle utilisée lors de l'apprentissage. La technique de la validation croisée stratifiée à dix tours a été employée afin d'évaluer et de vérifier l'efficacité des systèmes proposés. Les résultats des tests réalisés montrent le comportement robuste des algorithmes proposés.

Mots clés : Cris des nouveau-nés; Expiration; Inspiration; Modèles de Mélanges Gaussiennes; Modèles de Markov Cachés; Coefficients Cepstraux; Taux de Passage par Zéro; Fréquence Fondamental; Décomposition en Mode Empirique; Paquets d'ondelettes.

AUTOMATIC SEGMENTATION OF NEWBORNS' CRIES FOR THE EARLY SCREENING OF NEURO-PHYSIOLOGICAL HEALTH PROBLEMS

Lina ABOU-ABBAS

ABSTRACT

Several Studies have established the existence of a large number of information in an infant cry signal. Based on this assumption, many researches are devoted to the analysis of the cry signal in order to classify in one hand, the type of cry (birth cry, pain, hunger, discomfort, etc.) and in other hand the physical state of the newborn. This thesis describes the development and validation of an automated segmentation tool for the detection of vocal expiration and inspiration phases of newborn cries collected in a noisy hospital environment. This tool will be part of the preprocessing phase of newborn crying signals, prior to the automatic pathology classification system for newborns. As a first step, we have contributed to the establishment of a healthy and pathologic newborns' cries database, intended to be public, accessible at all times for multiple research purposes related to the health of infants. The implementation of the cry database fulfilled expectations, including: secure data archiving, easy, and fast retrieval of information by means of adequate and effective interface and fast downloading to different locations for different uses. A corpus of 1939 cry signals were collected. 769 babies participated of which 372 are suffering from various diseases such as respiratory diseases, cardiac diseases, and neurological diseases. In a second time, we used supervised learning methods, Gaussian Mixture Models and Hidden Markov Models, for the design of the automatic cry segmentation tool. Given the variability encountered in a real cry signals database, this tool is able to detect useful part of cries from other acoustic activities registered as the sounds of medical equipment, speech, noises at various levels and silence. Several signal processing and recognition tools have been investigated in this work in order to offer a fully automatic cries signals segmentation module robust towards noise and applicable in a real clinical environment and the most important, does not require a static threshold.

In order to exploit the information available in the cries signals in different ways, we have applied and compared the most used signal decomposition techniques namely Fast Fourier Transform, Wavelet Packet Transform, and Empirical Mode Decomposition. We extracted different features to characterize and model separately and efficiently each type of vocal expiration and inspiration. The third area of focus, and in order to improve the results obtained from the supervised approaches by reducing boundary detection errors of useful segments, we integrated a post-processing stage to take full advantage of the time information of the signal. The full architecture realized is based on two consecutive modules. The first module uses cepstral features and traditional statistical approaches to give first results' classification, and the second uses time and frequency features to correct errors and improve overall results. The various proposed approaches were tested on a different training and testing corpuses. The 10-fold cross validation technique is used to evaluate and verify the effectiveness of the proposed systems. The results of various tests show the robust performance of the proposed algorithms.

Keywords: Newborns cries signals; Expiration; Inspiration; Gaussian Mixture Models; Hidden Markov Models; Cepstral Coefficients; Zero Crossing rates; Fundamental Frequency; Empirical Mode Decomposition; Wavelet Packet.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ACC	Accuracy
AER	Accuracy Error Rate
ASR	Automatic Speech Recognition
BIP	Bip sound of medical instruments
BKG	Background
CER	Classification Error Rate
D	Number of Deletions Errors
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
EM	Expectation Maximisation
EMD	Empirical Mode Decomposition
EX	Expiration
EXP	Expiration
EXP2	Expiration during a Non-Cry Period
EXPN	Non Audible Expiration
F0	Fréquence Fondamentale
FFT	Fast Fourier Transform
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HPS	Harmonic Product Spectrum
HTK	Hidden Markov Model Toolkit
I	Number of Insertions Errors
IFFT	Inverse Fast Fourier Transform
IMF	Intrinsic Mode Functions
IN	Inspiration

INS	Silent Inspiration
INS2	Inspiration during a Non-Cry Period
INSV	Audible Inspiration or Inspiration with Vocalization
INT	Intensity
LFCC	Linear Frequency Cepstral Coefficients
MFCC	Mel -Frequency Cepstrum Coefficients
MFCC_E_D_Z	Mel-Frequency Cepstrum Coefficients with Energy Coefficients and Delta and with Zero Mean Static Coefficients
MFCC_E_D_A_Z	Mel-Frequency Cepstrum Coefficients with Energy Coefficients and Delta and Delta Delta Coefficients and with Zero Mean Static Coefficients
MFDWC	Mel-Frequency Discrete Wavelet Coefficients
N	Number of labels in the reference file
NOR	All None Cry labels
NCDS	Newborn Cry-Based Diagnostic System
NS	Noise
P	Pause
S	Number of Substitutions Errors
SBC	Subband based Cepstral Parameters
SI	Silence
SIFT	Simple Inverse Filter Tracking
SP	Speech
SVM	Support Vector Machine
STDLFCC	Segmental two-dimensional Linear Frequency Cepstral Coefficient
STE	Short Term Energy
TFCT	Transformée de Fourier à court terme
TP	True Positive
TPR	True Positive Rate
TN	True Negative
VAD	Voice Activity Detection

WE_DCT	Wavelet Energy based DCT
WPP	Wavelet Packet Parameters
WPT	Wavelet Packet Transform
ZCR	Zero Crossing Rate

INTRODUCTION

0.1 Mise en Contexte

Un dépistage précoce chez des nouveau-nés asymptotiques, dès les premiers jours de leur naissance, permettrait d'établir le diagnostic d'une pathologie grave. Ils pourront ainsi être pris en charge par des spécialistes, soit par un traitement spécifique, soit par des mesures susceptibles de mieux contrôler l'évolution de cette pathologie avant que les séquelles de celle-ci ne soient irréversibles (Cyril Goizet, 2004). Ceci souligne l'importance du dépistage précoce des différentes pathologies.

Le dépistage systématique, combiné aux meilleurs outils de diagnostic, peut s'avérer nécessaire afin de répondre aux enjeux de la médecine de demain en soutenant la prise de décision clinique. De ce fait, ces dernières années ont été caractérisées par un développement considérable de ces outils visant à améliorer le dépistage ou à mieux évaluer les symptômes.

Le signal du cri d'un nourrisson a suscité un fort intérêt de recherche depuis la dernière décennie (Alaie, Abou-Abbas et Tadj, 2016; Chang et al., 2016; Farsaie Alaie et Tadj, 2012; Hariharan et al., 2012; Rui et al., 2010; Várallyay, 2006; Várallyay Jr, Illényi et Benyó, 2008; Verduzco-Mendoza et al., 2012). Les chercheurs ont accumulé assez de données probantes pour conclure qu'un signal du cri contient une information pertinente sur l'état physique et physiologique de la santé du bébé. Ainsi, différentes relations formelles ont été établies entre les propriétés acoustiques des cris et l'état pathologique du bébé. C'est pour cette raison que plusieurs études en cours visent à mettre en place un outil de diagnostic des pathologies néonatales grâce à l'analyse automatique des cris.

Dans ce contexte, notre groupe de recherche s'intéresse à la conception d'un système automatique de diagnostic précoce qui permet d'identifier certaines pathologies chez les nouveau-nés à partir de leurs cris. Ce système vise à améliorer le diagnostic en soutenant la prise de décision clinique afin d'améliorer l'efficacité des traitements et d'éviter l'élaboration de pathologies. La réalisation de ce système de diagnostic requiert des contributions réparties sur trois axes : les outils de traitement des signaux audio, les outils d'extraction des caractéristiques des signaux traités, ainsi que les outils de modélisation et de classification selon les caractéristiques extraites.

En revanche, dans ce travail de recherche, nous nous sommes intéressés en grande partie aux outils de traitement des signaux audio de cris incluant les outils de prétraitement de ces signaux, la distinction entre les différentes activités acoustiques de la base de données ou plus clairement, la segmentation des signaux audio de cris.

Jusqu'à présent, et dans la plupart des travaux de recherche, la segmentation a été réalisée manuellement. Ceci représente une tâche fastidieuse, pouvant requérir des heures de travail par signal. Un tel délai est pratiquement inacceptable et la perspective d'un usage clinique du système de diagnostic visé impose qu'il soit plus raisonnable. Dans ces conditions, le but principal de la présente recherche est la conception d'un module de segmentation automatique des signaux de cris qui doit fonctionner d'une façon robuste, efficace et dans un contexte réel.

0.2 Problématique

Les problématiques de classification des pathologies des signaux de cris des nouveau-nés sont au cœur des préoccupations de notre groupe de recherche du projet NCDS (Newborn Cry-Based Diagnostic System). La Figure 0.1 représente l'architecture générale d'un système de diagnostic automatique.

La conception de ce système automatique de diagnostic requiert une base de données composée de centaines de signaux de cris. Une difficulté à surmonter lorsque nous travaillons avec une telle base de données est le problème de la diversité du contenu des signaux enregistrés. En d'autres mots, les enregistrements collectés à un temps donné contiennent différentes composantes acoustiques autres que les séquences de cris que nous devons directement fournir au système visé. Les signaux peuvent comporter des cris qui sont des suites d'expiration et d’inspiration, des paroles de qualités variables, des périodes de silence et de bruits, etc. Ainsi, les activités acoustiques inutiles nuisent aux processus de l’analyse et du traitement. D’une manière plus précise, le passage au système du diagnostic de toute composante acoustique autre que les expirations et inspirations sonores, qui constituent les parties principales d’un cri, peut provoquer des erreurs de classification des pathologies en réduisant les performances du système global de diagnostic.

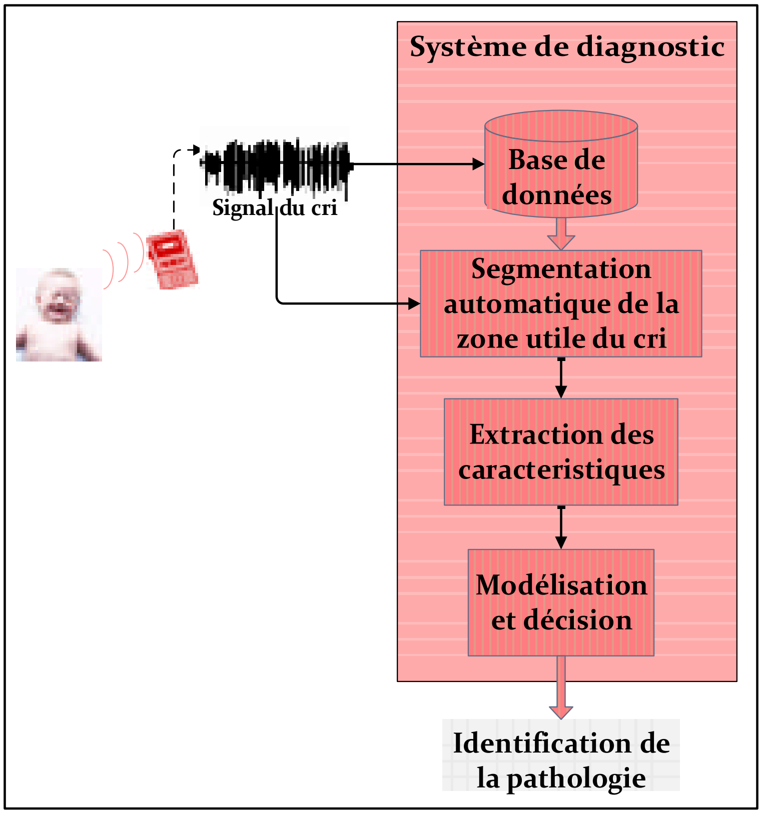

Système de diagnostic
Base de données
Signal du cri
Segmentation automatique de la zone utile du cri
Extraction des caracteristiques
Modélisation et décision
Identification de la pathologie

Figure 0. 1 Schéma global d’un système de diagnostic de la pathologie chez les nouveau-nés à partir de leurs cris

Ainsi, le premier outil indispensable à un projet tel que le projet NCDS consiste en la segmentation automatique des phases d’expiration et d’inspiration dans un flux audio enregistré du cri d’un nouveau-né qui se trouve dans un environnement plus ou moins bruité. Dans les divers travaux antérieurs et jusqu’à l’heure actuelle, la segmentation d’un signal enregistré du cri a souvent été faite manuellement : un opérateur humain écoute et surveille le signal enregistré de façon à ne sélectionner que les parties importantes pour pouvoir ensuite les analyser. D'où l'importance d'automatiser cette tâche considérant le coût élevé en terme de temps écoulé durant la segmentation manuelle ; de plus, le succès de cette dernière dépend de la perception subjective de la personne en charge. Finalement, la non-disponibilité d'un outil de segmentation automatique rend le système de diagnostic inutilisable.

De ce fait, nous nous sommes intéressés, dans un premier temps, à concevoir un module de segmentation automatique afin d'isoler les parties d'expiration et d'inspiration.

Il est important de noter que nous nous sommes basés par la suite sur l'annotation « Expiration » et « Inspiration » afin de désigner les parties des signaux de cris audibles sans tenir compte des inspirations et des expirations non perceptibles, ou plus précisément non vocalisés. Nous nous intéressons donc dans ce projet de recherche au signal cri des nouveau-nés pendant les phases d'expiration et d'inspiration de l'air durant un cri et non pas durant une phase de respiration normale.

Jusqu'à présent, nous ne retrouvons pas dans la littérature un système de segmentation automatique des parties utiles des signaux de cris assez efficace et robuste qui constitue un véritable outil de référence dans ce domaine. Dans ce cadre, la plupart des systèmes de segmentation réalisés sont « faibles » lors du passage des conditions de laboratoires aux conditions réelles. Ceci est expliqué par le fait que ces systèmes opèrent sous des paramètres distincts de ceux pour lesquels l'entraînement a été réalisé. En effet, le NCDS devra fonctionner en temps réel et devra être disposé dans un environnement médical bruité tel que la salle d'accouchement, salle de soins intensifs, autres salles d'hôpital et clinique d'urgence médicale.

C'est dans cet état d'esprit que nous nous sommes situés afin d'aborder d'une façon innovatrice la problématique de la segmentation automatique des cris des nourrissons qui ne possède pas, à ce jour, une solution satisfaisante lorsqu'ils se trouvent dans un contexte réel varié. Dans ce cadre, il est évident que la recherche visant la résolution de cette problématique est tout à fait justifiée. La question posée était la suivante : Est-t-il possible de concevoir un système de segmentation automatique ou un système de détection du cri robuste, capable de localiser précisément les zones importantes de cris parmi autres zones d'activités acoustiques?

0.3 Objectifs de la recherche

Ce travail de recherche a fait partie du développement d'un système de diagnostic précoce des pathologies chez les nouveau-nés à partir des signaux de cris. Ce système doit soutenir la prise de décision clinique, et éviter ainsi le développement des pathologies. Nos travaux dans ce projet ont été à différents niveaux :

1. Créer et gérer un espace collaboratif afin de partager une base de données composée de milliers d'enregistrements de cris provenant des nouveau-nés. Cette base de données servira à mener de multiples études dans le domaine du traitement du signal de cri d'un nouveau-né;

2. Identifier les caractéristiques appropriées capables de distinguer les parties utiles du cris des autres activités acoustiques et surtout de la parole;

3. Identifier les paramètres capables de différencier l'expiration de l'inspiration;

4. Localiser précisément le début et la fin des segments utiles;

5. Développer un outil de segmentation complètement automatique robuste vis-à-vis du bruit et capable de détecter et bien localiser les zones utiles de cris.

0.4 Méthodologie

Pour la réalisation de nos objectifs, nous avons procédé à la démarche suivante :

Dans le premier axe, il était requis de collecter des enregistrements de cris réels en collaborant avec plusieurs hôpitaux. Ces enregistrements doivent être sauvegardés dans un espace partagé aux fins de recherches multiples relatives à la santé des nourrissons. Nous avons consacré en effet d'importantes ressources pour la collecte, l'archivage, la mise à jour, l'accès et la gestion de notre base de données. Au terme de cette partie, nous avons mis à notre disposition un espace collaboratif publié accessible en permanence, contenant une base de données de cris réels qui sera utilisée dans la conception du système du diagnostic précoce prévu.

Dans le deuxième axe, nous nous sommes intéressés tout d'abord à l'étude et à la comparaison des différentes méthodes de segmentation disponibles dans le domaine de traitement des signaux audio afin d'identifier les meilleurs méthodes et techniques qui pourraient être adaptées au signal du cri. Suite à cette étude, nous avons développé une méthode de segmentation robuste et spécifique exploitant la particularité de ce type de signal.

0.4.1 Conception de la base de données

Le travail de recherche dans le cadre du projet NCDS requiert une base de données composée d'un nombre important d'enregistrements audio de cris de nouveau-nés. Malheureusement, et jusqu'à l'heure actuelle, il n'existe aucun corpus constitué d'un tel type de données et dont le contenu est partagé entre les chercheurs pour leur servir dans leurs recherches. Pour cette raison, il a fallu constituer une banque de données formée des enregistrements recueillis de cris des nouveau-nés dans un contexte réel. La formation d'une telle banque pourra être utile pour des chercheurs en donnant la possibilité d'utiliser des données stockées aux fins de diverses recherches relatives à la santé des nourrissons.

0.4.2 Collecte des données

Pour réaliser la collecte des données, un appel a été fait à différents hôpitaux qui acceptent de telles recherches et qui s'intéressent au thème du projet global. Notre étude porte sur les nouveau-nés de six semaines ou moins, sains ou malades, à termes ou prématurés et qui rencontrent des critères d'admissibilité donnés par les responsables médicaux du projet. Aussi, une autorisation signée des parents est requise. Pour chaque participant, il est demandé de recueillir trois enregistrements d'une durée de deux à trois minutes à chaque session. En effet, les collaborateurs médicaux peuvent choisir la situation dans laquelle l'enregistrement se fait et la raison pour laquelle le nouveau-né crie. (Par exemple changement de couche, faim, douleur, etc.) Pour recueillir les cris, un enregistreur Olympus, de très petite taille, est utilisé par l'infirmière de recherche en le plaçant à une distance de 10 cm à 30 cm du bébé.

Toutes les informations pertinentes à la recherche concernant le nouveau-né participant, son état médical ainsi que les cris recueillis doivent être notées pour être ensuite stockées dans la banque de données. Nous pouvons noter par exemple la date de naissance, le poids du bébé, l'âge gestationnel, les pathologies détectées, la date du cri, la raison du cri, etc.

0.4.3 Vue globale pour la construction d'un espace collaboratif

Le corpus qui est utilisé dans le cadre du projet NCDS se compose de centaines d'enregistrements audio des signaux de cris de nouveau-nés. Le moyen le plus efficace de travailler avec une telle base de données est d'utiliser des grilles. En d'autres termes, il est primordial d'afficher les données sous forme de tableaux qui doivent permettre aux utilisateurs de choisir, filtrer, manipuler facilement et travailler efficacement. Il a donc fallu consacrer d'importantes ressources pour la collecte, l'archivage, la mise à jour, l'accès et la gestion de notre base de données. Il était important de considérer les points suivants :

1. Le besoin du stockage de données concerne diverses catégories de personnel. Nous pouvons noter : le groupe du projet NCDS dans les laboratoires et les collaborateurs médicaux chargés de la collecte des signaux de cris dans de multiples localisations (infirmières de recherche dans divers hôpitaux);

2. La nécessité de stocker et sauvegarder cette base de données d'une façon commune et centralisée afin de répondre aux besoins des membres du projet;

3. Il est important que ce volume de stockage dédié à l'équipe du projet NCDS soit disponible en permanence, qu'il soit fonctionnel et en sécurité;

4. La solution de stockage doit être accessible de l'extérieur du laboratoire au travers d'une authentification;

5. Du fait que notre base de données est en cours d'évolution, cette solution de stockage doit prendre en compte l'augmentation des données durant les années courantes.

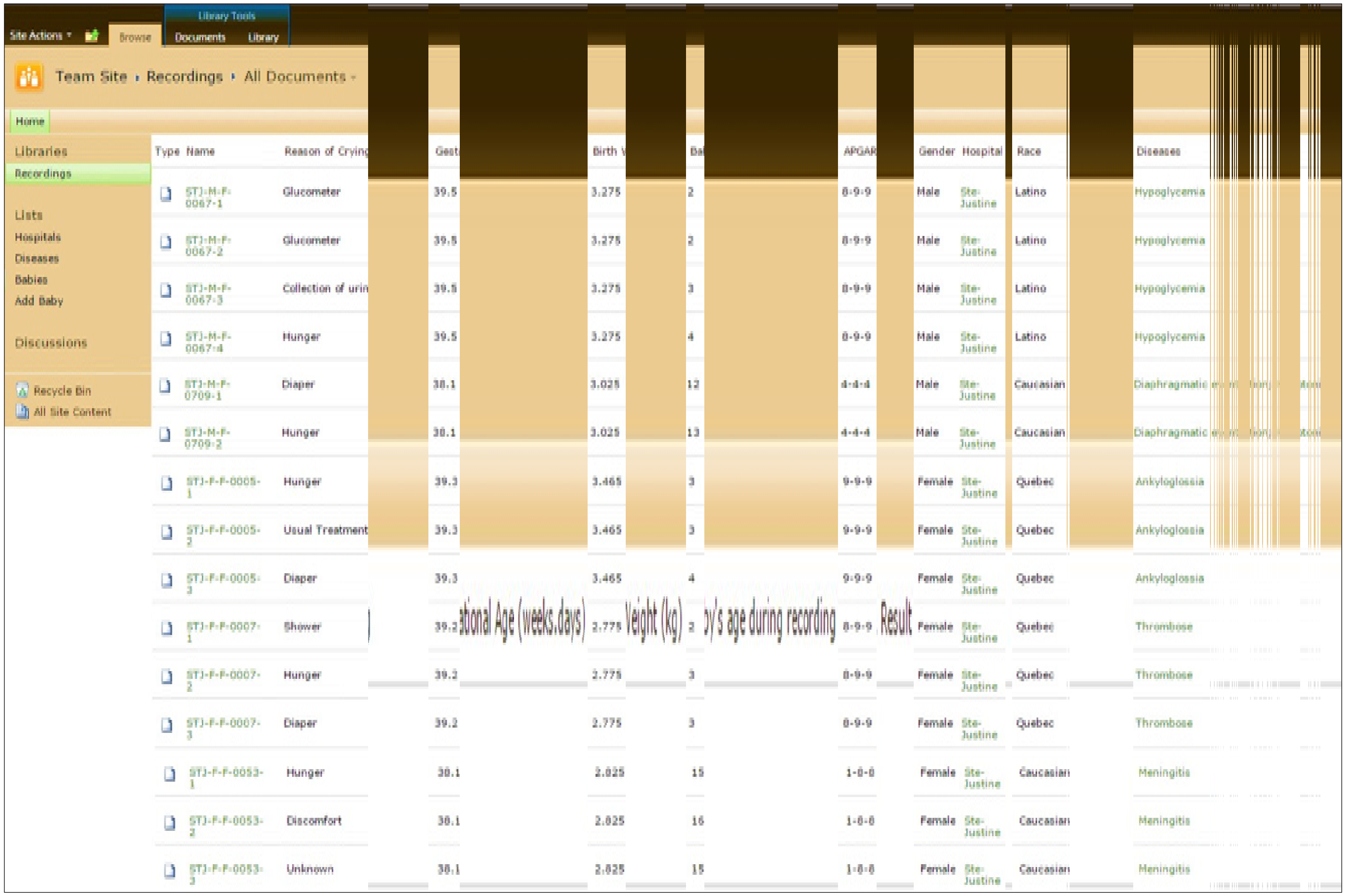
Enfin, il faut prendre en compte que les données ne doivent pas seulement être stockées, mais aussi converties en informations précieuses d'une manière à pouvoir les exploiter en totalité.

0.4.4 Outils utilisés

Afin d'atteindre les objectifs bien ciblés de notre base de données, il était indispensable de mettre en place un serveur de base de données qui permet un stockage direct à haute capacité et d'une façon organisée, structurée et sécurisée. Pour prendre en compte notre besoin d'archivage et de stockage des données centralisées qui doivent être disponibles par réseau sur plusieurs années, nous avons eu recours à la création d'un site SharePoint pour bénéficier des fonctionnalités et capacités de ce dernier, telles que la collaboration, la gestion des données et la recherche. SharePoint, intégré ainsi avec SQL Server qui est un système de gestion des données, a prouvé être un outil de travail collaboratif en fournissant un moyen efficace de partager les données entre un nombre important d'utilisateurs et en améliorant la qualité et la rapidité de la recherche d'informations.

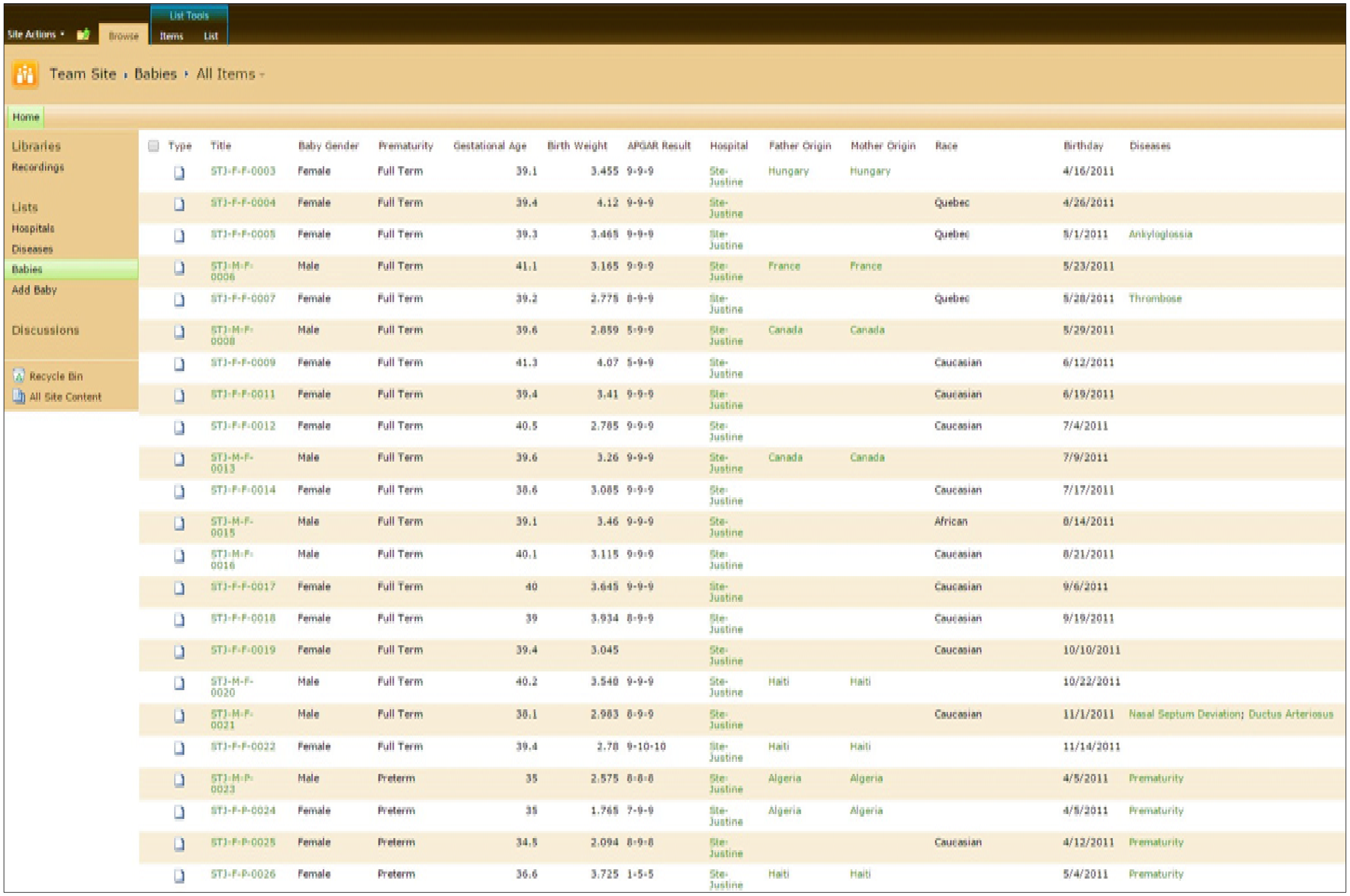
0.4.5 Description fonctionnelle de l'application web

Dans cette phase de travail, nous avons pu mettre à la disposition des personnes désignées pour le travail dans le projet NCDS, une application web accessible en permanence. Cette application permet à un utilisateur, selon sa classe d'authentification, d'accéder, par le biais d’internet, à une interface qui répond à ses besoins. Figures 0.2 et 0.3 ci-dessous présentent un aperçu des sections reliées aux enregistrements et aux bébés respectivement.



Type	Name	Reason of Crying	Gest	Birth W	Ba	APGAR	Gender	Hospital	Race	Diseases
	STJ-M-F-0067-1	Glucometer	39.5	3.275	2	8-9-9	Male	Ste-Justine	Latino	Hypoglycemia
	STJ-M-F-0067-2	Glucometer	39.5	3.275	2	8-9-9	Male	Ste-Justine	Latino	Hypoglycemia
	STJ-M-F-0067-3	Collection of urin	39.5	3.275	3	8-9-9	Male	Ste-Justine	Latino	Hypoglycemia
	STJ-M-F-0067-4	Hunger	39.5	3.275	4	8-9-9	Male	Ste-Justine	Latino	Hypoglycemia
	STJ-M-F-0709-1	Diaper	38.1	3.025	12	4-4-4	Male	Ste-Justine	Caucasian	Diaphragmatic
	STJ-M-F-0709-2	Hunger	38.1	3.025	13	4-4-4	Male	Ste-Justine	Caucasian	Diaphragmatic
	STJ-F-F-0005-1	Hunger	39.3	3.465	3	9-9-9	Female	Ste-Justine	Quebec	Ankyloglossia
	STJ-F-F-0005-2	Usual Treatment	39.3	3.465	3	9-9-9	Female	Ste-Justine	Quebec	Ankyloglossia
	STJ-F-F-0005-3	Diaper	39.3	3.465	4	9-9-9	Female	Ste-Justine	Quebec	Ankyloglossia
	STJ-F-F-0007-1	Shower	39.2	2.775	2	8-9-9	Female	Ste-Justine	Quebec	Thrombose
	STJ-F-F-0007-2	Hunger	39.2	2.775	3	8-9-9	Female	Ste-Justine	Quebec	Thrombose
	STJ-F-F-0007-3	Diaper	39.2	2.775	3	8-9-9	Female	Ste-Justine	Quebec	Thrombose
	STJ-F-F-0053-1	Hunger	38.1	2.025	15	1-8-8	Female	Ste-Justine	Caucasian	Meningitis
	STJ-F-F-0053-2	Discomfort	38.1	2.025	16	1-8-8	Female	Ste-Justine	Caucasian	Meningitis
	STJ-F-F-0053-3	Unknown	38.1	2.025	15	1-8-8	Female	Ste-Justine	Caucasian	Meningitis

Figure 0. 2 Aperçu de l’interface affichant les enregistrements sauvegardés



□ Type	Title	Baby Gender	Prematurity	Gestational Age	Birth Weight	APGAR Result	Hospital	Father Origin	Mother Origin	Race	Birthday	Diseases
🗋	STJ-F-F-0003	Female	Full Term	39.1	3.455	9-9-9	Ste-Justine	Hungary	Hungary		4/16/2011	
🗋	STJ-F-F-0004	Female	Full Term	39.4	4.12	9-9-9	Ste-Justine			Quebec	4/26/2011	
🗋	STJ-F-F-0005	Female	Full Term	39.3	3.465	9-9-9	Ste-Justine			Quebec	5/1/2011	Ankyloglossia
🗋	STJ-M-F-0006	Male	Full Term	41.1	3.165	9-9-9	Ste-Justine	France	France		5/23/2011	
🗋	STJ-F-F-0007	Female	Full Term	39.2	2.775	8-9-9	Ste-Justine			Quebec	5/28/2011	Thrombose
🗋	STJ-M-F-0008	Male	Full Term	39.6	2.859	5-9-9	Ste-Justine	Canada	Canada		5/29/2011	
🗋	STJ-F-F-0009	Female	Full Term	41.3	4.07	5-9-9	Ste-Justine			Caucasian	6/12/2011	
🗋	STJ-F-F-0011	Female	Full Term	39.4	3.41	9-9-9	Ste-Justine			Caucasian	6/19/2011	
🗋	STJ-F-F-0012	Female	Full Term	40.5	2.785	9-9-9	Ste-Justine			Caucasian	7/4/2011	
🗋	STJ-M-F-0013	Male	Full Term	39.6	3.26	9-9-9	Ste-Justine	Canada	Canada		7/9/2011	
🗋	STJ-F-F-0014	Female	Full Term	38.6	3.085	9-9-9	Ste-Justine			Caucasian	7/17/2011	
🗋	STJ-M-F-0015	Male	Full Term	39.1	3.46	9-9-9	Ste-Justine			African	8/14/2011	
🗋	STJ-M-F-0016	Male	Full Term	40.1	3.115	9-9-9	Ste-Justine			Caucasian	8/21/2011	
🗋	STJ-F-F-0017	Female	Full Term	40	3.645	9-9-9	Ste-Justine			Caucasian	9/6/2011	
🗋	STJ-F-F-0018	Female	Full Term	39	3.934	8-9-9	Ste-Justine			Caucasian	9/19/2011	
🗋	STJ-F-F-0019	Female	Full Term	39.4	3.045		Ste-Justine			Caucasian	10/10/2011	
🗋	STJ-M-F-0020	Male	Full Term	40.2	3.540	9-9-9	Ste-Justine	Haiti	Haiti		10/22/2011	
🗋	STJ-M-F-0021	Male	Full Term	38.1	2.983	8-9-9	Ste-Justine			Caucasian	11/1/2011	Nasal Septum Deviation; Ductus Arteriosus
🗋	STJ-F-F-0022	Female	Full Term	39.4	2.78	9-10-10	Ste-Justine	Haiti	Haiti		11/14/2011	
🗋	STJ-M-P-0023	Male	Preterm	35	2.575	8-8-8	Ste-Justine	Algeria	Algeria		4/5/2011	Prematurity
🗋	STJ-F-P-0024	Female	Preterm	35	1.765	7-9-9	Ste-Justine	Algeria	Algeria		4/5/2011	Prematurity
🗋	STJ-F-P-0025	Female	Preterm	34.5	2.094	8-9-8	Ste-Justine			Caucasian	4/12/2011	Prematurity
🗋	STJ-F-P-0026	Female	Preterm	36.6	3.725	1-5-5	Ste-Justine	Haiti	Haiti		5/4/2011	Prematurity

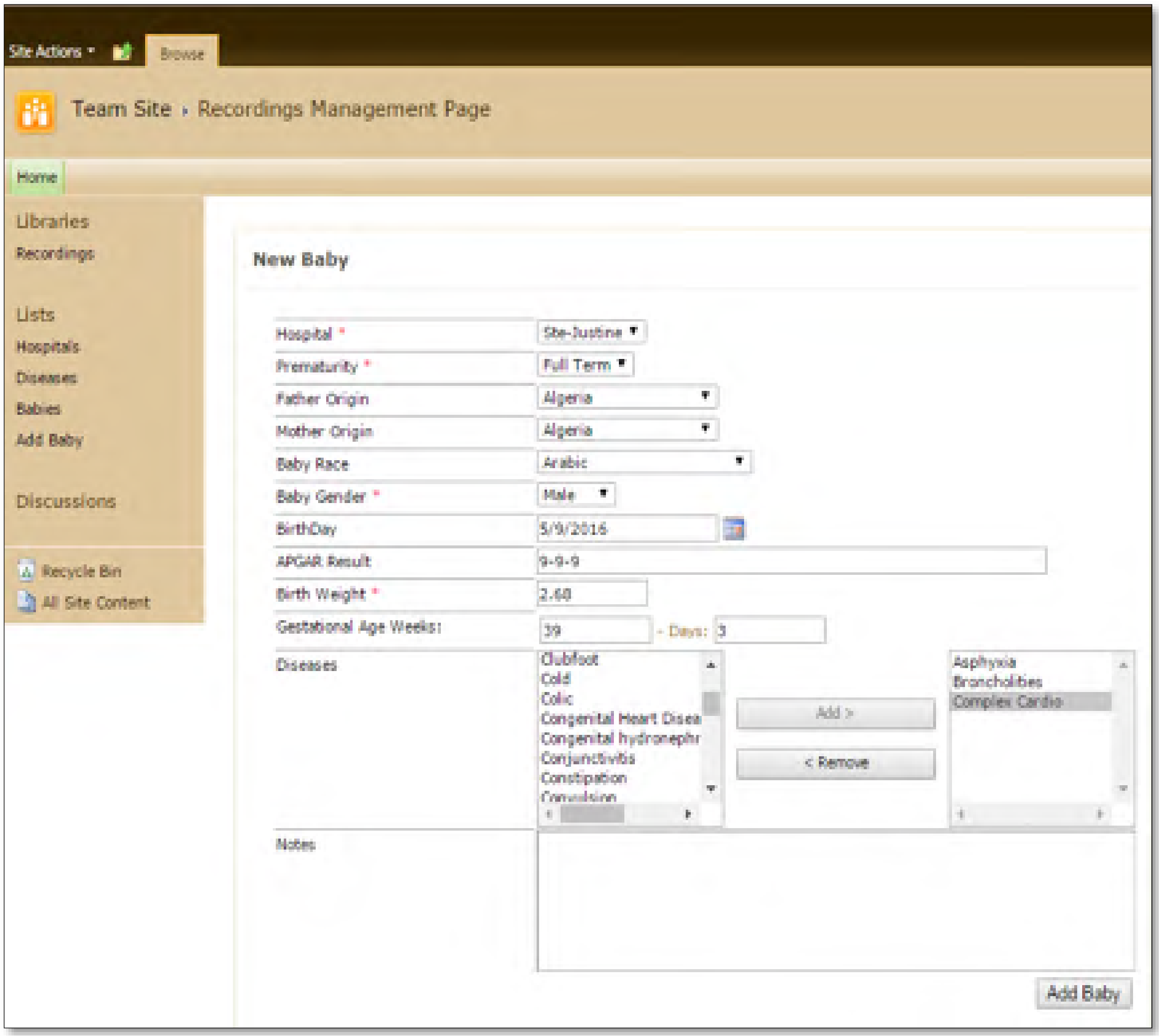
Figure 0. 3 Aperçu de l’interface affichant les informations sur les bébés participants

En utilisant le navigateur web, les membres du groupe du projet NCDS doivent pouvoir :

1. Consulter et télécharger des fichiers audio enregistrés des cris des nourrissons d'une façon sécurisée. (Un compte d'utilisateur est requis pour accéder à la base de données);

2. Trier et filtrer les enregistrements selon des critères spécifiques (par exemple : enregistrements des bébés à terme ou prématurés, des bébés sains ou malades, etc.);

3. Ajouter des documents traités liés aux enregistrements sur l'espace collaboratif.



New Baby

Hospital *	Ste-Justine
Prematurity *	Full Term
Father Origin	Algeria
Mother Origin	Algeria
Baby Race	Arabic
Baby Gender *	Male
BirthDay	5/9/2016
APGAR Result	9-9-9
Birth Weight *	2.60
Gestational Age Weeks:	39 - Days: 3
Diseases	Asphyxia Broncholities Complex Cardio
Notes	

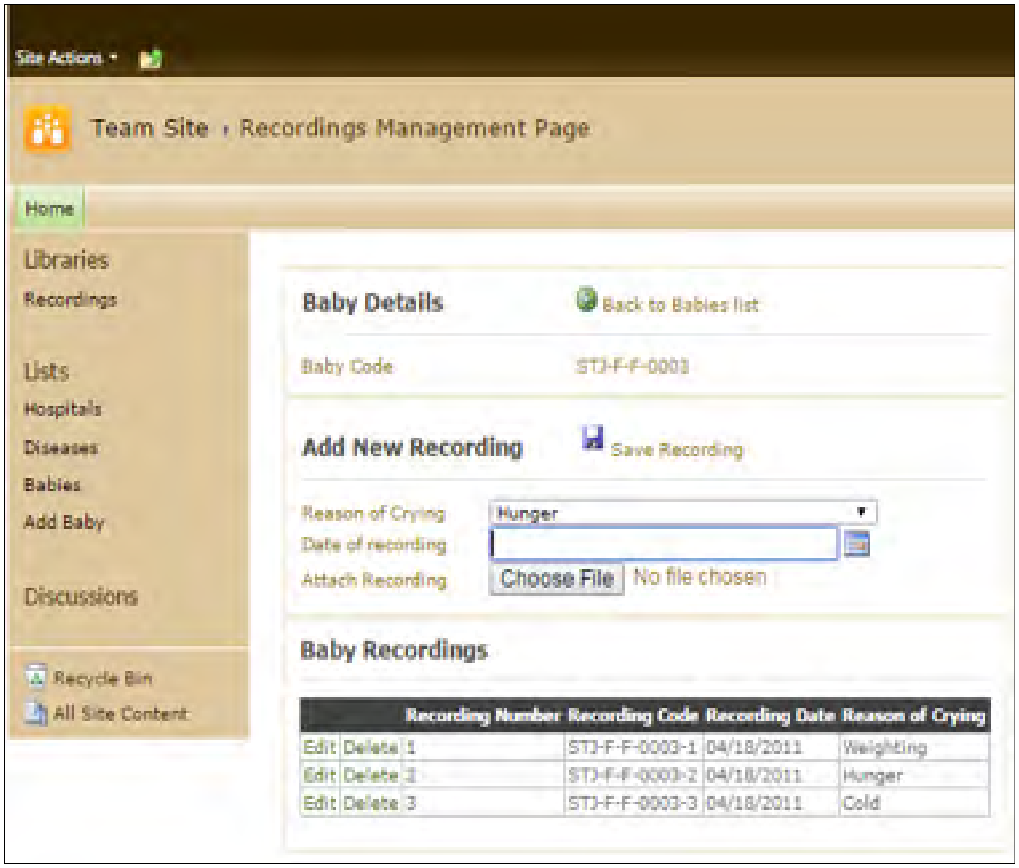
Add Baby

Figure 0. 4 Interface du site web réservée à l'entrée des informations reliées aux bébés par l'infirmière

D’autre part, les infirmières de recherche chargées de la collecte des signaux de cris doivent être capable de :

1. Stocker les informations des nouveau-nés participant dans le projet même si elles proviennent de plusieurs endroits. (dans notre cas : hôpital Saint-Justine au Canada, hôpital général Sahel au Liban). Voir Figure 0.4;

2. Télécharger vers le serveur les enregistrements captés des nouveau-nés sans perdre aucune information. Voir Figure 0.5;



Baby Details

Baby Code STJ-F-F-0003

Add New Recording

Reason of Crying: Hunger
Date of recording:
Attach Recording: No file chosen

Baby Recordings

		Recording Number	Recording Code	Recording Date	Reason of Crying
Edit	Delete	1	STJ-F-F-0003-1	04/18/2011	Weighting
Edit	Delete	2	STJ-F-F-0003-2	04/18/2011	Hunger
Edit	Delete	3	STJ-F-F-0003-3	04/18/2011	Cold

Figure 0. 5 Interface du site web pour le téléchargement des enregistrements pour chaque bébé

3. Mettre à jour l’information du suivi du bébé à l’âge de 6 mois pour la validation de l’étude ciblée par ce projet.

0.4.6 Conception du module de segmentation automatique

Notre objectif principal a été la réalisation d'un système de segmentation automatique des signaux de cris le plus efficace possible en terme de précision temporelle. Ce système consiste à détecter d'une manière précise les zones pertinentes du cri qui sont les expirations et les inspirations audibles. Ce système doit servir de nombreux systèmes d'analyse d'un cri. Notre travail a été divisé en deux parties. La première partie a été consacrée à la réalisation d'un système basé sur des approches supervisées. Figure 0.6 ci-dessous illustre le schéma global d'un système de segmentation. Le détail des méthodes utilisées pour réaliser l'étape I de travail se trouve dans les deux premiers articles de journaux publiés (Chapitres 2 et 3 de cette thèse) ainsi que dans les deux articles de conférence publiés (Annexes I et II).

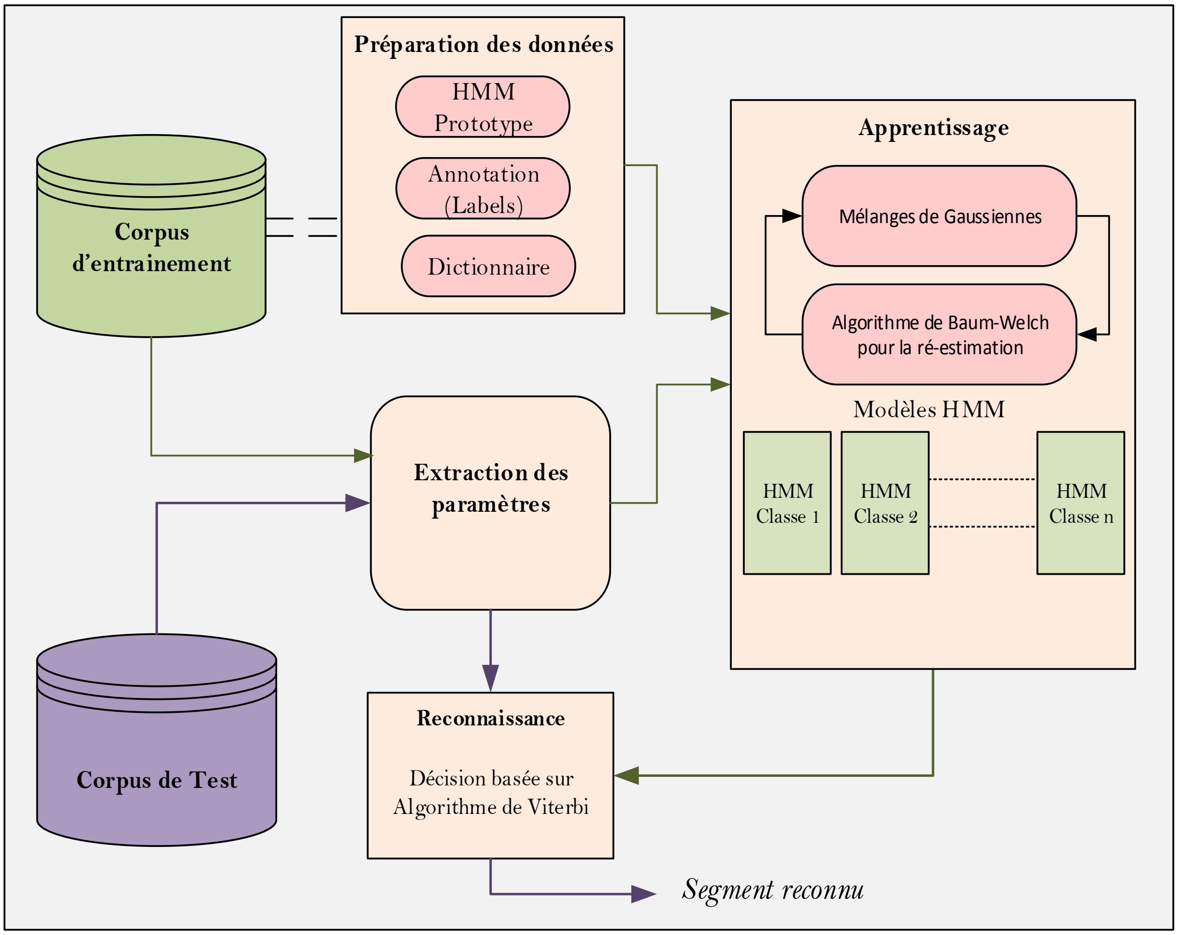


Figure 0. 6 Schéma global d'un système de segmentation basé sur une approche supervisée

La deuxième partie a été consacrée à l'ajout de la partie post-traitement afin d'assurer une segmentation robuste. La Figure 0.7 illustre l'architecture générale du module de segmentation basée sur les approches Markoviennes avec la phase de post-traitement. La phase de prétraitement consiste en la normalisation ainsi que le fenêtrage du signal à traiter. En effet, un signal du cri comme un signal de parole peut être perçu comme un processus aléatoire non stationnaire. Pour cette raison, afin de le considérer comme stationnaire, ce signal doit être traité dans une fenêtre temporelle comprise normalement entre 10 et 50 ms. Ce fait a permis ainsi de mettre en œuvre des méthodes robustes d'analyse et de modélisation du signal stationnaire. La phase de post-traitement a été requise à ce niveau afin 1) d'ajuster les bornes des segments expiration et inspiration et 2) conserver une séquence de cris y compris la suite des expirations et des inspirations alternées par des périodes de silence. La raison de ce choix est provenu du fait qu'il est nécessaire de calculer le temps du cri, le nombre d'expirations et d'inspirations ainsi que le temps de la latence du cri et d'autres mesures. Le détail de cette étape de travail est présenté dans notre troisième article de journal (Chapitre 4).

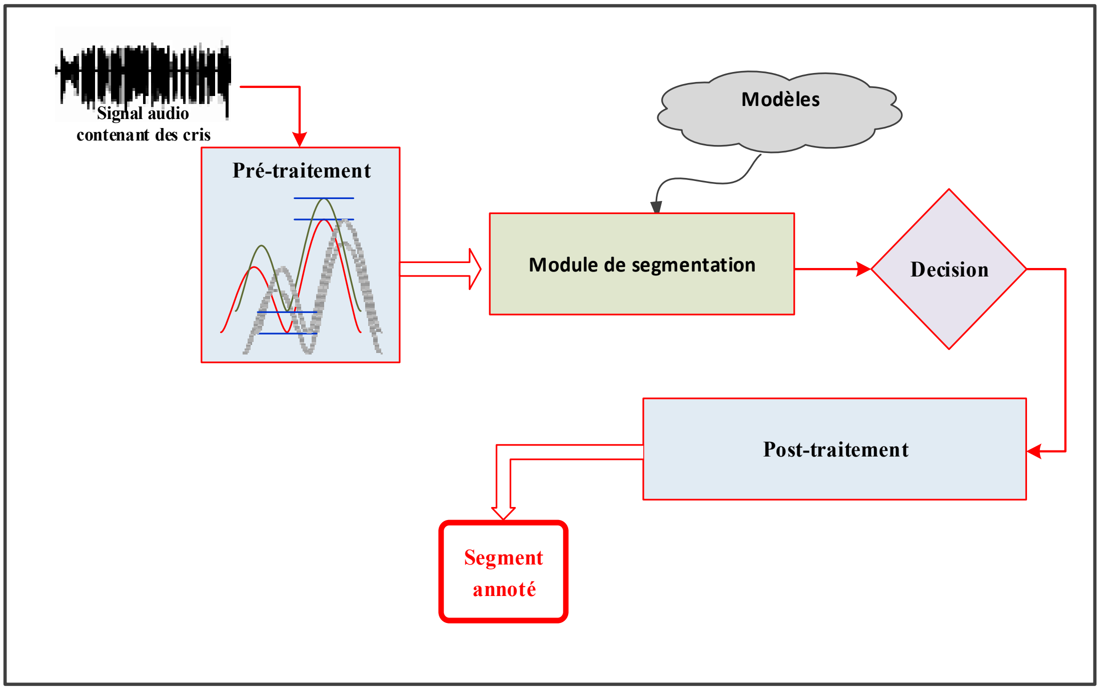


Figure 0. 7 Architecture générale du module de segmentation avec une phase de post-traitement

Le module de segmentation adopté dans notre approche est divisé en deux étapes principales : l'extraction des caractéristiques et la classification.

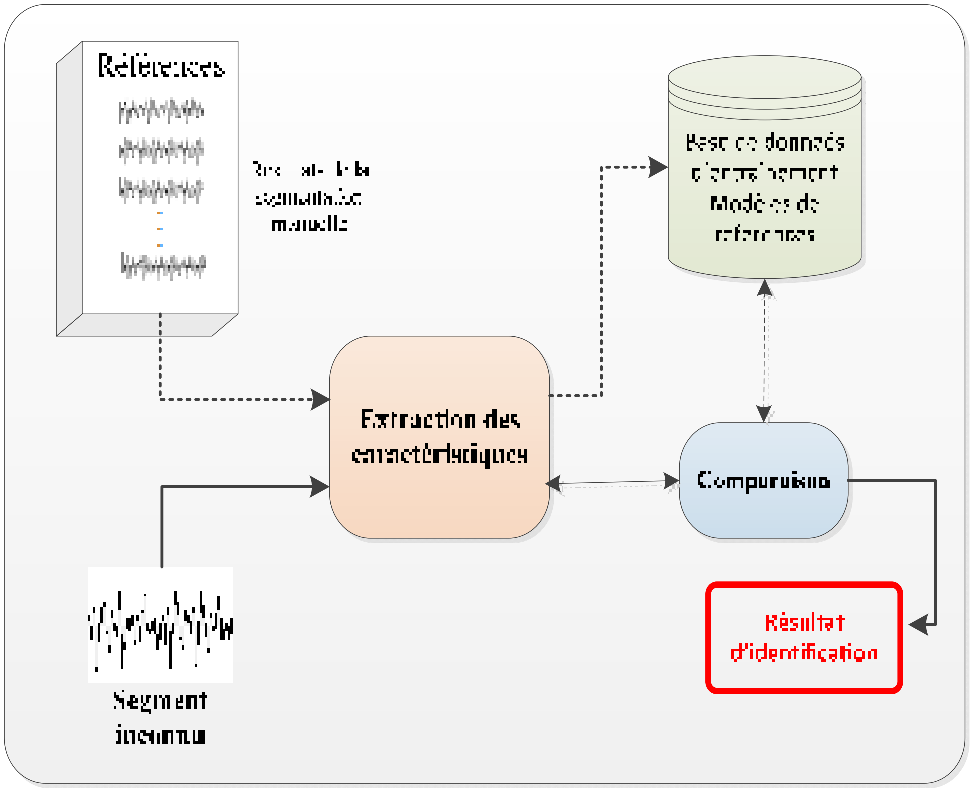

Références
Résultats de la segmentation manuelle
Base de données d'entraînement Modèles de références
Extraction des caractéristiques
Comparaison
Segment inconnu
Résultat d'identification

Figure 0. 8 Schéma général du module de segmentation

Nous avons vu qu'il existe de nombreuses méthodes de paramétrisation et de classification utilisées dans la littérature aux fins du sujet de segmentation audio. Chacune de ces méthodes présente des avantages et des inconvénients. Ainsi en s'inspirant de ces méthodes, cette section de travail a porté sur la recherche des meilleurs paramètres qui sont aptes à améliorer les performances de la tâche de segmentation des signaux de cris pour la détection des segments expirations et inspirations audibles. Les signaux enregistrés sont convertis en une série de vecteurs d'observation et ainsi des modèles HMM sont construits (par exemple, EXP-INS-AUTRE, etc.). Une phase de reconnaissance et d'évaluation de la performance sur un corpus de test a suivi. Voir Figure 0.9.

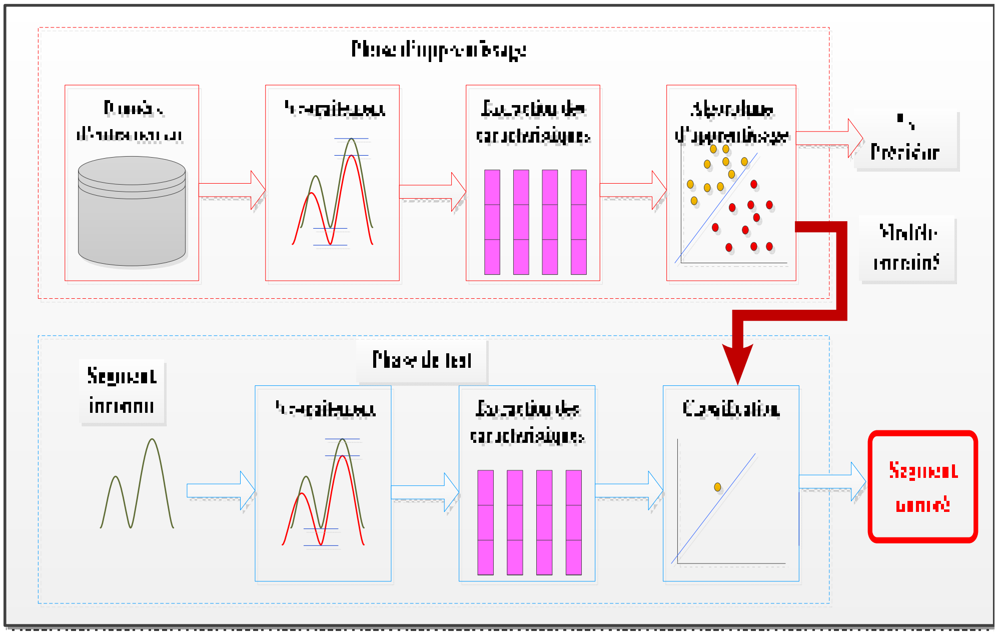


Figure 0. 9 Phase d'apprentissage et d'entrainement d'un module de segmentation

0.4.7 Méthode d'évaluation des performances de la segmentation automatique

L'évaluation des performances est nécessaire d'une part pour permettre une comparaison entre les différentes méthodes de segmentation, y compris la segmentation manuelle et d'autre part, pour mettre en évidence les points qui nécessitent une amélioration afin d'obtenir de meilleurs résultats. Ainsi, l'évaluation que nous avons utilisée est celle qui consiste à calculer l'exactitude et la précision résultant de la comparaison entre la segmentation automatique et la segmentation manuelle.

0.4.8 Outils utilisés

Nous avons choisi de mener notre étude en utilisant la plateforme logicielle HTK (Hidden Markov Model Toolkit). Cet outil constitue un environnement de recherche assez efficace dans le domaine de la reconnaissance de la parole. HTK forme un véritable banc d'essais en réduisant au minimum la tâche de programmation des différentes phases d'un système de segmentation à base des modèles de Markov cachés.

Deux phases principales ont été requises : phase d'apprentissage ou d'estimation de modèles et phase de test. Pour la partie apprentissage, une phase de travail laborieuse d'étiquetage manuel a été requise. Nous avons eu recours à un outil logiciel WaveSurfer qui permet de visualiser la forme d'onde et le spectrogramme du signal, en s'aidant en même temps de l'écoute en boucle afin d'annoter les transcriptions sur ce même signal et prendre une décision sur l'emplacement final des bornes. Figure 0.10 présente un exemple d'étiquetage.

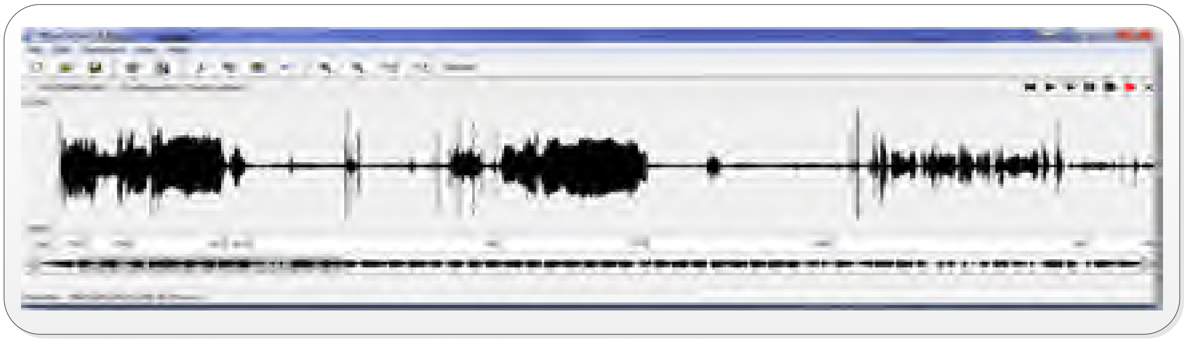
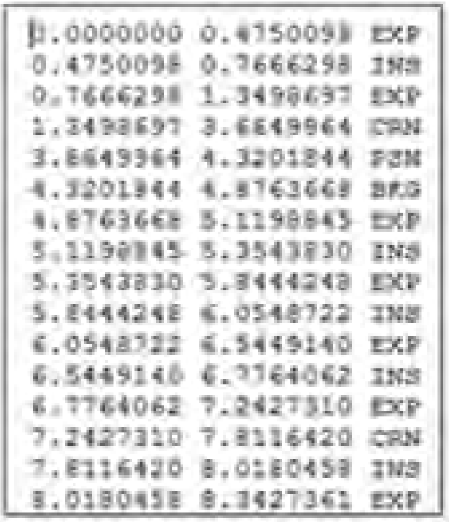


Figure 0. 10 Exemple d'étiquetage manuel en utilisant Wavesurfer

À la fin de cette procédure, nous avons obtenu une liste de segments dont les temps de début et de fin en millisecondes avec l'étiquette indiquant la catégorie sont indiqués. Voir Figure 0.11.



```
0.0000000 0.4750098 EXP
0.4750098 0.7666298 INS
0.7666298 1.3498697 EXP
1.3498697 3.8649964 CRN
3.8649964 4.3201844 PSM
4.3201844 4.8763668 BRG
4.8763668 5.1190845 EXP
5.1190845 5.3543830 INS
5.3543830 5.8444248 EXP
5.8444248 6.0548722 INS
6.0548722 6.5449140 EXP
6.5449140 6.7764062 INS
6.7764062 7.2427310 EXP
7.2427310 7.8116420 CRN
7.8116420 8.0180458 INS
8.0180458 8.3427361 EXP
```

Figure 0. 11 Exemple du contenu
d'un fichier segmenté

0.5 Organisation de la thèse

Les travaux réalisés lors de cette thèse sont présentés sous forme de deux articles publiés et un article soumis pour publication dans des revues scientifiques ainsi que deux articles de conférence publiés avec comité de lecture.

Cette thèse est structurée comme suit. Une introduction est consacrée à la description du projet global, la situation de notre problématique de recherche, les objectifs ainsi que la méthodologie suivie durant nos parcours d'étude. La revue de la littérature est présentée dans le chapitre 1. Les trois chapitres qui suivent sont des articles soumis et/ou publiés dans des revues internationales. Dans le deuxième chapitre, nous présentons notre premier article publié dans le journal « Biomedical Signal Processing and Control » :

Abou-Abbas, Lina, Hesam Fersaie Alaie et Chakib Tadj. 2015. « Automatic detection of the expiratory and inspiratory phases in newborn cry signals ». *Biomedical Signal Processing and Control,* vol. 19, p. 35-43.

Dans cet article, nous décrivons nos travaux de recherche établissant un système automatique de segmentation des cris de nouveau-nés enregistrés, basé sur les modèles de Markov cachés en utilisant HTK. Le système présenté dans ce travail est capable de détecter les deux constituants de base d'un cri, qui sont les parties acoustiques d'expiration et d'inspiration, en utilisant une architecture de reconnaissance audio à deux étapes. Le système est entrainé et testé sur une base de données de cris recueillis auprès des nouveau-nés sains et pathologiques. Les résultats expérimentaux indiquent que le système donne une précision allant jusqu'à de 83,79%.

Le troisième chapitre concerne notre second article publié dans « Voice Journal »

Abou-Abbas, Lina, Chakib Tadj, Christian Gargour et Leila Montazeri. 2016. « Expiratory and Inspiratory Cries Detection Using Different Signals' Decomposition Techniques ». *Voice Journal*, in Press.

L'approche utilisée dans ce travail est composée de trois étapes : décomposition du signal, extraction des caractéristiques et classification. Dans la première phase, la transformée de Fourier rapide, la décomposition en modes empiriques et la transformée en ondelettes ont été considérées. Dans la seconde étape, divers paramètres ont été extraits et investigués et dans la troisième étape, deux méthodes d'apprentissage supervisées, GMM et HMM avec 4 et 5 états ont été aussi discutées. Les expériences ont été répétées plusieurs fois pour différents groupes de données d'apprentissage et de test, aléatoirement choisis en utilisant la technique de validation croisée 10 fois. Un taux d'erreurs global de classification de 8,9% a été obtenu.

Le quatrième chapitre présente notre troisième article soumis au « The Journal of Acoustical Society of America » en juin 2016.

Abou-Abbas, Lina, Chakib Tadj et Hesam Fersaie Alaie. June 2016. «A Fully Automated Approach for Baby Cry Sound Segmentation in a Realistic Clinical Environment and Boundary Detection of Corresponding Expiratory and Inspiratory Episodes ». *The Journal of Acoustical Society of America, submitted*

Dans cet article, nous avons proposé une approche complète de segmentation automatique des cris qui tient compte de la détection précise des bornes de début et fin des segments expiration et inspiration. Nous avons étudié la contribution de certaines caractéristiques temporelles et fréquentielles sur la robustesse d'un système de segmentation des cris basé sur des approches supervisées. L'objectif principal de cet article était d'étendre les systèmes développés dans nos travaux précédents pour inclure une étape de post-traitement avec un ensemble d'outils de correction pour améliorer les performances de classification. Cette approche complète permet de localiser précisément les segments expirations et inspirations audibles, permettant aux segments détectés d'être utilisés dans de nombreuses applications nécessitant une analyse du cri. Les résultats expérimentaux ont montré l'efficacité de la solution proposée avec des taux de détection d'expiration et d'inspiration sonores d'environ 94,29 et 92,16%, respectivement, en appliquant une technique de validation croisée dix fois.

A la fin de cette thèse, nous résumons nos travaux de recherche, les contributions achevées ainsi que les perspectives recommandées dans la conclusion.

En Annexe I, nous présentons notre premier article de conférence publié.

Abou-Abbas, Lina, Hesam Fersaei Alaei et Chakib Tadj. 2015. « Segmentation of voiced newborns' cry sounds using Wavelet Packet based features ». In *Electrical and Computer Engineering (CCECE), 2015 IEEE 28th Canadian Conference* (Halifax, Canada, 3-6 May).

Dans cet article, nous avons proposé une méthode de segmentation automatique des signaux de cris des nouveau-nés en se basant essentiellement sur l'analyse par paquets d'ondelettes en raison de sa richesse dans la résolution fréquentielle. Par rapport à l'énergie calculée dans chaque bande de fréquence, nous avons constaté que l'opérateur Teager Energy Kaiser donne des résultats plus efficaces dans la segmentation. Nous avons testé des décompositions en paquets d'ondelette de niveaux 5, 6 et 7. Nous avons montré qu'il est possible de détecter les phases d'expirations et d'inspirations audibles du cri avec un taux de précision allant jusqu'à 84.08%.

L'Annexe II concerne notre second article de conférence publié.

Abou-Abbas, Lina., Leila. Montazeri, Christian. Gargour et Chakib. Tadj. 2015. « On the use of EMD for automatic newborn cry segmentation ». In *Advances in Biomedical Engineering (ICABME), 2015 International Conference on*. (16-18 Sept. 2015), p. 262-265.

Dans cet article, nous avons exploité la nouvelle technique de décomposition des signaux non stationnaires, la décomposition en mode empirique (EMD), appliquée à des signaux de cris. Nous avons montré que les coefficients cepstraux extraits des fonctions modales intrinsèques (IMF) sont efficaces pour distinguer les parties de cris voisées parmi les autres activités acoustiques.

CHAPITRE 1

DEFINITIONS ET REVUE DE LA LITTÉRATURE

1.1 Le cri d'un nouveau-né

Du point de vue acoustique, une séquence de cris constitue une série de longues expirations séparées par des intervalles brefs d'inspirations.

Un cri est le résultat du bruit que génèrent les cordes vocales en vibrant sous la pression de l'air. Différentes études sur les cris des nouveau-nés ont montré un certain nombre de constantes dans les valeurs des caractéristiques acoustiques. L'intensité d'un cri peut dépasser 80 décibels, sa fréquence moyenne se situe aux alentours de 440 Hz et la modulation de l'intensité et de la fréquence évolue selon une courbe ascendante descendante (Zei 1995). Cependant, des études conduites dès 1968, prouvent qu'il est évident d'observer des particularités comme le timbre, la durée et l'intensité en fonction de l'âge, la maturité, le genre ou bien la nature de ce qui déclenche un cri comme par exemple, un cri de naissance, la douleur, la faim et le plaisir (Wasz-Höckert et al., 1968). Par exemple, un cri de naissance se distingue par le fait qu'il correspond à l'entrée de l'air pour la première fois dans le corps du nouveau-né. Ce cri débute par une inspiration puissante bien supérieure à celle d'une inspiration ordinaire. À cette inspiration, suit une très forte expiration. Le cri est normalement aigu, non voisé et contient des explosives. Un cri de douleur est particulier et attire l'attention de l'entourage tout de suite. Il commence par une expiration d'une durée de plus de 4 secondes suivie par un silence puis d'une inspiration déclenchant une série de cris. Son pitch est normalement élevé. Le cri de la faim est en moyenne d'une durée d'une seconde. Il est généralement précédé par une inspiration courte et aigüe et recommence après une pause. Le cri du plaisir est propre à chaque bébé, d'une intensité et d'une tonalité variables (Hubin-Gayte, 2012). Une analyse spectrale de différentes caractéristiques acoustiques de divers cris pathologiques a été introduite pour la première fois par Michelsson (Michelsson et Michelsson, 1999). Les chercheurs sont parvenus à séparer les diverses catégories de cris en fonction de la tonalité, de la fréquence, de l'intensité et d'autres attributs acoustiques. En effet, les expirations et les inspirations audibles constituent les parties utiles pour le traitement d'un signal du cri dans différents contextes. D'après l'encyclopédie médiathèque, une inspiration correspond à « une action qui consiste, à l'aide du diaphragme, à rassembler dans les poumons, via la trachée l'air nécessaire à l'expiration et donc à la phonation », qui est à son tour le phénomène de la production des sons par les organes vocaux. Par contre, l'expiration correspond à l'action par laquelle l'air est expulsé des poumons et qui a toujours lieu après un mouvement d'inspiration. Golub a classifié les parties utiles du cri, relatives aux expirations et inspirations, sous quatre catégories principales : (Golub, 1979)

1. Phonation expiratoire;
2. Hyperphonie expiratoire;
3. Dysphonie expiratoire;
4. Phonation inspiratoire.

Ainsi, une phonation expiratoire ou inspiratoire correspond aux parties acoustiques audibles du cri durant une phase d'expiration ou d'inspiration. La phonation expiratoire est normalement voisée et résultante de la vibration des cordes vocales avec une fréquence fondamentale F0 pouvant varier de 250 jusqu'à 750 Hz. La phonation inspiratoire correspond à la production d'un son ou plutôt à la vibration des cordes vocales durant une inspiration. Elle est appelée chez les adultes voix inspiratoire. Dans les archives générales de médecine, une voix inspiratoire a été définie comme étant « un bruit étouffé qui n'est engendré qu'avec effort ». Chez les hommes, cette voix s'observe pendant le rire, le hoquet ou le pleur. La voix ou phonation inspiratoire, chez les hommes comme chez les nouveau-nés, représente des sons d'une extrême acuité et d'un pitch très élevé. Une hyperphonie est un type d'expiration audible qui se caractérise par une fréquence fondamentale très élevée, entre 1000 et 2000 Hz. Enfin, une dysphonie ou encore appelée turbulence, contient à la fois des sources de sons périodiques et apériodiques et se produit quand un bruit de turbulence est généré par les cordes vocales.

1.2 Revue de la littérature

Plusieurs recherches ont été menées sur le sujet d'analyse des cris du nourrisson dans différents contextes (classification du type de cri, classification de la pathologie à partir du signal du cri, etc.). Une phase commune aux différents sujets est celle de la segmentation du signal. À l'exception des travaux de certains auteurs (Aucouturier et al., 2011; Orlandi et al., 2013; Várallyay Jr, Illényi et Benyó, 2008) que nous allons aborder plus tard, la segmentation d'un signal cri n'a pas attiré l'intérêt des chercheurs. Dans la plupart des travaux, la segmentation des signaux de cris a été réalisée manuellement (Messaoud et Tadj, 2010; Michelsson et Michelsson, 1999; Proctor, 1984; Wermke et al., 2002). Dans d'autres recherches menées sur l'analyse des cris, une phase de segmentation automatique commune a précédé les étapes de traitement du signal. Cette phase commune entre les différentes recherches a souvent été basée sur les méthodes connues de détection de l'activité vocale (VAD, Voice Activity Detection). VAD permet uniquement de discriminer entre le bruit et les parties vocales. Dans le cas d'un contexte réel, un VAD n'est pas en mesure de distinguer d'une part, entre des signaux de paroles et de cris et d'autre part, entre des segments d'expiration et d'inspiration. D'autres méthodes temporelles basées sur la détection de silences ont été fréquemment utilisées dans la littérature. Ces méthodes reposent en général sur le calcul de l'énergie du signal. Le taux de passage par zéro à court terme ZCR (Zero Crossing Rate) et l'énergie à court terme STE (Short Time Energy) sont les deux descripteurs les plus couramment utilisés pour la segmentation, en introduisant quelques modifications sur le choix des seuils (Kuo, 2010; Rui et al., 2010; Várallyay, 2006; Zabidi et al., 2009a). L'utilisation de ces deux méthodes, employées seule ou de façon combinée, ne permet pas de distinguer les parties voisées ou non d'un signal du cri comme cela a déjà été mentionné par la plupart des auteurs. D'autre part, dans ces travaux, la base de données employée était composée uniquement de cris de nouveau-nés alternés par des silences ou bien de faibles bruits d'environnement (ou ceux dus aux microphones), qui ont été éliminés par des filtres appliqués en amont du système d'analyse. Alors, le but des auteurs dans ces travaux était seulement d'éliminer les parties de silence (ou encore appelées non voisées) sans distinguer entre les parties expiration et inspiration d'un cri.

De rares études ont été menées spécifiquement sur la segmentation automatique des signaux de cris. Nous pouvons nous arrêter sur les études conduites par Varallyay en 2008 et 2009 (Várallyay, Illényi et Benyó, 2009; Várallyay Jr, Illényi et Benyó, 2008). Dans ces travaux, l'auteur a utilisé la méthode de la détection de la fréquence fondamentale HPS (Harmonic Product Spectrum). L'idée de base était de vérifier si le signal présente des harmoniques ou non pour conclure s'il provient d'un conduit vocal humain ou si c'est le résultat des effets indésirables ou du bruit. Dans ce dernier cas, la partie du signal doit être rejetée. L'auteur a montré qu'il est capable de classifier le contenu des signaux de cris afin de distinguer entre les segments du cri (qui sont les parties de l'expiration résultant de la vibration des cordes vocales) et les segments de l'inspiration (Várallyay Jr, Illényi et Benyó, 2008). Varallyay n'a pas pris en compte les inspirations sonores qui peuvent être générées fréquemment par des bébés malades et qui peuvent contenir une information importante sur l'état pathologique du bébé.

En effet, nous ne pouvons pas ignorer les segments dysphoniques dans les signaux de cris. Normalement, ces segments représentent clairement des perturbations (ou encore des bruits) durant un cri et il est très fréquent qu'ils n'aient pas une forme harmonique.

Une autre recherche portant spécifiquement sur la segmentation des signaux de cris a été menée en 2012. Le but était de marquer chaque segment selon les trois catégories : cri/non cri/non-activité. De même, cette recherche était basée sur l'étude du contenu spectral et l'harmonicité du signal sans tenir compte des segments dysphoniques ni des segments d'inspirations audibles qui peuvent être pertinents pour la partie analyse du signal d'un cri pathologique.

Une étude intéressante a été réalisée par J.Aucouturier (Aucouturier et al., 2011). L'auteur a utilisé une modélisation markovienne pour la reconnaissance ou la segmentation des signaux de cris. Il a modélisé trois catégories : Expiration, Inspiration et Silence. La base de données utilisée dans ce travail était formée seulement des cris qui sont des suites d'expirations et d'inspirations qui peuvent être alternées par des périodes de silence ou de bruit de l'environnement. Chaque modèle était représenté par un GMM. Pour la phase d'apprentissage, au lieu d'utiliser l'algorithme de Baum-Welch, l'auteur a choisi de comparer les performances des deux techniques SVM (Machines à vecteurs de support) et GMM (modèles de mixture gaussienne). Puisque les composantes acoustiques de la base de données utilisée dans cette recherche sont très limitées, les résultats obtenus étaient élevés.

Une autre étude réalisée par Orlandi et al. (Orlandi et al., 2013) semble intéressante. Dans ce travail, les auteurs ont proposé un système automatique pour la detection des événements sonores en appliquant un sous-echantillonnage suivi d'un filtrage et d'un calcul de l'energie à court-terme. Cette méthode encore ne peut détecter que les segments voisés.

Il est évident qu'afin de manipuler notre système, il a fallu s'inspirer des travaux de recherche présents dans la littérature, associés au traitement des signaux audio dans différentes applications, par exemple : segmentation de la parole en mots isolés, segmentation de la parole en syllabes, segmentation en phonèmes, segmentation en son voisé/non voisé, segmentation de la parole parmi plusieurs types de contenu sonore comme la musique, la voix téléphonique, la voix dans un fond bruité, segmentation parole/silence, segmentation en locuteurs et tours de parole, et encore segmentation des signaux cardiaques.

Le problème de la segmentation automatique de la parole se trouve sous différents noms dans la littérature; par exemple « Voice Activity Detection », « Speech Detection » et « Endpoint Detection ». Plusieurs méthodes de VAD ont été proposées dans la littérature, plus spécifiquement pour la détection de la parole. Nous citons par exemple les travaux suivants: (Dongwen et al., 2011; Lu, Jiang et Zhang, 2001) utilisent l'énergie ; (L.R. Rabiner, 1975) se base sur le calcul de STE et ZCR ; (Ramírez et al., 2004) calculent la divergence spectrale à long terme entre la parole et le bruit pour effectuer la segmentation ; (Saunders, 1996) a utilisé le ZCR et le centroide spectral afin de séparer le bruit des parties voisées; (Davis, Nordholm et Togneri, 2006) opèrent la segmentation en se basant sur la mesure du rapport signal sur bruit. En effet, parmi les VAD proposées, la segmentation s'appuyant sur le calcul de la STE et le ZCR demeure la plus simple et la plus utilisée, mais présente l'inconvénient de non-robustesse face au bruit. D'autres paramètres ont été définis. Nous retrouvons par exemple l'entropie : (Shen, Hung et Lee, 1998) ont montré que l'entropie d'un segment de parole diffère significativement d'un segment non-parole ; (Huang et Yang, 2000) se sont basés à la fois sur l'entropie et l'énergie. Les paramètres MFCC (Mel-Frequency Cepstrum Coefficients) ont été très largement utilisés dans la majorité des articles traitant le problème de la segmentation parole/musique, parole/non-parole, classification des genres musicaux et autres (Carey, Parris et Lloyd-Thomas, 1999; Gauvain, Lamel et Adda, 2002; Wang, Xu et Li, 2011; West et Cox, 2004; Williams et Ellis, 1999; Woodland et al., 1998) ont proposé une technique VAD basée sur la mesure de la distance entre un vecteur MFCC d'un segment de parole et celui d'un segment de bruit. Une autre approche de segmentation utilisant les paramètres fondés sur les « Ondelettes » a été proposée dans différentes études. L'avantage de l'utilisation de cette paramétrisation est qu'elle permet d'extraire à la fois les caractéristiques fréquentielles et temporelles du signal en fournissant ainsi une représentation compacte du spectre du signal. Les ondelettes sont plus robustes au bruit et à la non-stationnarité du signal et donnent des résultats plus efficaces en comparaison avec les MFCC (Alani et Deriche, 1999; Galka et Ziolko, 2008; Ziółko et al., 2006). Scheirer dans (Scheirer et Slaney, 1997) a présenté une collection de descripteurs connus pour leur bonne performance dans la discrimination parole/musique comme le ZCR, les moments spectraux, la modulation de l'énergie à 4 Hz ainsi que la variation de la magnitude spectrale, appelée aussi le flux spectral, qui est apte à distinguer les continuités harmoniques d'un signal.

La technique de décomposition en mode empirique (EMD) a été déployée récemment sur plusieurs types de signaux comme les signaux cardiaques, les signaux de paroles ainsi que d'autres signaux biomédicaux. L'EMD a montré ses capacités d'analyse sur des signaux non-stationnaires et des résultats prometteurs ont été obtenus (Charleston-Villalobos, Aljama-Corrales et Gonzalez-Camarena, 2006; He et al., 2011; Liu et al., 2010; Moukadem et al., 2011; Moukadem, Dieterlen et Brandt, 2012) .

Concernant le problème de la classification, de nombreuses approches ont été communément utilisées pour la classification ; les approches statistiques et les modèles probabilistes sont largement employés de nos jours dans les systèmes de reconnaissance et de segmentation de la parole. Ces approches, notamment celles basées sur les GMM et les HMM, ont atteint des performances notables et une robustesse au bruit remarquable (Carey, Parris et Lloyd-Thomas, 1999; J. Pinquier, 2002b; Rabiner et Juang, 1993a; Razik et al., 2003; Scheirer et Slaney, 1997; Zhang Tong, 1998). Plusieurs autres travaux se sont intéressés au problème de l'indexation audio, plus généralement, la classification de documents sonores. Une indexation basée sur une modélisation différentiée a été adaptée par Pinquier (Pinquier et Chambert, 2001); séparation de type classe/non-classe (parole/non-parole et musique/non-musique) en utilisant différents paramètres pour chaque type de classification. La phase de classification de cette approche a été fondée sur les GMMs et a donné des résultats satisfaisants.

Après ce tour d'horizon autour des revues bibliographiques en segmentation d'un signal audio, nous retenons cependant les points suivants :

- les paramètres MFCC sont les plus déployés dans la littérature et peuvent servir pour la manipulation d'un système de référence, le plus proche des systèmes de segmentation audio existants dans l'état de l'art ;

- les techniques de VAD, ZCR et STE, sont les plus importants, parmi les autres, pour les étapes de post ou pré traitement d'un système de segmentation ;

- les paramètres fondés sur les « ondelettes » et sur les EMD ont montré leur efficacité récemment dans le domaine de la reconnaissance des signaux audio. À notre connaissance, ces paramètres n'ont jamais été utilisés pour la tâche de segmentation d'un signal de cri en parties expiration et inspiration ;

- les approches HMM ont prouvé une robustesse au bruit remarquable pour la classification et la segmentation.

1.3 Architecture générale d'un système de segmentation

La détection des segments, expirations et inspirations, audibles, portant une information considérée pertinente sur l'état du nouveau-né est le but principal de notre étude. Ainsi, la distinction de ces deux composantes acoustiques des autres activités inutiles nécessite deux étapes principales : la paramétrisation et la classification du segment. Voir Figure 1.1. Une telle architecture correspond à celle de type « reconnaissance des formes ».

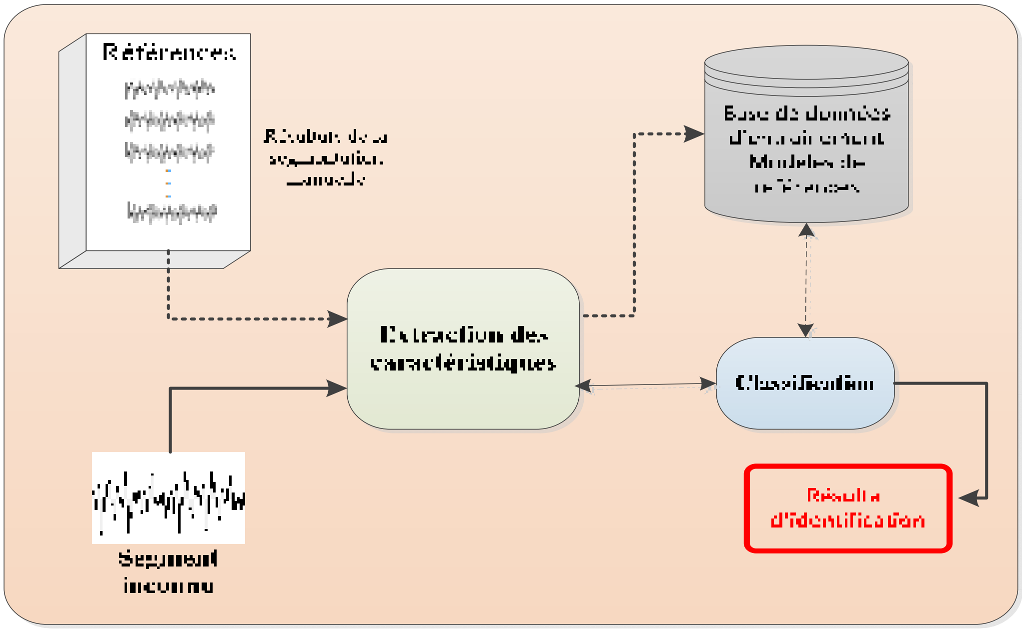

Références
Résultats de la segmentation manuelle
Base de données d'entraînement Modèles de références
Extraction des caractéristiques
Classification
Résultat d'identification
Segment inconnu

Figure 1.1 Architecture globale d'un système de segmentation automatique

La paramétrisation appelée aussi phase de l'extraction des caractéristiques consiste à extraire, d'un segment inconnu à l'entrée, les caractéristiques les plus discriminantes ou encore une information pertinente fournissant une description la plus représentative possible. Celle-ci permettra ensuite d'attribuer ce segment d'entrée à sa classe, dans notre cas : expiration, inspiration, background ou bruit.

1.4 Paramétrisation

La segmentation d'un signal audio de cris des nouveau-nés dans un contexte réel nécessite une description compacte et représentative des différentes activités acoustiques collectées en même temps que le cri. La phase de paramétrisation consiste à réduire la quantité d'information contenue dans le signal à traiter.

Dans ce contexte, dans les sections suivantes, un cadre scientifique est exposé en lien avec les divers descripteurs utilisés pour la segmentation dans différents contextes.

1.4.1 Le taux de passage par zéro (ZCR) et l'énergie à court terme (STE)

Il s'agit de deux paramètres qui se calculent à partir de la représentation temporelle du signal acoustique. Ces deux paramètres sont les plus fréquemment utilisés dans la phase de prétraitement d'un système de reconnaissance jusqu'à maintenant.

Le taux de passage par zéro est considéré un paramètre robuste pour une classification voisée (valeur ZCR élevée) / non voisée (valeur ZCR faible). Le ZCR représente la fréquence du passage de l'onde temporelle par l'axe d'amplitude nulle. Ainsi, il y a passage par zéro lorsque deux échantillons qui se suivent sont de signes opposés (Boite et Kunt, 1987).

Le calcul de ZCR se fait normalement pour un segment qui peut aller de 10 à 50 ms durant lequel le signal audio est supposé être quasi stationnaire :

$$ZCR(i) = \frac{1}{2N} \sum_{n=1}^{N} \operatorname{sgn}[x_n(i)] - \operatorname{sgn}[x_{n-1}(i)] \tag{1.1}$$

$$\text{Avec } \operatorname{sgn}[x(i)] = \begin{cases} 1, si\ x(n) \geq 0 \\ -1, si\ x(n) < 0 \end{cases} \tag{1.2}$$

$x_n(i)$ représente le nième échantillon du segment i et N le nombre d'échantillons dans ce même segment i.

Cet outil a été fréquemment utilisé pour une segmentation dans diverses applications. La méthode ne nécessite pas beaucoup de calculs et peut donner une idée sur le voisement du signal (Scheirer et Slaney, 1997; Zhang Tong, 1998) (Saunders, 1996), (Huiqun et O'Shaughnessy, 2007; Khaled, 2004).

L’énergie à court terme est l’un des outils qui donnent la représentation la plus fidèle de l’évolution temporelle du signal vocal (Khaled, 2004). Normalement, l’énergie est faible durant les périodes de silence et élevée en présence des activités vocales. Elle dépend de la catégorie et de l’amplitude du son émis.

L’énergie d’un signal échantillonné est définie par :

$$E\ (i) = \sum_{n=1}^{N} x_n^2(i) \tag{1.3}$$

Dans plusieurs travaux de segmentation et plus précisément dans la partie prétraitement du signal, le calcul de l’énergie et sa comparaison à un seuil bien défini était une étape préliminaire afin de détecter les segments de silences et ainsi ne traiter que les zones d’activités acoustiques (J. Pinquier, 2002b; Lu, Jiang et Zhang, 2001; Zhang Tong, 1998), (Wang, Gao et Ying, 2003).

Le travail de Saunders dans (Saunders, 1996) était le premier à montrer que le contour énergétique d’un signal acoustique est capable de distinguer entre les segments de parole et de musique. L’énergie subit des variations claires pour un signal de parole dues aux alternances voisées et non voisées alors qu’elle est généralement constante pour la musique.
Ces deux paramètres font toujours appel à un seuil qui dépend du contexte audio considéré ; ceci limite l’utilisation de ces deux paramètres.

1.4.2 L’entropie moyenne par trame

C’est une mesure qui représente le « degré du désordre » d’une observation (Papoulis, 1991). Ce paramètre a été employé dans plusieurs travaux sous différentes formes.

La mesure de l’entropie donne une information sur l’analogie entre le modèle observé et le modèle acoustique entraîné (Ajmera, McCowan et Bourlard, 2002), (Ajmera, McCowan et Bourlard, 2003).

Les probabilités a posteriori des phonèmes reconnus sont généralement élevées ; l’entropie moyenne pour des segments de parole (degré de désordre) est plus faible que pour des segments de la non-parole ou de musique (Ajmera, McCowan et Bourlard, 2003).

Pinquier, dans ses travaux (J. Pinquier, 2002a; J. Pinquier, 2002b), a employé un nouveau paramètre qui est la modulation de l’entropie au lieu de l’entropie moyenne. Cette mesure est obtenue à partir du calcul de la variance de l’entropie pendant une durée d’une seconde du signal. Au début, il a calculé l’entropie des segments de durée 16 ms et ainsi, dans des segments d’une seconde, il obtient 62 valeurs d’entropie dont il leur a calculé la variance pour avoir la modulation d’entropie. L’auteur a déduit que la modulation d’entropie aura une valeur plus élevée pour des segments de parole qui permet de les distinguer entre autres segments acoustiques.

1.4.3 Le centroide spectral

C’est le centre de gravité fréquentiel de la densité spectrale de puissance. Il peut être considéré comme le point d’équilibre du spectre (Didiot, 2007). Pinquier et Didiot ont noté que le centroide spectral a une valeur supérieure et constante pour la musique plutôt que pour la parole. En outre, l’alternance voisée/non voisée d’un signal de parole est caractérisée par une variation importante du centroide spectral (Pinquier, 2004).

1.4.4 Le point spectral de coupure

« *C’est le point en dessous duquel est contenue 95 % de la puissance du spectre* » (Razik, 2003). Ce paramètre permet de faire la discrimination voisée/non voisée d’un signal de parole. Pour les sons voisés, ce point a une valeur inférieure, car l’énergie est contenue dans les fréquences basses. Pour les sons non voisés, la valeur du point spectral de coupure est supérieure car une portion importante de l'énergie est concentrée dans les fréquences les plus élevées (Scheirer et Slaney, 1997).

1.4.5 La fréquence fondamentale

« *Elle correspond à la fréquence de vibration des cordes vocales* » (Boite et Kunt, 1987). Dans la littérature, plusieurs algorithmes de détection de la fréquence fondamentale ont été proposés. Zhang a mentionné que la valeur de la fréquence fondamentale donne une information sur le type du signal acoustique étudié en le comparant à une référence (Zhang Tong, 1998). Dans (G Várallyay Jr., 2008), l'auteur a utilisé la méthode HPS d'estimation de la fréquence fondamentale pour définir deux autres paramètres pertinents pour la classification des segments du cri pur d'un bébé et d'autres composantes du même signal.

1.4.6 Paramètres MFCC

Les coefficients cepstraux sur l'échelle de Mels (MFCC) sont les caractéristiques les plus utilisées dans le traitement d'un signal audio pour différentes applications : reconnaissance automatique de la parole, reconnaissance automatique des émotions, reconnaissance du locuteur, segmentation parole/musique, etc. (Carey, Parris et Lloyd-Thomas, 1999; Gauvain, Lamel et Adda, 2002; West et Cox, 2004; Williams et Ellis, 1999; Woodland et al., 1998). Le principe de calcul des coefficients MFCC provient de l'hypothèse selon laquelle le modèle de génération de la parole est le produit de la convolution d'une excitation (les cordes vocales) et d'un filtre impulsionnel (le conduit vocal) (Rabiner et Juang, 1993a). Une étape d'extraction des caractéristiques MFCC consiste à transformer le signal d'entrée en une suite de vecteurs d'observation riches en informations utiles pour une étape suivante de reconnaissance, de classification ou encore de segmentation (Figure 1.2). Les différentes étapes de calcul des coefficients MFCC sont représentées dans la Figure 1.3 (Rabiner et Juang, 1993a). Pour extraire les coefficients MFCC, le signal d'entrée est pré accentué afin de faire ressortir les hautes fréquences avec un coefficient de préaccentuation de 0.97.

Un signal de cri, comme tout signal vocal, est normalement un signal non-stationnaire. (Lederman, 2010; Saraswathy et al., 2012). Afin d'appliquer les techniques d'analyse et de paramétrisation déployées dans la littérature, il fallait respecter la notion de la stationnarité du signal audio à étudier. L'analyse doit donc être effectuée en utilisant des fenêtres sur des courts intervalles, pouvant aller de 5 à 50 ms (Huang, Acero et Hon, 2001), et qui subdivisent le signal audio en une suite de segments stationnaires.

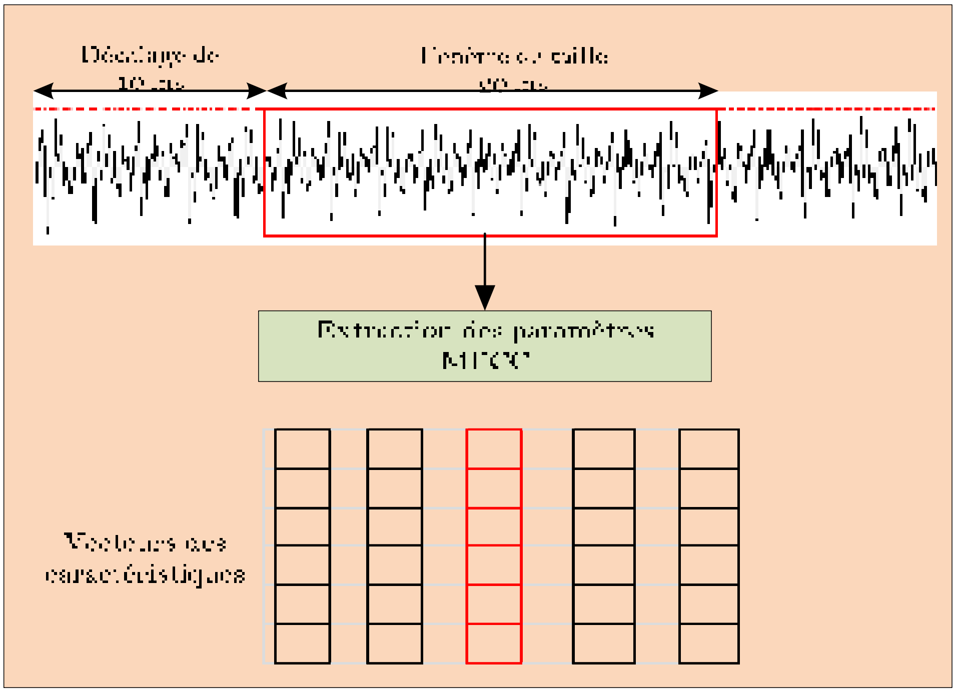


Figure 1.2 Extraction des caractéristiques MFCC

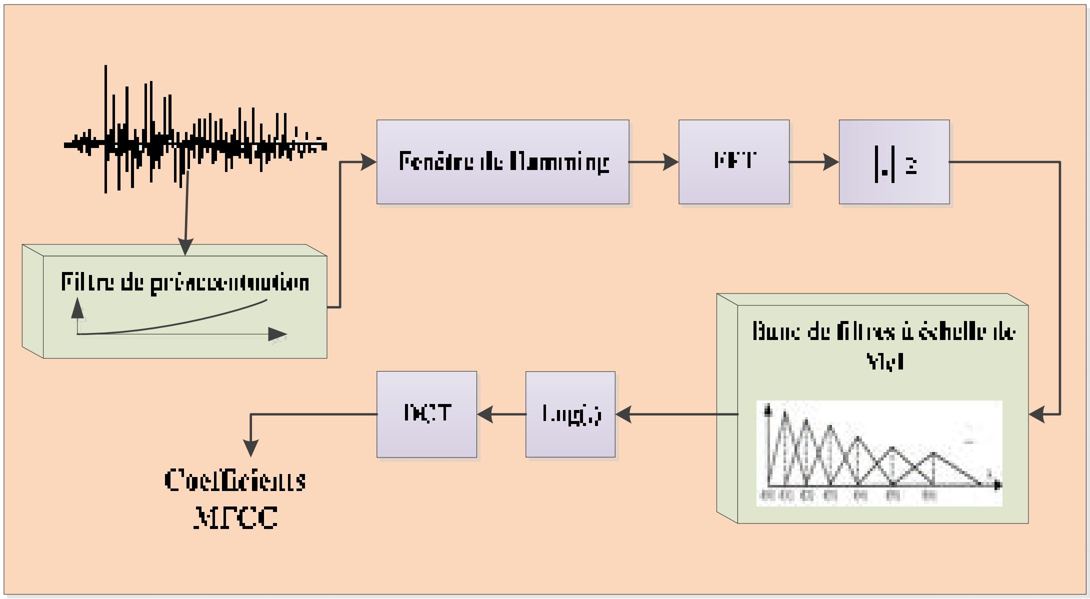


Figure 1.3 Étapes de calcul des coefficients MFCC
Adaptée de Rabiner et Juang (1993)

La fenêtre de Hamming est utilisée pour ce but, avec un recouvrement fixé à l'avance. Cette fenêtre permet de réduire le niveau de discontinuités au niveau des extrémités des segments (Farsaie Alaie et Tadj, 2012). La transformée de Fourier rapide est requise à cette étape pour convertir les segments d'entrée, fenêtrés par Hamming, du domaine temporel au domaine fréquentiel. Le module au carré de la FFT est appliqué à l'entrée d'un banc de filtres triangulaires sur l'échelle de Mel. Le choix de cette échelle provient du fait qu'elle peut simuler le mieux la perception de l'oreille humaine. L'échelle de Mel utilisée est définie par :

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{1.4}$$

Alors, pour extraire les coefficients MFCC, il s'agit d'un calcul direct de la transformée de cosinus discrète du logarithme du spectre de l'énergie du signal d'entrée correspondant à la sortie du banc des filtres.

1.4.7 Paramètres fondés sur les ondelettes

La transformée en ondelettes représente une autre technique d'analyse des signaux similaire à la transformée de Fourier à court terme (TFCT). À la différence de la TFCT, elle permet de contrôler à la fois les variables temps et fréquence d'un signal. Les ondelettes sont robustes à la non-stationnarité d'un signal acoustique (Didiot et al., 2010).

Les techniques fondées sur les ondelettes ont été utilisées avec succès dans les domaines du débruitage de la parole et de la reconnaissance automatique de la parole (Alani et Deriche, 1999; Didiot et al., 2010; Didiot et al., 2006; Galka et Ziolko, 2008; Ziółko et al., 2006). L'approche de décomposition en ondelettes consiste à décomposer un signal en coefficients d'approximation et coefficients de détail. Les coefficients d'approximation représentent les moyennes locales du signal alors que les coefficients de détail, dit coefficients d'ondelettes, correspondent aux différences entre deux moyennes locales. L'utilisation de l'ondulation « dyadique » demeure la plus fréquente, du fait qu'elle peut simuler le mieux l'oreille humaine en fournissant ainsi une approximation à l'échelle de Mel (Didiot et al., 2010; Didiot et al., 2006).

La transformée en ondelette dyadique est définie par l'équation suivante :

$$W(k,j) = \sum_{j}\sum_{k} x(n) 2^{\frac{-j}{2}} \Psi(2^{-j} n - k) \tag{1.5}$$

Avec x(n) le signal à l'instant n et Ψ () la fonction appelée ondelette mère.

En effet, nous nous sommes intéressés à l'utilisation des coefficients d'ondelettes qui serviront comme vecteurs de caractéristiques dans la phase de paramétrisation d'un signal acoustique. Dans chaque bande de fréquence, différents paramètres d'énergie pourront être calculés à partir des coefficients d'ondelettes. Nous citons par exemple :

- le logarithme de l'énergie instantanée :

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=1}^{N_j}\left(w_k^j\right)^2\right) \tag{1.6}$$

- les coefficients MFDWC (Mel-Frequency Discrete Wavelet Coefficients) qui sont définis de la même manière que les coefficients MFCC décrits ci-dessus. Il s'agit d'un calcul direct de la transformée en cosinus discrète du logarithme du spectre de l'énergie calculée à partir des coefficients d'ondelettes.

Il existe plusieurs familles d'ondelettes dont les plus connues dans le domaine de la reconnaissance de la parole sont : Daubechie, Symlet et Coiflet (Gemello et al., 2001).

1.5 Classification

Une fois le choix des paramètres acoustiques qui contribuent à donner l'information jugée pertinente sur le contenu du segment effectué, la question qui se pose concerne le pouvoir de discrimination entre les différentes classes acoustiques. Il faut alors choisir un classifieur qui permettra d'effectuer la classification du signal étudié en segments catégorisés suivant les classes que nous voulons distinguer, notamment : expiration, inspiration, parole, « bip » et bruit.

Deux stratégies fondamentales de classification ont été employées fréquemment dans les travaux récents : classification non supervisée et classification supervisée. Nous nous intéressons ici à la classification supervisée.

En effet, l'usage des approches de classification supervisée, basées sur les modèles statistiques, issus du domaine de la reconnaissance de la parole, intervient dans la plupart des domaines du traitement automatique d'un signal audio. Ces approches nécessitent deux phases : apprentissage et test. Elles sont basées sur le calcul de la probabilité d'appartenance à une classe acoustique connue d'avance. Ainsi, la probabilité qu'un segment de test appartenant à une classe acoustique sera calculée et le segment test sera attribué à la classe la plus probable.

L’apprentissage de tels modèles nécessite un grand nombre d’observations, ce qui implique la mise en place d’un corpus d’entrainement de grande taille (Pinquier, 2004). La classification supervisée s’applique alors dans le cas où les vecteurs d’entrées sont arrangés selon des classes connues à l’avance (Ramona, 2010).

1.5.1 Les mélanges de modèles gaussiens GMM

Le principe de base des GMM est de modéliser les fonctions de densité de probabilité des données observées en se basant sur la somme pondérée de N lois gaussiennes puisqu’une seule loi gaussienne n’est pas suffisante.

Une phase d’apprentissage est obligatoirement indispensable pour fixer les paramètres des GMM basés sur l’algorithme EM (expectation maximisation) dont le principe est d’estimer et de maximiser la vraisemblance des vecteurs d’apprentissage aux vecteurs observés de manière itérative jusqu’à un point de stabilité.

La phase de reconnaissance est basée sur le principe du maximum de vraisemblance. Ainsi, chaque nouveau vecteur d’entrée est attribué à une classe selon la probabilité la plus élevée (Didiot, 2007; Ramona, 2010; Razik, 2003).

La grande capacité de modélisation des GMM leur permet d’être « des approximateurs universels » qui sont fréquemment utilisés dans le domaine de la reconnaissance de la parole plus précisément dans le sujet de la segmentation des signaux acoustiques (Saunders, 1996; Scheirer et Slaney, 1997).

1.5.2 Les modèles de Markov cachés

Les modèles de Markov cachés sont apparus dans la problématique de la reconnaissance de la parole depuis plusieurs décennies. Ils ont prouvé leur efficacité dans de nombreux domaines d’applications.

« Un modèle de Markov est un automate probabiliste qui permet de donner une représentation statistique d'un évènement ». (Razik, 2003)

Il est composé d'un nombre d'états et de transitions. Il se base sur l'hypothèse de Markov selon laquelle un état futur ne dépend que de l'état présent. Les Modèles de Markov cachés sont une extension des chaînes de Markov. Ces modèles se basent sur un processus stochastique double. En effet, la séquence des états du processus ne peut pas être directement observée; la séquence est cachée. Chaque état émet des observations. Nous nous servons alors de la séquence d'observations générée (Juang et Rabiner, 1991; Young et Evermann, 1996).

Un HMM pourra être utilisé comme un générateur d'une suite d'observation. La notation λ est souvent employée pour $O = O_1, O_2, \ldots, O_T$ désigner l'ensemble constitué des probabilités initiales, des probabilités de transition et des probabilités d'émission . La Figure 1.4 représente un HMM à états cachés. $\lambda = (A, B, \pi)$

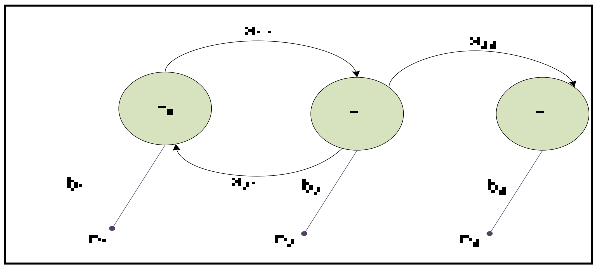


Figure 1.4 Chaine de Markov caché à 3 états

Les HMM ont été largement déployés pour la reconnaissance de la parole et ont prouvé leur efficacité et leur précision. Ceci a permis de développer des modèles de plus en plus précis avec le temps.

CHAPITRE 2

AUTOMATIC DETECTION OF THE EXPIRATORY AND INSPIRATORY PHASES IN NEWBORN CRY SIGNALS

L. Abou-Abbas a, H. Fersaie Alaie b and C Tadj [c]

[a,b,c] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montréal, QC, Canada H3C1K3

This paper is published
in *Biomedical Signal Processing and Control*, volume 19, p. 35-42

Highlights

- The paper presents a segmentation system of newborn cry signals based on Hidden Markov Models;
- The proposed system is able to detect the audible expiratory and inspiratory parts from other acoustic activities;
- Experimental results show the effect of certain parameters on the performance of the proposed system.

2.1 Abstract

An analysis of newborn cry signals, either for the early diagnosis of neonatal health problems or to determine the category of a cry (e.g., pain, discomfort, birth cry, and fear), requires a primary and preliminary preprocessing step to quantify the important expiratory and inspiratory parts of the audio recordings of newborn cries. Data typically contain clean cries interspersed with sections of other sounds (generally, the sounds of speech, noise, or medical equipment) or silence. The purpose of signal segmentation is to differentiate the important acoustic parts of the cry recordings from the unimportant acoustic activities that compose the audio signals. This paper reports on our research to establish an automatic segmentation system for newborn cry recordings based on Hidden Markov Models using the HTK (Hidden Markov

Model Toolkit). The system presented in this report is able to detect the two basic constituents of a cry, which are the audible expiratory and inspiratory parts, using a two-stage recognition architecture. The system is trained and tested on a real database collected from normal and pathological newborns. The experimental results indicate that the system yields accuracies of up to 83.79%.

Keywords: HMM; Automatic segmentation; Newborn cry signals; Mel Frequency Cepstral coefficients; Viterbi algorithm; Baum Welch algorithm.

2.2 Introduction

With early newborn screening, a serious illness can be diagnosed such that treatment can begin before severe problems appear, and in certain cases, sudden mortality or disability can be prevented. Clearly, the presence of disease must be detected at an early stage. Systematic screening combined with better diagnostic tools is therefore required to meet future medical challenges, with the aim of supporting clinical decision-making and improving the effectiveness of treatment (Orlandi et al., 2012). These tools have evolved considerably in recent years in terms of improving screening and symptom evaluation, and the newborn cry signal has been the object of strong research interest for the past three decades.

Researchers have amassed enough evidence to conclude that a cry signal contains relevant information on the psychological and physiological condition of the newborn; formal relationships have been established between the acoustic features extracted from the cries and the health problems of the child (Golub, 1979; Proctor, 1984; Rui et al., 2010; Várallyay, 2006). Various studies are currently under way to devise a tool that analyzes cries automatically, to diagnose neonatal pathologies (Farsaie Alaie et Tadj, 2012; Fort et Manfredi, 1998; Hariharan et al., 2012).

We are involved in the design of an automatic system for early diagnosis, called the Newborn Cry-based Diagnostic System (NCDS), which can detect certain pathologies in newborns at an early stage. The implementation of this system requires a database containing hundreds of cry signals.

The overwhelming problem that arises when working with such a database is the diversity of acoustic activities that compose the audio recordings, such as background noise, speech, the sound of medical equipment and silence. Such diversity could harm the analysis process, as the presence of any acoustic component other than the cry itself could result in the misclassification of pathologies by reducing the NCDS system performance. This is because the NCDS would decode every segment of the recording signal, whether it is part of a cry or not. In this case, unwanted segment insertion in essential crying segments would lengthen the process of classification unnecessarily and leave the system prone to error. An important subtask of the NCDS is the manipulation of the newborn cry sound, and what is needed to perform this subtask is a segmentation system. Until now, few works have been carried out in this area. In this paper, we propose an automatic segmentation module designed to isolate the audible expiration and inspiration parts of cry sounds to serve as a preprocessing step of our NCDS.

The rest of this paper is organized as follows: Related work is presented in section 2.3. The HMM and the HTK are reviewed briefly in section 2.4. The training corpus and the testing corpus are described in section 2.5. In section 2.6, the architecture of the proposed system is presented, and details of the individual blocks are described in five subsections. Section 2.7 contains the implementation of the system, the obtained results, and the discussion. Finally our conclusions are presented in section 2.8.

2.3 Related Work

Several studies have been conducted in which the infant cry is analyzed (categorization of the cry, disease classification based on the cry). In 1985, for example, Corwin and Golub outlined four acoustic categories composing a cry episode, which are: (a) expiratory phonation (with F0 ranging from 250-750 Hz), (b) expiratory hyperphonation (with F0 ranging from 1000-2000 Hz), (c) expiratory dysphonation (aperiodic expiratory segment); (d) inspiratory phonation (associated with any perceptually audible sound generated by the newborn during inspiration, or high-pitched cries during inspiration) (Golub, 1979; Grau, Robb et Cacace, 1995).

In most studies, the cry segmentation phase was performed manually; a human operator was asked to monitor the recorded audio signals and pick out only the important cry parts from the recordings (Michelsson et Michelsson, 1999; Proctor, 1984; Wermke et al., 2002). This manual task is tiresome and too time-consuming when the volume of data is large. The cry segmentation that serves the needs of a real-time diagnostic tool should be performed automatically.

In some studies, the authors have applied various voice activity detection software approaches such as the traditional methods of ZCR (Zero Crossing Rate) and STE (Short Time Energy), with some modification of the thresholds (Kuo, 2010; Rui et al., 2010; Várallyay, 2006; Zabidi et al., 2009b). In general, these methods are of limited use in this context, as speech and cry sounds have different features. With these methods, particularly in the search for the high-energy parts of the audio signals, not only are the meaningful parts of cry vocalizations found but also background noise, speech, and machine sounds. In other words, the typical voice activity detection methods alone are not suitable for segmenting a cry signal. The corpuses used to examine these methods (ZCR, STE) were composed only of cry sounds, which are sequences of expiration and inspiration, alternating with short periods of silence and background noise. The main goal of authors was to eliminate silence and background noise without affecting the audible expiration and inspiration phases.

Few studies have been conducted specifically on the automatic segmentation of cry signals (Aucouturier et al., 2011; Várallyay, Illényi et Benyó, 2009; Várallyay Jr, Illényi et Benyó, 2008). Two novel algorithms were introduced by modifying the Harmonic Product Spectrum (HPS) method (Várallyay, Illényi et Benyó, 2009). The HPS method was created to detect the fundamental frequency of an audio signal. The authors showed that it is possible to check the regularity structure of the spectrum using the HPS method and classify its content by detecting the meaningful parts of the cry sounds. Another study on the segmentation of cry signals was conducted in 2012 with the purpose of labeling each successive segment as a cry/non-cry/non-activity. However, with the methods presented, the inspiration parts as well as the dysphonic vocalizations of the cry spectrum that could be presented with irregular or non-harmonic structure were ignored.

Recent studies have shown that differentiated characteristics in expiratory and inspiratory vocalizations exist in adults as well in newborns (Orlikoff, Baken et Kraus, 1997).

Assuming that the inspiratory phase of a cry episode reflects a laryngeal contraction of the ingressive airstream, inspiratory vocalization has been proven to be useful in the identification of newborns at risk for various health conditions (Grau, Robb et Cacace, 1995). In fact, the amount of time the inspiratory phase lasts in newborns with respiratory disorders is greater than it is in normal newborns (Verduzco-Mendoza et al., 2012). Indeed, recent medical evidence confirms that a relationship exists between upper airway obstruction and sudden infant death syndrome and sleep apnea. Despite this evidence, it is surprising to find acoustic data that are limited to the expiratory phase alone (Grau, Robb et Cacace, 1995).

To create an effective diagnostic tool based on the cry signals, the involvement of both the expiratory and the inspiratory components is a prerequisite. The aim of this study is to identify and quantify both the audible inspiratory and expiratory components of a newborn cry automatically.

The work presented in this paper is based on the well-established and widely used Hidden Markov Model (HMM) statistical technique, which has been successfully applied in automatic speech recognition and segmentation systems.

To the best of our knowledge, no work has yet been carried out on the automatic segmentation of crying signals recorded in noisy environments without manually pre-processing the signals to remove at least irrelevant acoustic activities, such as speech and beep sounds around the infant.

In recent work (Aucouturier et al., 2011), authors applied an automatic segmentation approach based on a HMM classification tool to segment the expiratory and inspiratory sounds of cry signals. The difference with this recent approach compared to our approach is not only with the limited number of infants and the limited available acoustic activities types (due to the environment in which recordings are taking place) but also the way in which they applied the HMM. The authors of (Aucouturier et al., 2011) considered only three classes, Expiration (EX), Inspiration (IN) and Silence (SI). As a first stage, to train each class, they used different techniques such as Support Vector Machines (SVM) as well as Gaussian Mixture Models (GMM) consisting of 5 and 20 Gaussian components. To reduce errors by taking into account the arrangement in time between the three classes, the authors added a second stage using the Viterbi algorithm. The whole architecture of the approach in (Aucouturier et al., 2011) could be taken as an HMM architecture of three states. In fact, the segmentation approach presented in (Aucouturier et al., 2011) performed well, but its performance needs to be enhanced to segment audio signals recorded in a noisy environment (e.g., sounds of speech, medical equipment, noise, and silence).

To provide a better understanding of the context of this study, some important terms used must be predefined.

1. Inspiration is associated with inspiratory phonation as defined by Golub and Corwin (Golub, 1979);
2. Expiration is referred to the acoustic output during the expiration phase of a cry (it can be phonation, dysphonation, or hyperphonation), as well as any audible expiration sound generated by the infant outside its cry episodes. Note that we do not make a distinction here between the expiration phases that occur during or following a cry;

3. A cry sequence consists of long periods of expiratory crying separated by short inspiratory episodes.

We have avoided using the terms voiced inspiration and voiced expiration to describe the important parts of the cry. In fact, a dysphonation vocalization is characterized in earlier studies as an unvoiced part during a cry and it is considered one of the most useful vocalizations in the detection of newborns at risk of various health conditions (LaGasse, Neal et Lester, 2005). For this reason, we prefer using the terms audible inspiration and audible expiration.

2.4 Hidden Markov Model and the HTK

HMM underlie the most modern automatic speech recognition (ASR) systems. They have many potential applications in statistical signal processing and acoustic modeling, including the segmentation of recorded signals (Young et Evermann, 1996). The basic principles of any ASR system involve constructing and manipulating a series of statistical models that represent the various acoustic activities of the sounds to be recognized (Young et Evermann, 1996). Many studies have shown that speech, music, newborn cries, and other sounds can be represented as a sequence of feature vectors (temporal, spectral, or both), and HMMs could provide a very important and effective framework for building and implementing time-varying spectral vector sequences (Gales et Young, 2008b).

An HMM generates a sequence of observations $O=O_1$, O_2, …,O_T and is defined by the following parameters: number of hidden states, state transition probability distribution A, observation probability distribution B, and initial state distribution π. We denote the model parameters of the HMM as $\lambda= \{A, B, \pi\}$

(Kuo, 2010; Várallyay, 2005). These concepts are depicted in Figure 2.1.

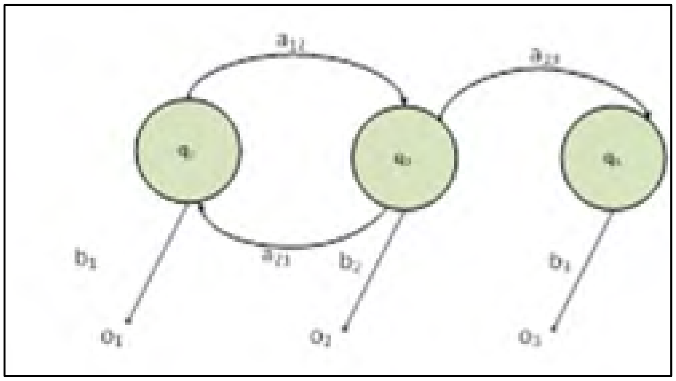


Figure 2.1 Hidden Markov model topology

To build and manipulate an HMM, three problems must be solved: the evaluation problem, the decoding problem, and the training problem. HMM theory, the aforementioned problems, and the proposed solutions are widely explained in the literature, especially in the well-known Rabiner tutorial.(Rabiner, 1989) The Viterbi algorithm is proposed as a decoding solution to find the most probable future state of the system based on its current state (Rabiner, 1989). The Baum Welch algorithm is an iterative procedure used to estimate the HMM parameters.

The HTK is the Hidden Markov Model Toolkit developed by Steve Young in the Cambridge University Engineering Department (CEUD). This toolkit is designed to build and manipulate HMMs using training observations from a sound corpus to decode unknown observations. It consists of a set of library modules and tools available in the C source code. Although the use of the HTK has been limited to speech recognition research, it is flexible enough to support the development of various HMM systems (Young et Evermann, 1996).

2.5 Materials

To develop our targeted diagnostic system (NCDS), we are working on building a very large corpus of cry signals. The recordings were made in the neonatology departments of several hospitals in Canada and Lebanon. The infants that were selected for the recording procedures are from 1 to 53 days old and were both preterm and full term. The database includes both healthy and sick babies and both males and females. The average duration of the newborn cry records is 90 seconds. The medical staff was put in charge of the following tasks: determining the type of cry being recorded, such as pain, hunger, diaper change and birth cry, writing down the date and time of the cry recording, and any useful information available about the babies (date of birth, gender, maturity, race, ethnicity, gestation, and known diseases). Three recording files were collected from the majority of the babies. All the recordings were acquired with an Olympus hand-held digital 2-channel recorder at a sampling frequency of 44.1 kHz and a sample resolution of 16 bits. The recorder was placed 10 to 30 cm from the newborn's mouth to be effective. The recorded audio signals are registered as WAV files. The newborns that were selected for the global NCDS project suffer from various pathological conditions.

The group of abnormal infants represents various types of serious conditions and diseases, chief among them being:

- diseases affecting the central nervous system (cerebral hemorrhage, meningitis, sepsis);
- blood disorders (anemia, hyperbilirubinemia, hemolytic disease, and hypoglycemia);
- congenital cardiac anomaly (ventricular septal defect, atrial septal defect, complex cardiovascular cases);
- diseases in which the respiratory system is directly involved (asphyxia, respiratory distress syndrome, apnea, Bronchopulmonary dysplasia, and pneumonitis);
- chromosomal abnormality.

Thus far the corpus collected includes infants' cries in different recording environments and conditions, from silent to very noisy. The background noises may be of many types, such as human speech (nurses, doctors, parents), the sounds of the recording device and medical equipment in the neonatal Intensive Care Unit (the beeping of machines), and the sounds made of doors opening and closing and running water. To build our segmentation module, we used signals produced by 64 newborns, including both normal and abnormal, for a total of 151 cry signals.

The total duration of the recordings in the training corpus and the testing corpus used here is 21900 seconds: 4 hours and 11 minutes are devoted to the training samples, and 1 hour and 54 minutes for the testing samples. It is important here to note that the babies chosen for the training phase are different from those chosen for the testing phase. See Table 2.1.

Table 2.1 Corpus statistics

			Number of babies	**Number of signals**
Male	Full term	Healthy	3	8
		Pathological	3	7
	Preterm	Healthy	3	6
		Pathological	4	7
Female	Full term	Healthy	22	50
		Pathological	12	31
	Preterm	Healthy	7	16
		Pathological	10	26
Total			**64**	**151**

This content of this database is unique and realistic. It contains long cry sequences (expiration phases alternating with short periods of inspiration episodes). The cry signals in both the training corpus and the testing corpus have been manually indexed – see Table 2.2 for details.

Table 2.2 Data used for training and testing corpuses

	Training Corpus in min	**Testing Corpus in min**	**Total Tim in min**
EX-Expiration	72.1	38.7	110.8
IN-Inspiration	8.1	3.6	11.7
SP-Speech	8.6	6.3	14.9
BIP- Beeping of machines	2.8	1.8	4.6
SI-Silence	58.4	29.5	87.9
NS-Noise	27.7	14.7	42.4

2.6 System Overview

In this paper, we introduce a new approach for high performance cry signal segmentation on realistic tasks related to our automatic Newborn Cry-based Diagnostic System (NCDS). We use a Cygwin interface, which enables HTK commands to be executed on the Windows platform. This segmentation tool is designed mainly to automatically segment and label expiration and inspiration phases taken from the audio recordings of newborn cries, in an attempt to find the best ways to resolve these two major issues:

1. Select the best parameters of the MFCC extraction procedure such as the number of MFCCs and the window size;

2. Design a robust classifier that can perform accurate segmentation.

Cry signal segmentation using the HMM approach can be viewed as similar in its implementation to cry signal classification problems (classification of diseases or types of cry).

.

To implement the HMM efficiently, we used a Cygwin interface and the HTK. The cry segmentation system architecture consists of the following modules: data preparation, feature extraction, training, and recognition. A block diagram of the proposed system is shown in Figure 2.2. Individual blocks are described in the following subsections.

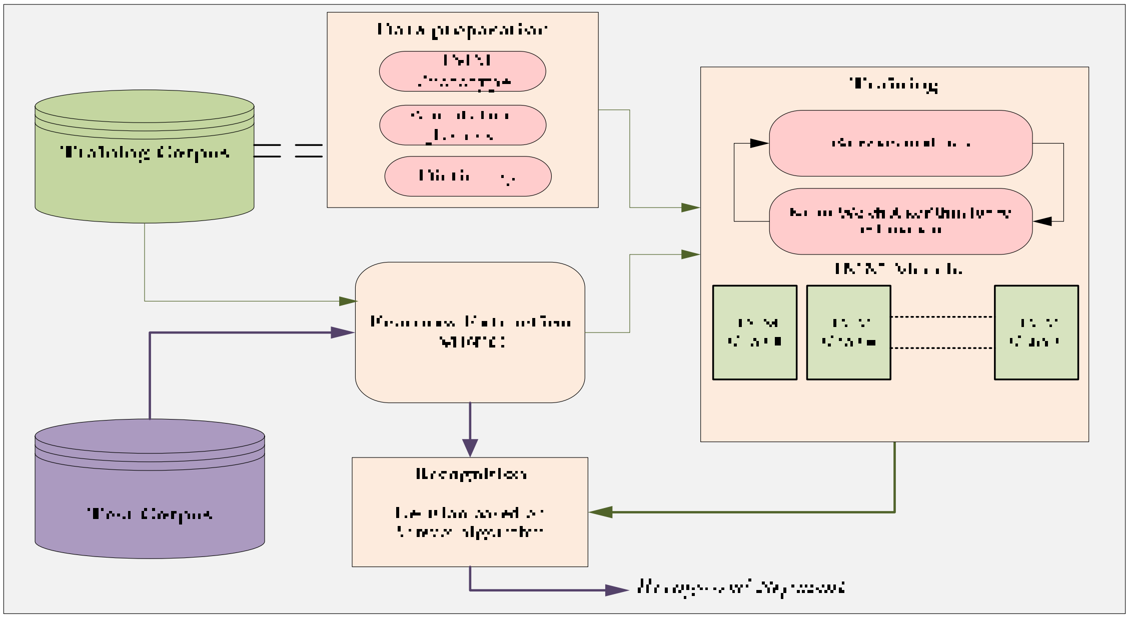


Figure 2.2 Automatic infant cry segmentation system architecture

2.7 Data preparation

The first step in any recognizer tool is data preparation, as data are needed for both training and testing. This phase consists of recording and labeling the cry signals. The recording task was discussed in the Materials section. Labeling is required so that the HMM models can be trained and the results of the proposed automatic segmentation system can be tested. For this task, which is performed manually, we used the Wave Surfer software (Sjölander et Beskow, 2000). Figure 2.3 is an example of manual segmentation. For the labeling task, a task dictionary in text format is required to describe the correspondence between the name of the class and the label. These text files are saved in the .lab format. We chose to build 6 HMMs, according to the various sound activities recorded, as follows:

- expiration class (EX): composed of the acoustic activities of the baby during expiration episodes;

- inspiration class (IN): composed of the vocal activity of the baby during the inspiration phase;

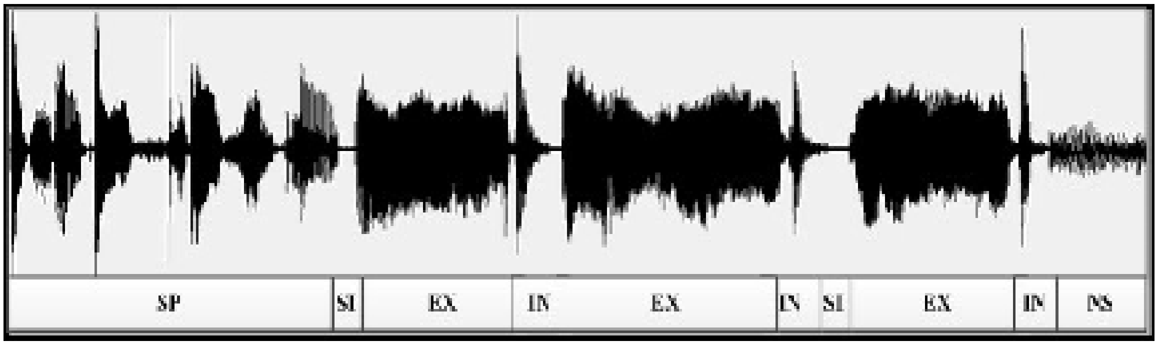


Figure 2.3 Example of manual segmentation using the Wavesurfer tool

- noise class (NS): composed of the sounds produced by the recording device and background sounds;

- speech class (SP): composed of the sounds made by speakers within the recording area;
- silence class (SI): composed of periods of lack of sound;
- bip Class (BIP): the sounds made by medical equipment, characterized by uniform energy.

2.8 Feature Extraction

The cry signals captured by the recorder are fed into the Feature Extraction module. At this stage, the input signal is first converted into a series of acoustic vectors, which are then be forwarded to the Recognition (decoding) phase (Rabiner et Juang, 1993a). In the feature extraction phase, the audio signals are expressed in spectral form by converting the raw audio signal into a sequence of acoustic feature vectors that carry acoustic information about the signal. MFCC (Mel Frequency Cepstral Coefficients) is one of the most efficient and widely used parameterization techniques used to produce the feature vectors. The audio waveforms, sampled at 44.1 KHz, are treated with a pre-emphasis coefficient of 0.97. The Hamming window is then applied, and then the Fourier Transform is calculated for each frame.
The obtained power spectrum is fed to the Mel-scale filter banks (24 channels) to yield more low frequency details, and the Mel coefficients are generated by applying the Discrete Cosine Transformation (DCT) to the log spectral representation (see Figure 2.4). MFCC can be computed using the following formula [26]:

$$MFCC(l) = \frac{1}{M}\sum_{i=0}^{M-1} \log(E(i)).\cos(\frac{2\pi}{M}(i+\frac{1}{2})l) \tag{2.1}$$

Where M represents the number of Mel filters, and l =0… M-1.

The feature vector that represents the distinctive properties of the audio signals is designed to be up to 39 MFCC parameters in length, consisting of 12 Cepstral coefficients and an energy component, along with their dynamic and acceleration coefficients (Δ and ΔΔ). The acoustic vector files (.mfcc) obtained in this step will be used in both the training and recognition phases of the system, and the extracted MFCCs will be useful for the NCDS processing phases, as no additional computation is required to extract the features. This makes our proposed NCDS a feasible and timesaving segmentation system.

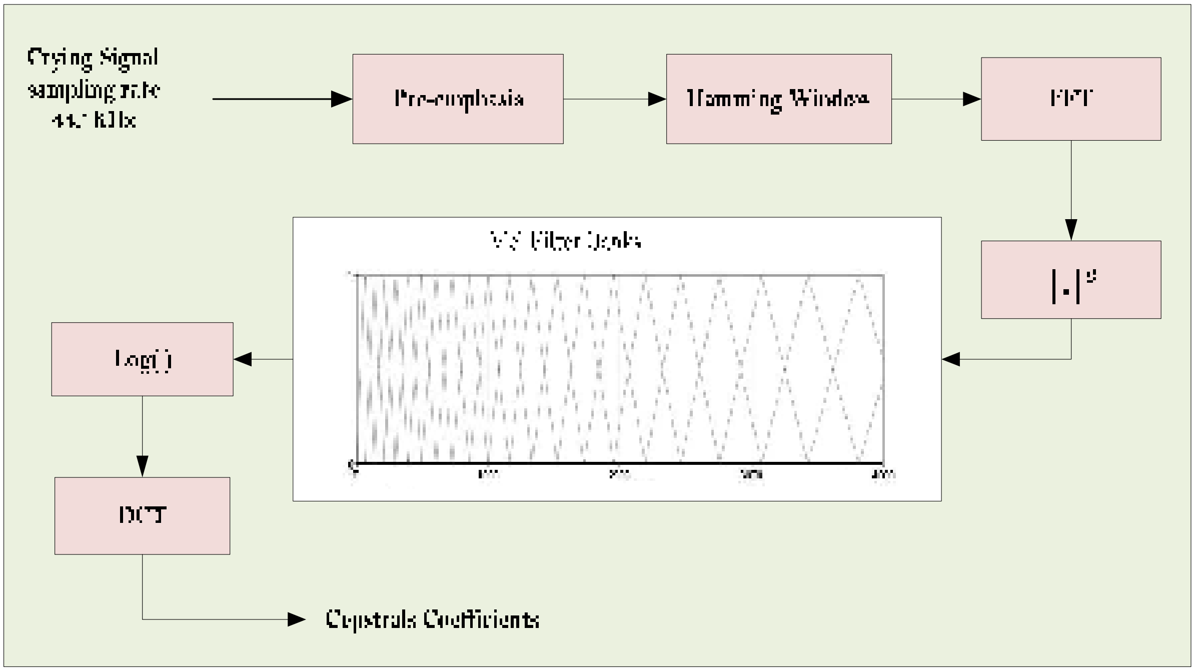


Figure 2.4 Extraction Mel frequency cepstral coefficients (MFCC) from the audio recording signals

2.9 Training

Our segmentation system consists of 6 classes, corresponding to the six types of sounds composing the audio recordings: expiration (EX), inspiration (IN), BIP sounds, Silence (SI), and noise (NS). The main goal in this step is to establish a consistent pattern representation for each class or label, which is also called a statistical model or HMM model. An HMM model is defined at the training stage as a reference model. This is because, during testing, a direct comparison should be made between the unknown label and each HMM-trained model to determine the most probable identity of this unknown label. A training phase for modeling all the acoustic activities is therefore essential.

For comparison purposes, we used 4-, 5-, and up to 8-state HMMs, in which the first and last states are non-emitting states. Moreover, multiple Gaussian distributions, varying from 1 to 100, with diagonal matrices are used and described by mean and variance vectors. The optimal values of the HMM parameters, the transition probability, mean, and variance of each observation, are estimated iteratively at this point. This is also called re-estimation, as the procedure is repeated many times for each HMM until convergence is reached. The Baum-Welch algorithm is used to estimate and re-estimate the mean and covariance of the each model.

2.10 Recognition stage

After the training stage has been completed, a different HMM is trained for each of the six classes. At the recognition stage, the trained HMMs are used to generate a set of transcriptions for unknown observations. Therefore, given an unknown observation or unknown segment of an audio recording, the unknown observation is converted into a series of feature vectors (.mfcc), in the same way as in data training. See Figure 2.5.

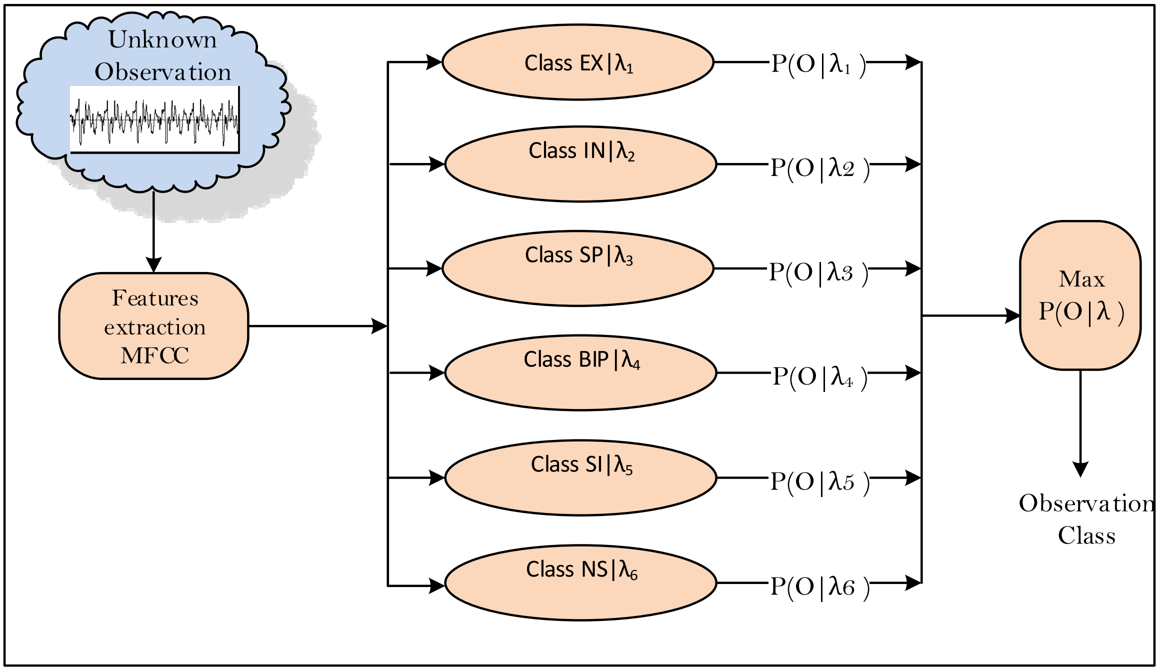


Figure 2.5 Recognition stage

This acoustic information, along with the reference HMM models, dictionary, and class names, are taken as input to create output in a label (.lab) file. Each HMM class produces a posterior probability estimate, and the HMM with the maximum probability is chosen as the most likely class. This phase is performed using the Viterbi algorithm (Rabiner, 1989).

2.11 Evaluation of system performance

The performance of the automatic segmentation system, such as that of any recognition system, is usually analyzed in terms of accuracy and speed. In this step, the output transcription file results are compared to the manually annotated files (reference labels N). Therefore, the comparison is made on a label-by-label basis by matching each of the recognized and reference label sequences.

Three possible types of errors should be calculated: number of insertions (I), deletions (D), and substitutions (S). In Figures 2.6, 2.7 and 2.8, we can see the difference between the three types of errors:

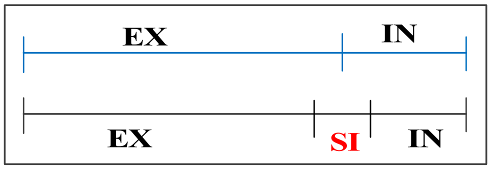


Figure 2.6 Example of an insertion error- reference labels are marked in blue and automatic labeling in black- insertion error in red

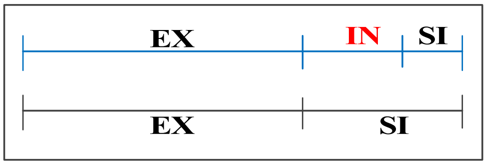


Figure 2.7 Example of a deletion error- manually annotated labels in blue- automatic labeling in black and deletion error in red

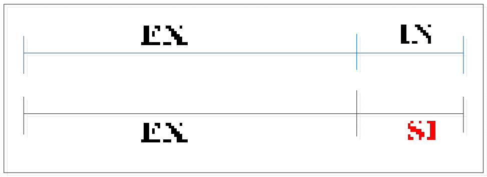


Figure 2.8 Example of a substitution error- manually annotated labels in blue- automatic labeling in black and deletion error in red

Once the alignment between the automatic transcription file and the reference file is done and both of three types of errors (D, I and S) mentioned above are calculated, the accuracy rate of the system could be estimated as follows: Accuracy Rate AR:

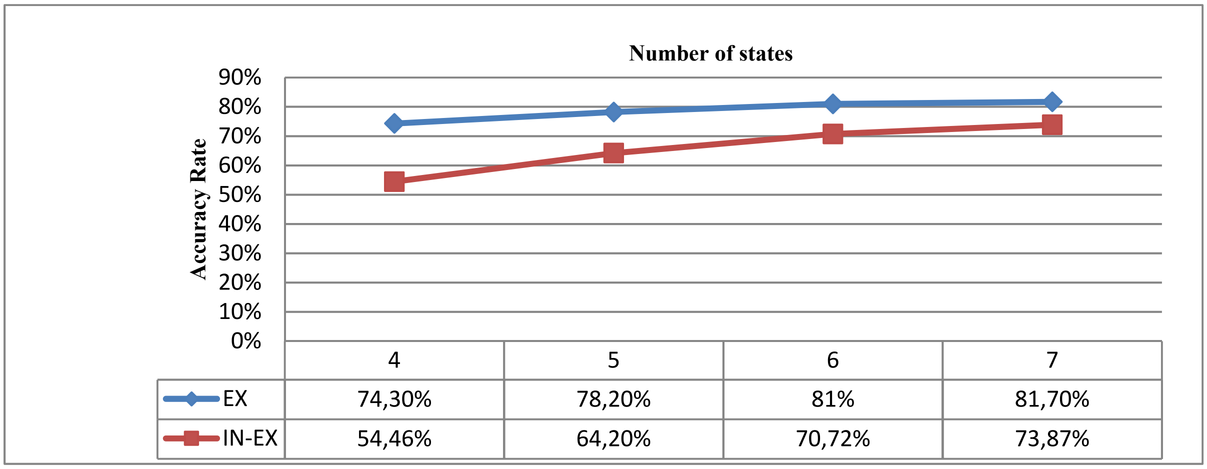
$$AR = \frac{N - D - I - S}{N} \times 100\% \tag{2.2}$$

Where N represents the total number of labels in the reference transcription files, D the number of deletions, S the number of substitutions, and I the number of insertions.

2.12 Results and Discussion

The basic function of this system is to differentiate the audible expiratory and inspiratory parts of newborn cries in audio recordings. The performance of this segmentation system has been evaluated on the testing corpus using a manually segmented cry corpus. We then measured the discrepancies between manual segmentation and automatic segmentation.

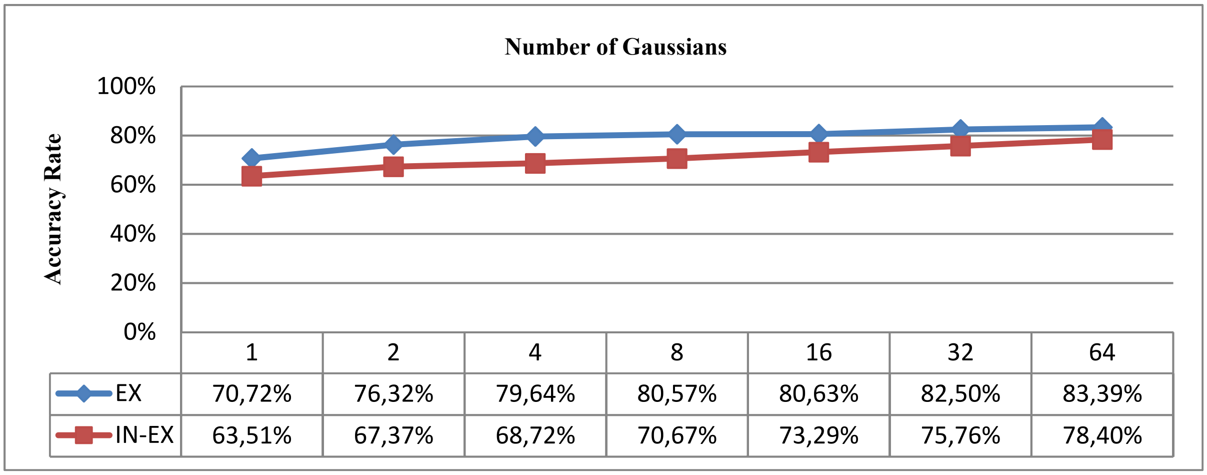
Various training strategies were evaluated, to select the most suitable reference system, and some important observations were made concerning the best segmentation performance. Here, we summarize the experiments that were carried out. The first step in the manipulation of our automatic newborn cry segmentation system is the selection of the best HMM model topology. A large number of experiments were performed to determine the optimal number of states and the number of mixtures for the training data. Figure 2.9 and Figure 2.10 illustrate the classification accuracy that results when these numbers are varied.



	4	5	6	7
EX	74,30%	78,20%	81%	81,70%
IN-EX	54,46%	64,20%	70,72%	73,87%

Figure 2.9 HMM classification results obtained with the various model topologies and a fixed window size of 30 ms and MFCC_E_D_A_Z (39 parameters)

It is essential to recall here that the main goal of this research is the detection and differentiation of the expiratory and inspiratory parts of newborn cries from cry recordings. To achieve this goal, we focus now on the accuracy rate of these two classes. In Figure 2.9, we show the accuracy rates obtained using HMM topologies varying from 4 to 7 states, with a fixed-length window of 30 ms with a 50% overlap and 39 dimensional feature vectors.



	1	2	4	8	16	32	64
EX	70,72%	76,32%	79,64%	80,57%	80,63%	82,50%	83,39%
IN-EX	63,51%	67,37%	68,72%	70,67%	73,29%	75,76%	78,40%

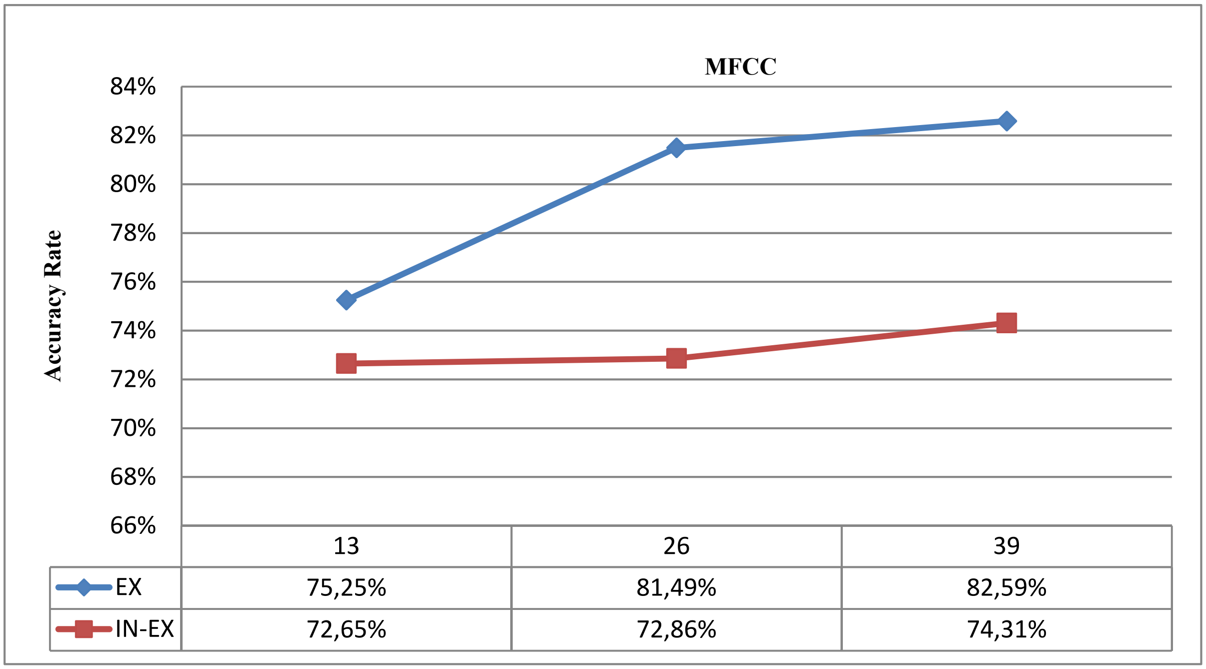
Figure 2.10 Relationship between the accuracy error rate and the number of mixtures for each state of a 7-state HMM with a fixed window size of 30 ms and MFCC_E_D_Z (26 parameters)

Figure 2.10 presents the accuracy results using a 7-state HMM. The observation distribution for each state is modeled by a different number of Gaussian mixtures, varying from 1 to 64. In the first two sets of experiments, we discovered that increasing the number of states and the number of mixtures has a significant impact on the system in terms of classification accuracy, returning rates of up to 83.39%.

Note that the accuracy rates are higher for the expiration phase, owing to the larger number of occurrences of this type of cry in the training corpus. The expiration parts of cry sounds was the easiest to classify, which makes intuitive sense because these parts are different from the other acoustical activities recorded. The purpose of conducting the third set of experiments was to determine the effect of the number of MFCCs on the performance of the system. As Figure 2.11 shows, the higher the number of MFCC parameters, the higher the classification accuracy results. These three tests were performed using the 5-state HMM, a fixed-length window of 50 ms with a 50% overlap.

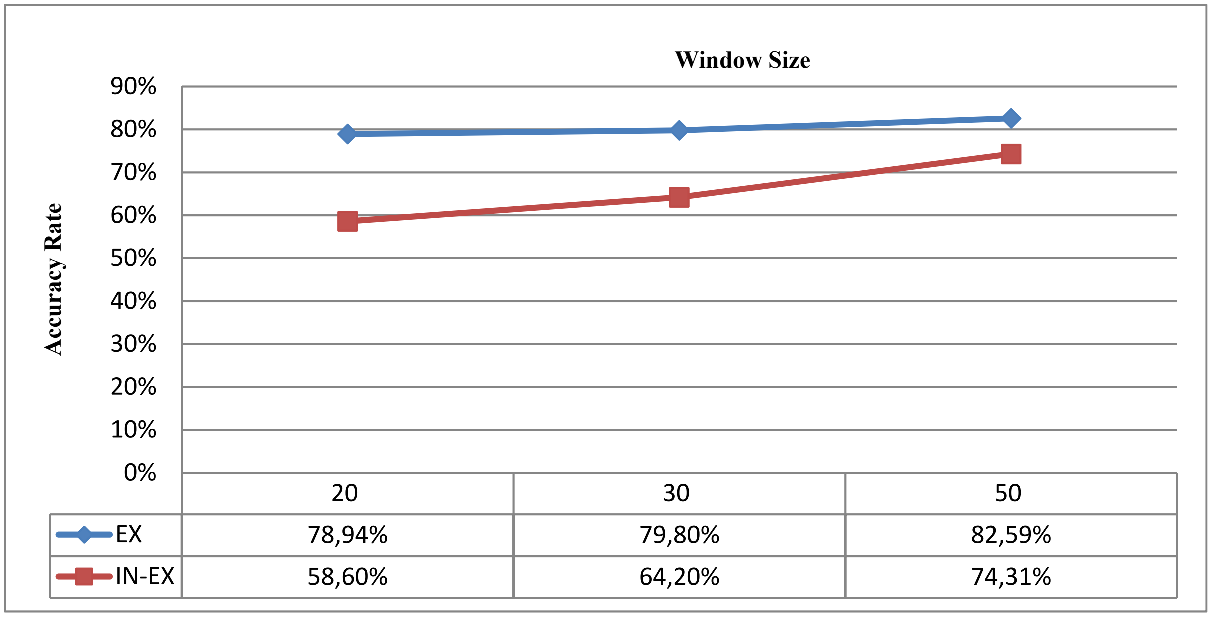
For comparison purposes, we also tested the impact of window size on system performance. As indicated in Figure 2.12, the best rate obtained for the classification of the expiratory parts was 82.59%. We used the same 5-state HMM topology as in the previous experiment, but with a fixed number of feature vectors (39).

We looked at the way in which the performance of the proposed method varied with changes in a large number of its components.



	13	26	39
EX	75,25%	81,49%	82,59%
IN-EX	72,65%	72,86%	74,31%

Figure 2.11 Relationship between the accuracy error rate and the number of MFCC parameters for a 5-state HMM and a window size of 50 ms



	20	30	50
EX	78,94%	79,80%	82,59%
IN-EX	58,60%	64,20%	74,31%

Figure 2.12 Relationship between the accuracy error rate and the window size using a 5-state HMM and MFCC_E_D_A_Z (39 parameters)

Table 2.3 illustrates the accuracy results obtained as a function of the HMM topology, the number of Gaussian mixtures, the number of filter banks of the 12 MFCC, and the window size. The experiments show that the best accuracy is achieved using a 7-state HMM and feature vector with 39 components and a window size of 50 ms. This accuracy is as high as 83.79% for the expiration episodes and 77.93% for the classification of both the expiration and inspiration sounds.

Table 2.3 Overall system accuracy obtained for the MFCC with 39 parameters

	MFCC_E_D_A_Z (39 parameters)					
	20 ms		**30 ms**		**50 ms**	
N0 of States	**INS-EXP**	**EXP**	**INS-EXP**	**EXP**	**INS-EXP**	**EXP**
4	51.76	75.15	54.46	74.3	64.68	79.68
5	58.6	78.94	64.2	78.2	74.31	82.59
6	67.15	79.27	70.72	81	75.76	83.39
7	74.24	80.1	73.87	81.7	**77.93**	**83.79**

On the whole, the best performance is achieved with a window length of 50 ms, and using a 7-state HMM and an MFCC with 13 parameters, along with its energy component, delta, and accelerations.

Typically, the approach wrongly estimates, either positively or negatively, the starting and ending points of the expiratory and inspiratory episodes.

Figure 2.13 illustrates a comparison between a manual segmentation (top) and the output of the proposed segmentation system. This lack of precision in boundary selection should be treated in a future work using temporal approaches.

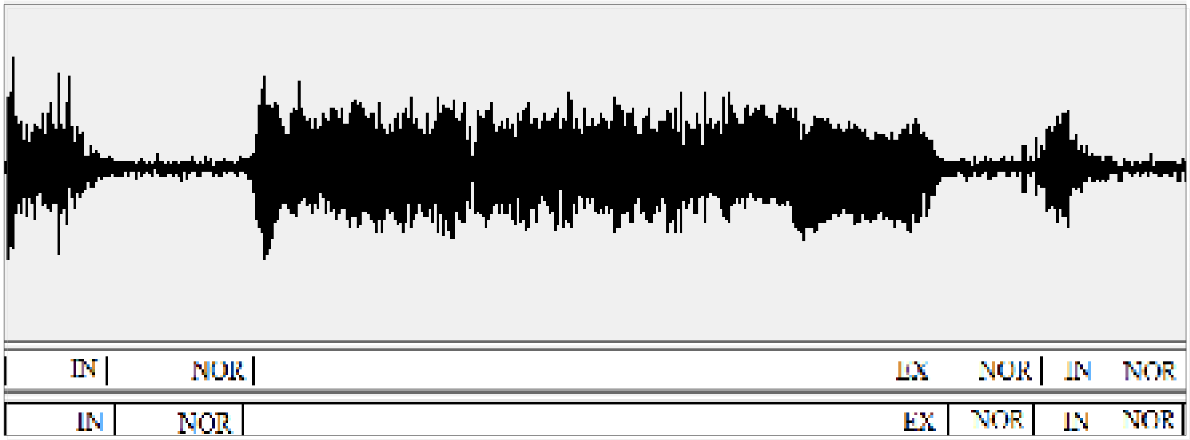


Figure 2.13 Illustration of the lack of precision in boundary selection between the manual segmentation.(top) and the output of the segmentation system

2.13 Conclusion

The paper presents our preliminary results in our ongoing work to develop a tool for the automatic segmentation of newborn cry signals designed to detect the audible expiratory and inspiratory components of the newborn cry.

The authors have demonstrated the effect of some parameters on the performance of the proposed segmentation system. The difficult issues inherent in the manual labeling process are somewhat alleviated in the automated procedure. The system has two main stages, which are feature extraction and training and recognition, which both are performed using the HTK.

This research has revealed that the performance of our automatic newborn cry segmentation system is enhanced by making a number of design improvements:

- use of an HMM topology with a larger number of states to obtain a higher accuracy rate;

- use of multiple mixtures of Gaussian components for a more robust segmentation process;

- use of a larger number of filter banks and MFCC parameters to achieve better trajectory modeling;

- consideration of the impact of window size.

We applied several parameter variations to find the optimal configuration for our segmentation system. By comparing the experimental sets, we found that the best system performance is achieved using the following set of parameters: an MFCC feature vectors with 39 parameters, and an HMM of window size of 50 ms and 7 states. We conclude that the current system's performance can be considered satisfactory. The automatic segmentation system is convenient to operate and gives consistent results. In general, it could be used to create a starting point for further manual refinement if desired or a post-processing stage. The system has been used on recordings of varying types of content with an acceptable degree of success.

Our system gives an accuracy of more than 83%; however, it has some limitations owing to a lack of training data. Consequently, additional manually segmented data need to be added to the training corpus, which should improve the accuracy rate of the system. In the future, its applicability and performance are expected to continue to grow. Based on the results we have obtained, future work can be aimed at improving the system's segmentation performance, which could be achieved by better training of some unit models and adding some post-processing techniques to the system. Many more parameters should be incorporated in any future study to address the problematic issues that were noted in this study.

The tool that we have developed is also part of the effort to develop our automatic Newborn Cry-based Diagnostic System (NCDS). It will serve as the front-end processing unit for the NCDS to improve this system's recognition performance.

Acknowledgments

The authors acknowledge The Bill and Melinda Gates Foundation, which supported the Newborn Cry-based Diagnostic System project through the Grand Challenges Explorations Initiative. The authors are grateful to the staff of the neonatology departments at Saint-Justine Hospital in Montreal and Sahel Hospital in Lebanon for their collaboration.

CHAPITRE 3

EXPIRATORY AND INSPIRATORY CRIES DETECTION USING DIFFERENT SIGNALS’ DECOMPOSITION TECHNIQUES

L. Abou-Abbas [a], C, Tadj [b], C. Gargour [c], and L. Montazeri [d]

[a,b,c] Department of Electrical Engineering, Ecole de Technologie Supérieure,
1100 Rue Notre-Dame West, Montreal, Quebec, Canada H3C 1K3
[d] Department of Electrical Engineering, Polytechnique Montreal, Quebec, Canada H3T 1J4

This article is accepted in “Journal of Voice”, in May 2016

3.1 Abstract

This paper addresses the problem of automatic cry signal segmentation for the purposes of infant cry analysis. The main goal is to automatically detect expiratory and inspiratory phases from recorded cry signals. The approach used in this paper is made up of three stages: signal decomposition, features extraction, and classification. In the first stage, Short Time Fourier Transform, Empirical Mode Decomposition (EMD), and Wavelet Packet Transform have been considered. In the second stage, various set of features have been extracted and in the third stage, two supervised learning methods, Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) with 4 and 5 states have been discussed as well. The main goal of this work is to investigate the EMD performance and to compare it to the other standard decomposition techniques. A combination of two and three intrinsic mode functions (IMF) resulted from EMD has been used to represent cry signal. The performance of nine different segmentation systems has been evaluated. The experiments for each system have been repeated several times with different training and testing data sets, randomly chosen using a 10-fold cross validation procedure. The lowest global classification error rates of around 8.9% and 11.06 % have been achieved using a GMM classifier and HMM classifier respectively. Among all IMFs combination, the winner combination is (IMF3+IMF4+IMF5).

Keywords: Automatic segmentation; Empirical mode decomposition; Wavelet Packet Transform; Mel-frequency Cepstral coefficients; Gaussian Mixture Models, Hidden Markov Models.

3.2 Introduction

Cry is the only possible way for newborns to express their needs and their physical conditions since they are not able to communicate with words. Cry signals have been studied for many years and it has become evident that cry signals can provide valuable information concerning physiological and physical states of infants. Most of researches on infant cry focused on extracting information from infant cry signals with known medical problems such as prematurity asphyxia, hypoglycemia, Down syndrome, and meningitis. For example, the cries of infants with neonatal asphyxia and meningitis are high-pitched and the cry duration is very short or unusually long with melody type rising or falling-rising in compared to healthy infants. Preterm babies have higher minimum fundamental frequency compared to normal babies.

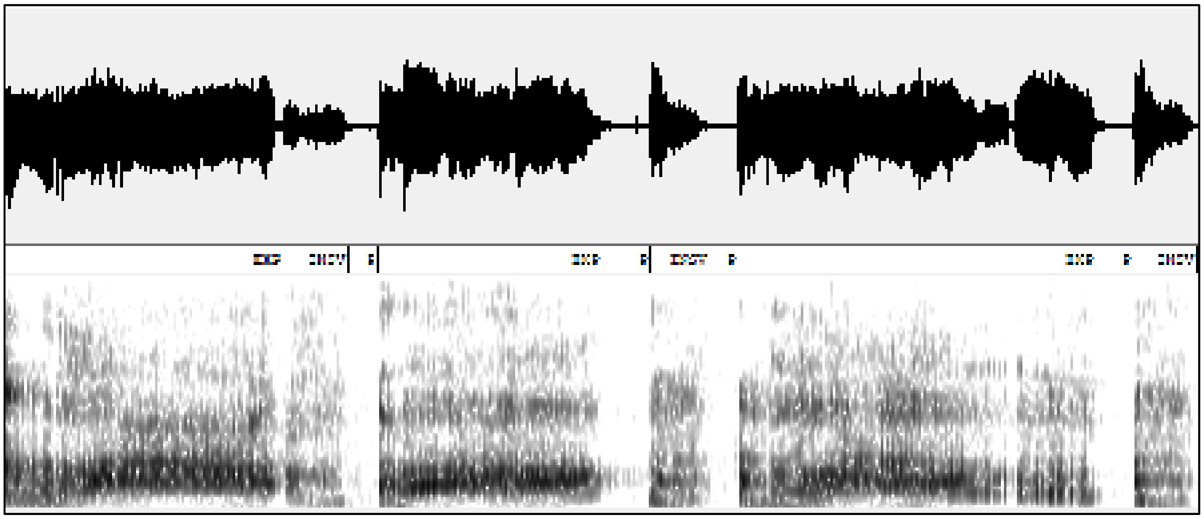


Figure 3.1 An example of a portion of cry signal with its corresponding components expiration (EXP), audible inspiration (INSV), and pauses (P)

Cries of infants with hyperbilirubinemia have significant changes in f0 over a 100ms period. For the reason of cries, features such as pitch, and loudness are able to distinguish hunger cry from pain cry.

(Amaro-Camargo et Reyes-García, 2007; CanoOrtiz, Beceiro et Ekkel; Corwin, Lester et Golub, 1996; Wasz-Höckert, Michelsson et Lind, 1985).

Given these evidences, many researchers have suggested an automatic system to classify infant cries which is more like a pattern recognition problem, similar to Automatic Speech

Recognition (ASR) systems. The aim of the automatic classification system is to give clinician an early diagnostic result if the baby may have high probability to get specific type of medical diseases. As any ASR system, a cry classification system needs a segmentation module which role is to detect useful part of recorded signal and reject other acoustic activities to be thereafter classified. Infant cry signals consist of a sequence of audible expiratory and inspiratory phases separated by a period of silence or by unvoiced phases of cry (inaudible expiratory and inspiratory phases during cry). A cry signal recorded in a real environment usually contains different acoustic activities other than the cry such as background noise, speech, sound of medical equipment and silence. This work aims to retrieve most relevant and audible sections from cry signals recorded in a realistic clinical environment as well as distinction between expiratory and inspiratory phases of the cries. One way to address this problem is to manually segment recorded audio signals and pick out important cry parts. However this manual task is tiresome and prone to errors when the volume of data is large. It is therefore essential to design a segmentation system able to automate this tedious task and being implemented in a real-time clinical decision support tool. A typical waveforms of a cry signal, expiratory phase, as well as inspiratory phase of cry signals are shown in Figures 3.1, 3.2 and 3.3, respectively.

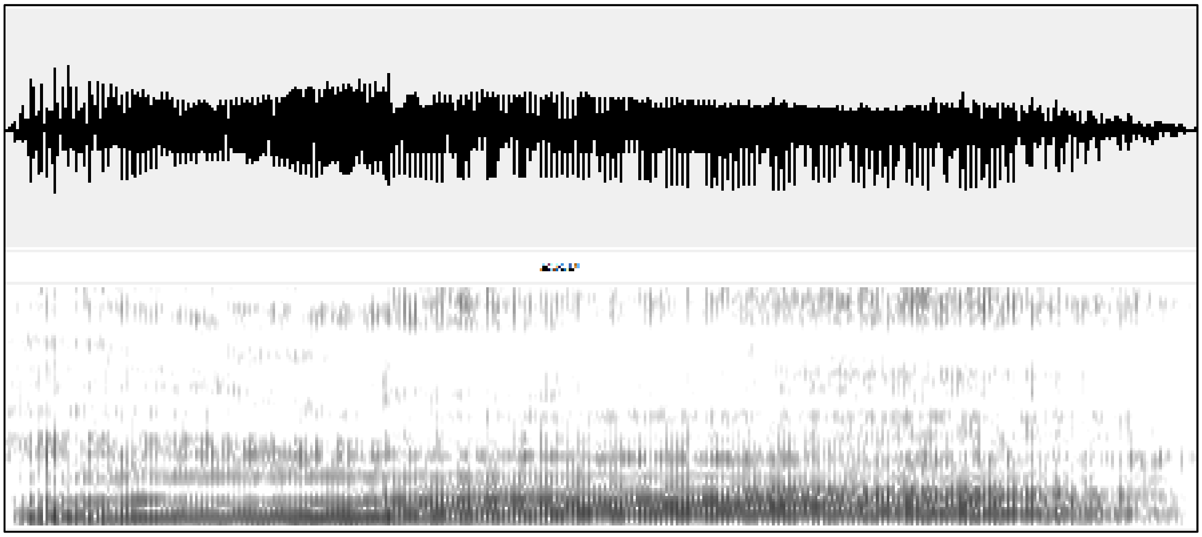


Figure 3.2 An example of a waveform and spectrogram of an expiration phase

Some attempts to segment cry signals have been reported in the literature. Many studies used the spectrogram to segment cry signals manually through visual and audio monitoring (Messaoud et Tadj, 2010).

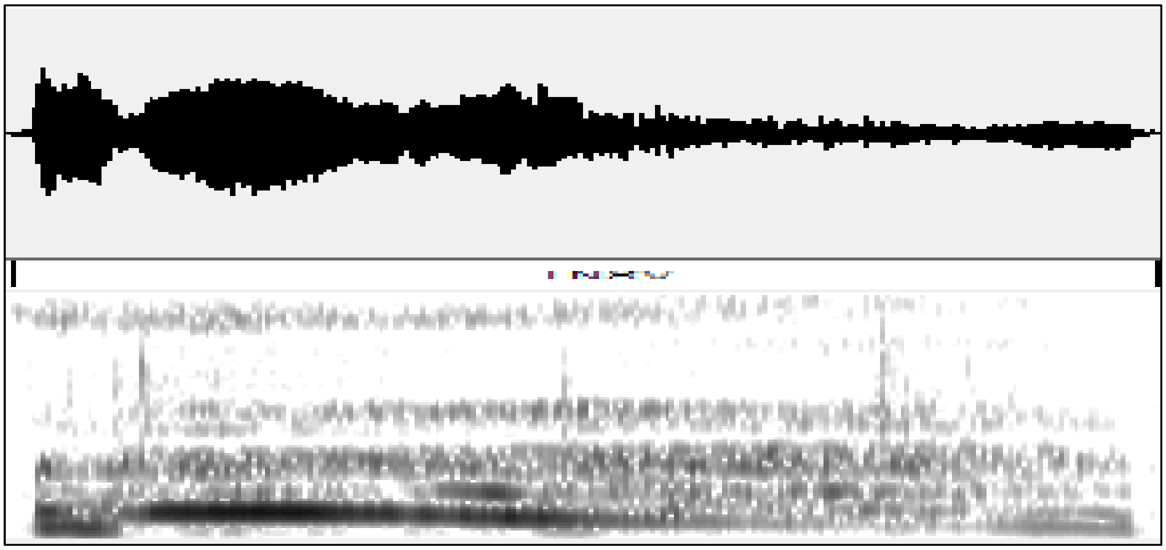


Figure 3.3 An example of a waveform and spectrogram of an inspiration phase

In one hand, automatic segmentation is often desired to manipulate all automated diagnostic system and in the other hand, since the manual segmentation is an extremely long, tedious task and prone to errors especially when the amount of data is large. A number of recent works have been done on infant cry segmentation based on the time domain characteristics of the signal. The problem of cry segmentation was being considered as the problem of Voice Activity Detection. (Kuo, 2010; Várallyay, 2006) used high pass filter to reduce most of the background noise and in order to distinguish between important and less important parts of the cry signals, they applied Short Term Energy or/and Zero Crossing Rate by using a satisfactory threshold. However, these methods perform well when cries have been recorded within a laboratory environment and fail under noisy or clinical environment.

In other research efforts, the cry detection problem was considered as the problem of start and end points detection of a cry unit. Based on the hypothesis that cry segments have four times more energy than unvoiced segments, authors in (Rui et al., 2010; Rúız, Reyes et Altamirano,

2012) defined some guidelines to detect cry units based on a dynamic threshold for each record. In these works, authors eliminate not only useless sounds from the signals but also inspiratory sounds of the cry. Another technique used in (Várallyay, Illényi et Benyó, 2009), considers the problem of cry segmentation as the problem of Voiced/Unvoiced decision. In (Várallyay, Illényi et Benyó, 2009) authors modified a well-known fundamental frequency estimation method, the Harmonic product spectrum to check the regularity of the segment analyzed in order to classify it as important or not important cry parts. In (Manfredi et al., 2009) authors used the Simple Inverse Filter Tracking (SIFT) algorithm to detect the voiced frames of cry based on a threshold of the autocorrelation function. The use of threshold limits the attractiveness of the mentioned approaches and decreases their performance in low signal-to-noise ratio levels.

Inspiratory cry parts have been proven to be important in identification of newborns at risk for various health conditions. Despite this evidence, it is thus surprising that in most of researches on analyzing cry signals, the inspiratory parts of cry were ignored and not being considered in analysis while the main focus has been only on extraction acoustical data of expiratory parts.

A cry sound segmentation system has been implemented in this work. The proposed system has the capability of detecting three different acoustic classes: Audible expiration, Audible inspiration, and others (including unimportant acoustics like speech, medical machines sounds, noise, etc). Different signal decomposition techniques such as Wavelet Packet Transform (WPT) and Empirical Mode Decomposition (EMD) have been examined for the features extraction phase.

The WPT has been widely and successfully used in various applications in the voice signal processing domain. It decomposes cry signal into sub-bands in order to give better resolution. The EMD has been successfully used in denoising and characterizing of non-stationary and multicomponent signals such as heart sound signal. (Becerra et al., 2012; Chu et al., 2013; Saïdi, Pietquin et André-Obrecht, 2010; Shi, Xiong et Chen, 2014; Tu et Yu, 2012).

Statistic generative models such as GMM and HMM have been also chosen as classifiers to distinguish between the three different classes. Recently, GMM and HMM techniques proved to be very successful by many researchers especially in speaker recognition. These models provide a robust solution when a large amount of training data is available.

The remainder of the paper is organized as follows: Recording Procedure and Cry Database are presented in section 3.3. Proposed Methodology is described in section 3.4. Mathematical backgrounds of signal decompositions methods, features extraction, modeling, and classification approaches used in this work are addressed in section 3.5. An evaluation of the proposed methods and results obtained are reported in section 3.6. Finally, section 3.6 concludes the paper offering a list of suggestions for further research.

3.3 Recording Procedure and Cry Database

Data used in this research has been obtained from the Newborn Cry-based Diagnostic System (NCDS) database. A description about the data collection technique was presented in the previous work (Abou-Abbas, Fersaie Alaie et Tadj, 2015). 507 cry signals were randomly picked up from the database. Cry signals were recorded with a sampling rate of 44.1 KHz and a sample resolution of 16 bits. The 507 cry signals of average duration of 90 seconds have been recorded from 203 babies including both normal and pathological cases.

The constructed data set contains different kinds of cries such as pain, hunger, birth cry, etc. It also includes infants’ cries in different recording environment and conditions, from silent to very noisy combined with different acoustic activities like speech, machines’ sounds, noise, silence, etc. Cry signals have been manually segmented and labeled using Wave Surfer application (Sjölander et Beskow). To divide the dataset between the training and the testing sets, the ten-fold cross validation was carried out. The dataset was partitioned into ten folds: 9 folds for the training set and the remaining fold for the testing set. 10 tests were conducted with different choice of folds. Data base statistics and details about average time of each class in the testing and training datasets are presented in Table 3.1 and 3.2, respectively.

Table 3.1 Database statistics

			Number of babies	**Number of signals**
Female	Full term	Healthy	56	141
		Pathological	34	94
	Preterm	Healthy	20	23
		Pathological	17	49
Male	Full term	Healthy	4	11
		Pathological	54	146
	Preterm	Healthy	5	11
		Pathological	13	32
Total			**203**	**507**

Table 3.2 Data used for training and testing corpuses

Classes	**Time in sec**	**Average time for training corpus/sec**	**Average time for testing corpus/sec**
Expiration	21414	19348	2066
Inspiration	2154.8	1930	224.8

3.4 Proposed Methodology

The basic contribution of this paper is the proposition of a practical cry sounds segmentation systems with the ability to detect audible expiratory and inspiratory cry episodes. This section describes the modules required for developing the proposed system. A block diagram of the general system architecture is presented in Figure 3.4. The framework is based on supervised pattern classification and it consists of two stages: training stage and testing stage. In either stages, signal decomposition module receives the input cry signal. It converts the original signal from time domain to another domain in order to better characterize it. Training and testing stages also share the same features extraction module.

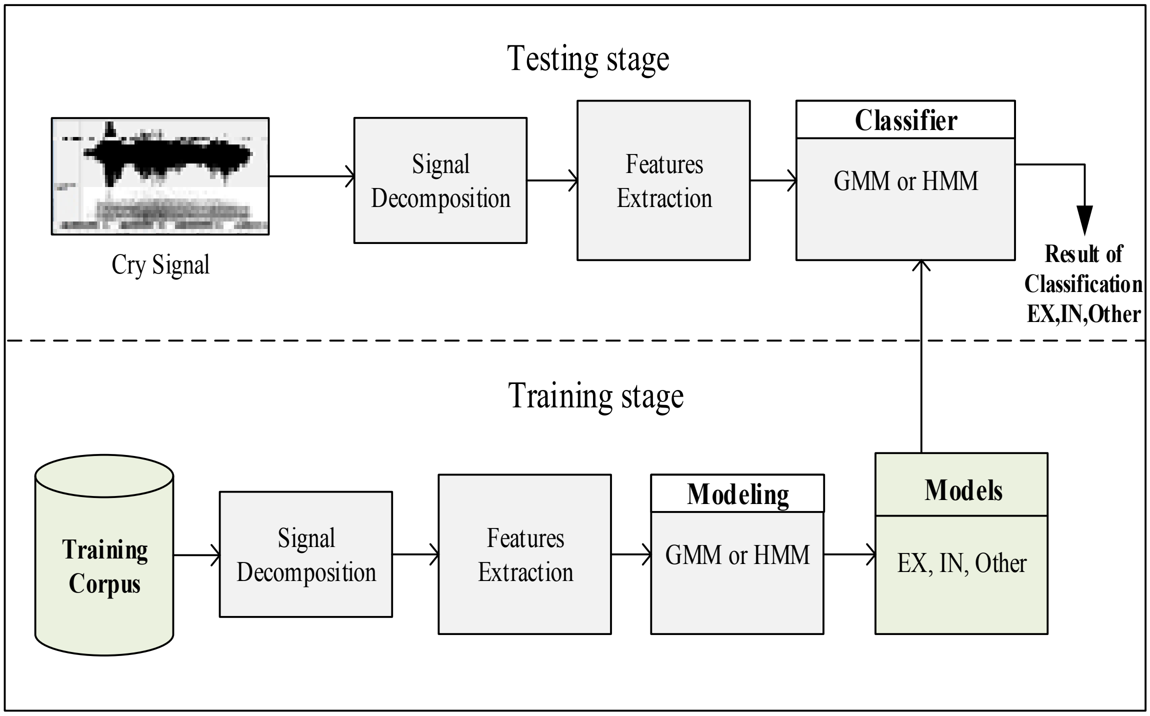


Figure 3.4 Block diagram of the system architecture

This module receives the decomposed signal as input and extracts important acoustic information within each frame to form a set of feature vectors. Training involves learning the system and creating an acoustic model for each class based on the acoustic training data set. Re-estimation algorithms are used after the initial training to adapt models' parameters to various conditions. Subsequently, the created models, stored in a database as reference models, are used to classify testing dataset and to measure the system performance during the testing stage. A description of each module is described in the following sub-sections.

3.5 Mathematical Background

3.5.1 Signal Decomposition

Signal decomposition, also referred to the front end module, is the first step in the proposed method. Since most of the audio signals are nonlinear and non-stationary, a time series and/or frequency analysis of the signals are needed. Fourier transform, Wavelet Packet Transform, and Empirical mode decomposition are the most common analysis techniques addressed in the literature. In this paper, two cry segmentation systems based on WPT and EMD are designed and compared using the system already designed based on Fast Fourier Transform in our previous work (Abou-Abbas, Fersaie Alaie et Tadj, 2015).

3.5.1.1 Wavelet Packet Transform

The main objective of the wavelet analysis is to apply varying size windowing technique on the signal under study. In low-frequency band study, a large window size should be used while in high-frequency band study, small windows size should be employed (Misiti et al., 1996). Wavelet Packet Transform (WPT) represents a generalization of wavelet decomposition that could offer a more precise signal analysis by considering both low and high pass results. WPT decomposes the original signal into different sub-bands in order to get better resolution. Each WPT is associated with a level j which each splits the frequency band [0, fs/2] to 2^j equal bands by decomposing both low and high frequency components called approximation and detail coefficients, respectively. The result of this decomposition is a balanced tree structure. WPT has been widely and successfully used in various applications in voice signal processing domain. Based on experiences achieved during this work, WPT level 5 on different orders of Daubechies wavelet db1, db10 and db20 is employed in this study. In Figure 3.5 examples of some wavelets functions from the Daubechies family are shown.

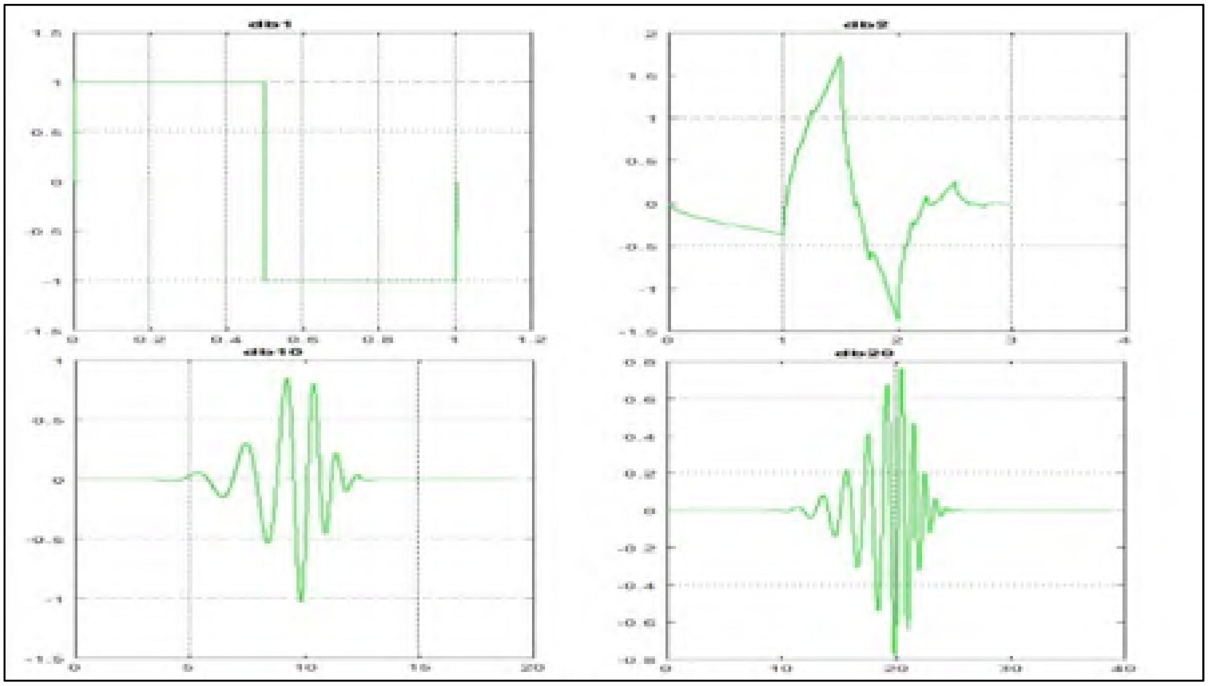


Figure 3.5 Waveforms of some versions of Daubechies wavelet

Considering that h(n) is the low pass filter of length 2N called also scaling filter, and g(n) is the high pass filter of length 2N called also wavelet filter, Wavelet Packet functions are estimated using the following equations:

$$W_{2n}(x) = \sqrt{2}\sum_{k=0}^{2N-1} h(k)W_n(2x-k) \tag{3.1}$$

$$W_{2n+1}(x) = \sqrt{2}\sum_{k=0}^{2N-1} g(k)W_n(2x-k) \tag{3.2}$$

Where $W_0(x) = \varphi(x)$ is the scaling function and $W_1(x) = \Psi(x)$ is the wavelet function. For more details about wavelet coefficients calculation, readers are referred to the publication of

Mallat (Mallat, 2000). An example of a Wavelet packet Tree decomposition of level 5 and the corresponding frequency intervals at each level is given in Figure 3.6.

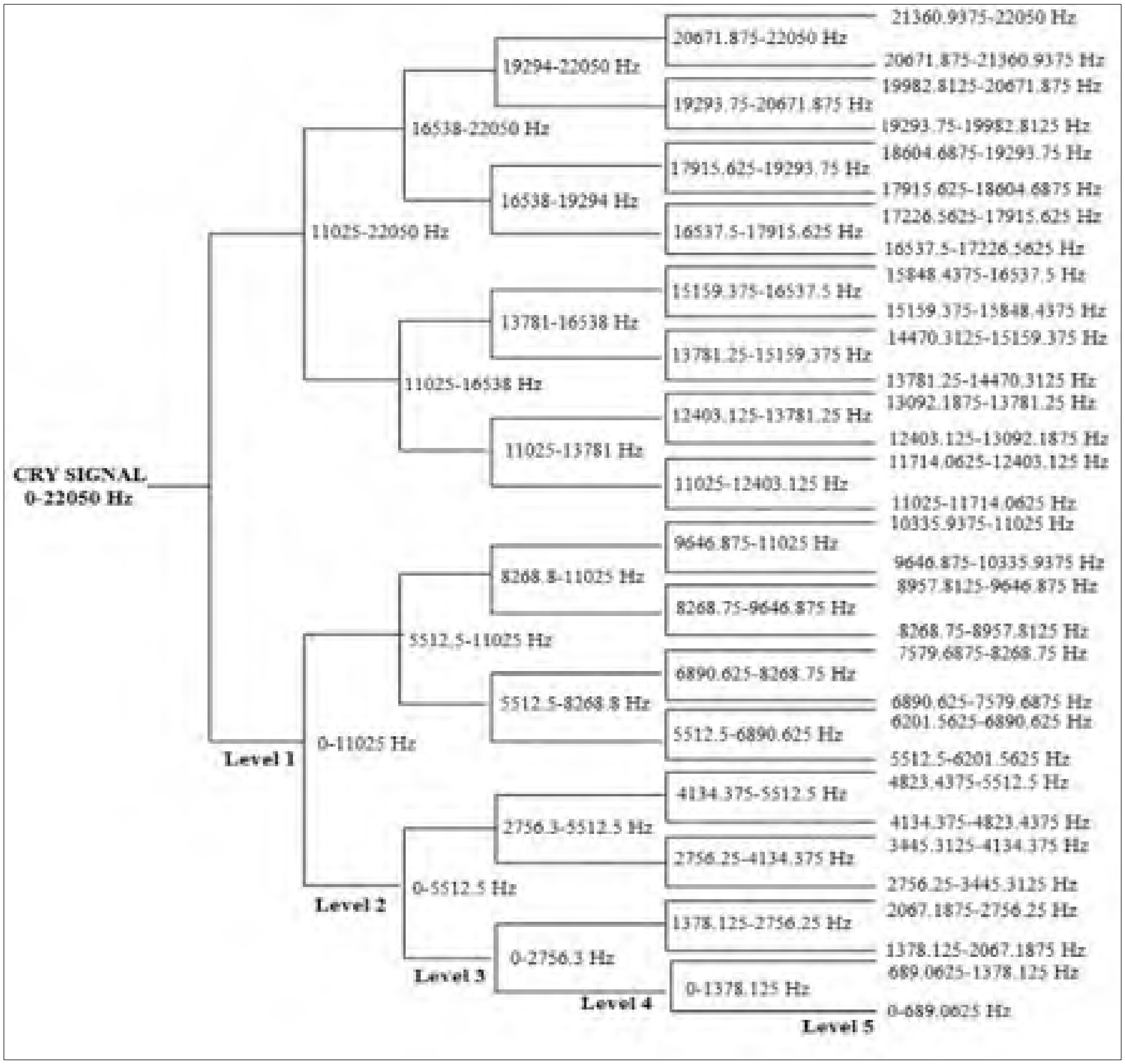


Figure 3.6 Example of wavelet packet decomposition level 5 of a cry signal at a sampling frequency of 44100 HZ

The Sampling frequency used in this work is 44100HZ. Figures 3.7 and 3.8 are examples of details and approximations coefficients at level 4 of inspiration and expiration phases, respectively.

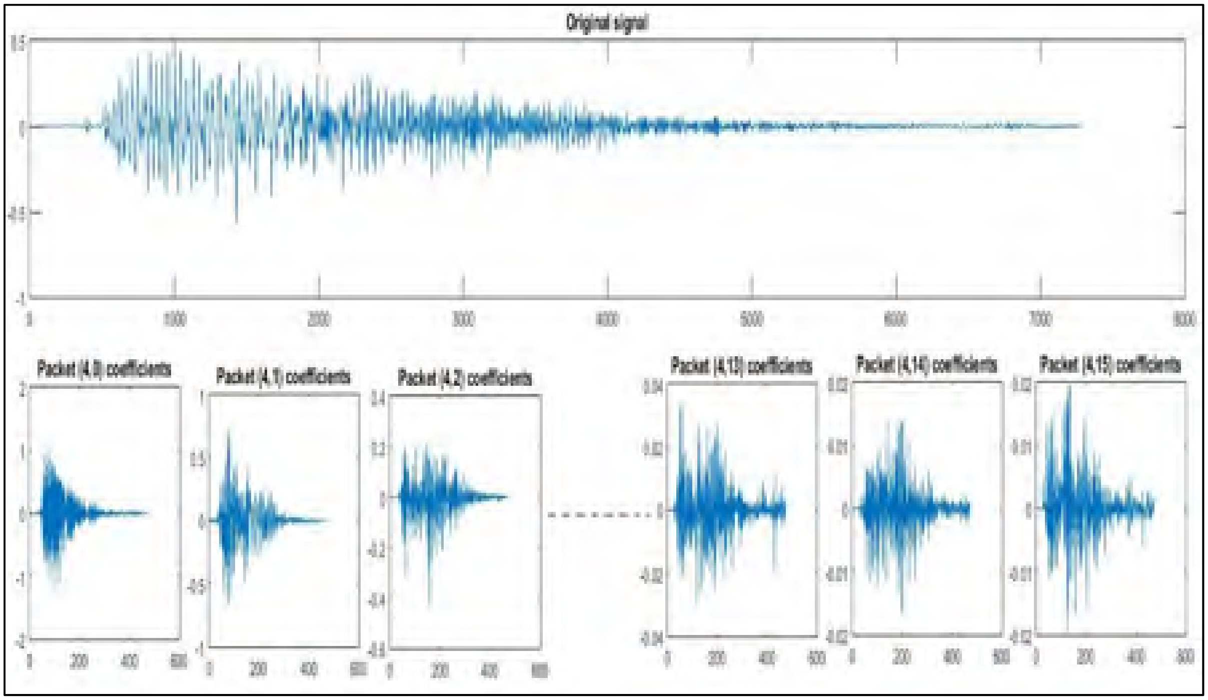


Figure 3.7 Level 4 of wavelet packet decomposition of an inspiration

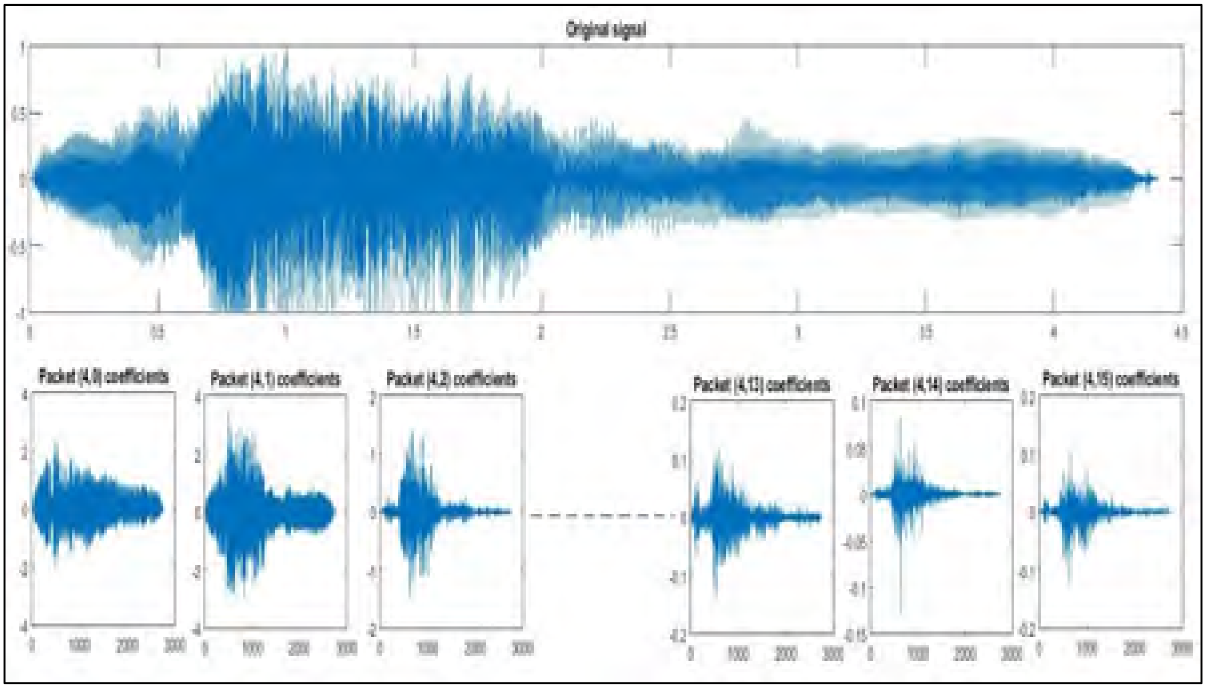


Figure 3.8 Level 4 of a wavelet packet decomposition of an expiration

3.5.2 Empirical Mode Decomposition

Empirical mode decomposition (EMD) algorithm proposed by Huang et al. in 1998 as an efficient tool to analyze natural signals which are mostly non-linear and non-stationary.

This method decomposes the given signal into a set of functions in time domain and of the same length of the original signal allowing for preserving the frequency variation in time. This is the key feature of the EMD algorithm which helps to characterize natural signals being produced by various causes at certain time intervals.

EMD algorithm applies a sifting process in order to break down the given signal into a set of intrinsic mode functions (IMFs) which represents simple oscillatory mode of the original signal.

Sifting process is an iterative process during which smooth envelopes are formed by local minima and maxima of the signal and their mean is subsequently subtracted from the initial signal to finally produce an IMF satisfying two criteria:

1. The number of extremes and the number of zero crossings in the whole sequence of data are equal or differ by one;

2. The mean value of the envelops of local extremes is zero at all points. Examples of extracted IMFs from expiratory and inspiratory parts of cry signal using EMD are depicted in Figures 3.9 and 3.10, respectively.

The following sifting approach has been adopted in this work to extract IMFs from a cry sign $x(t)$al:

1. Identify the local minima and local maxima of the given signal;

2. Interpolate the local maxima using cubic splines interpolation method to form the upper envelope $Env_U\ (t)$;

3. Interpolate the local minima using cubic splines interpolation method to form the lower envelope $Env_L\ (t)$;

4. Obtain the mean envelope of the upper and lower envelopes:

$$Env_m = \frac{Env_U(t) + Env_L(t)}{2} \tag{3.3}$$

5. Subtract the mean envelope from the signal:

$$h(t) = x(t) - \ Env_m(t) \tag{3.4}$$

6. Iterate with x(t) = h(t) until h(t) satisfies the IMF criteria;

7. Calculate the residue by subtracting the obtained IMF from the signal:

$$r(t) = x(t) - \ h(t) \tag{3.5}$$

8. Repeat the process by considering the residue as the new signal: $x(t) = r(t)$ until the termination condition is satisfied.

The original signal can be reconstructed by summing up the obtained IMFs and the residue:

$$x(t) = \sum_{i=1}^{n} C_i(t) + \ r_n(t) \tag{3.6}$$

where $C_i(t)$ and $r_n(t)$ represent the i-th IMF and the residue function, respectively. The number of IMFs extracted from the original signal is also represented by n.

The adopted termination condition in this work is the minimum number of extrema in the residue signal. However, usually a certain number of IMFs which contain more important information are used in next steps. It has been proven through several experiments in this work that the first five IMFs of cry signals have the most important information.

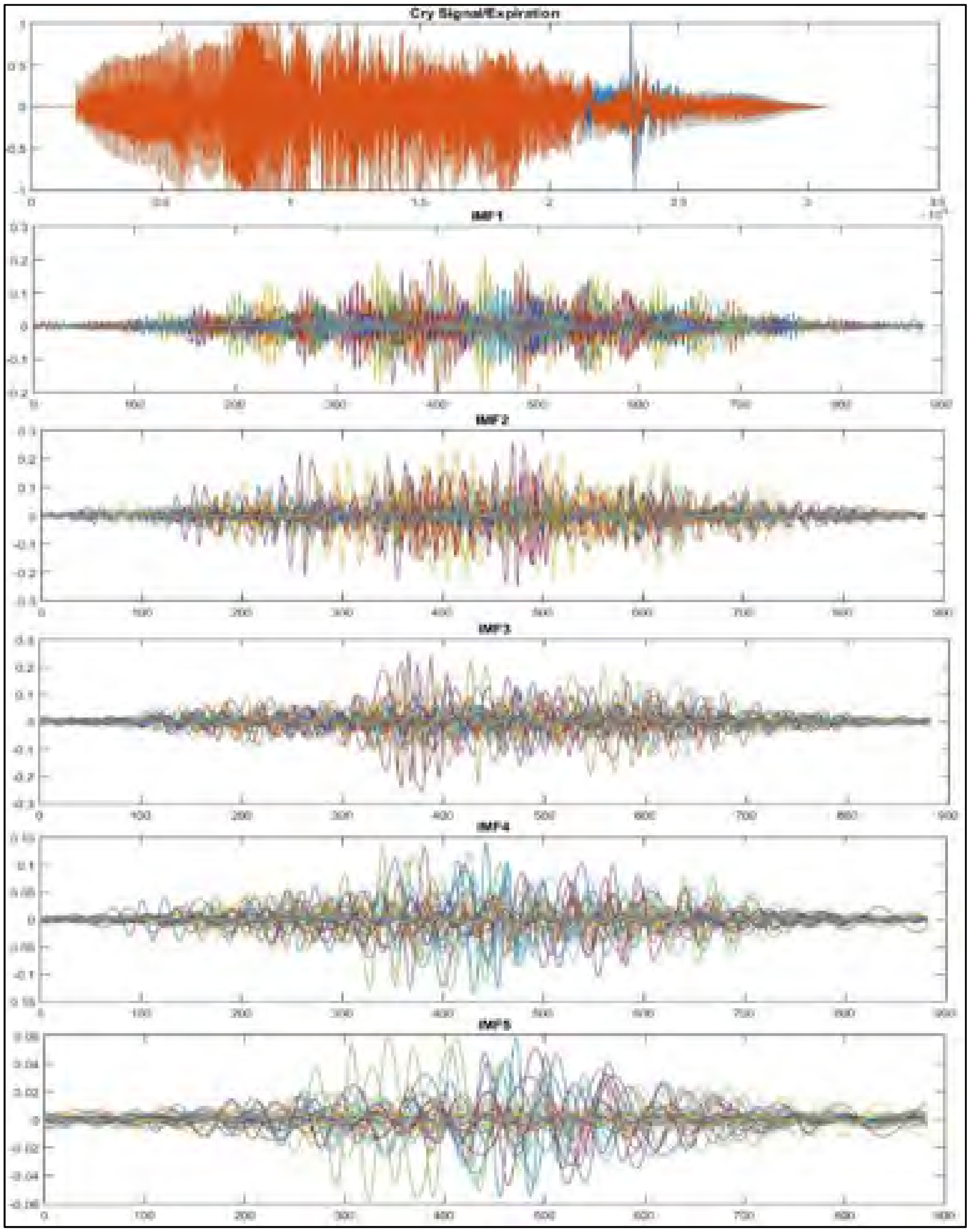


Figure 3.7 Example of IMF functions of an expiration

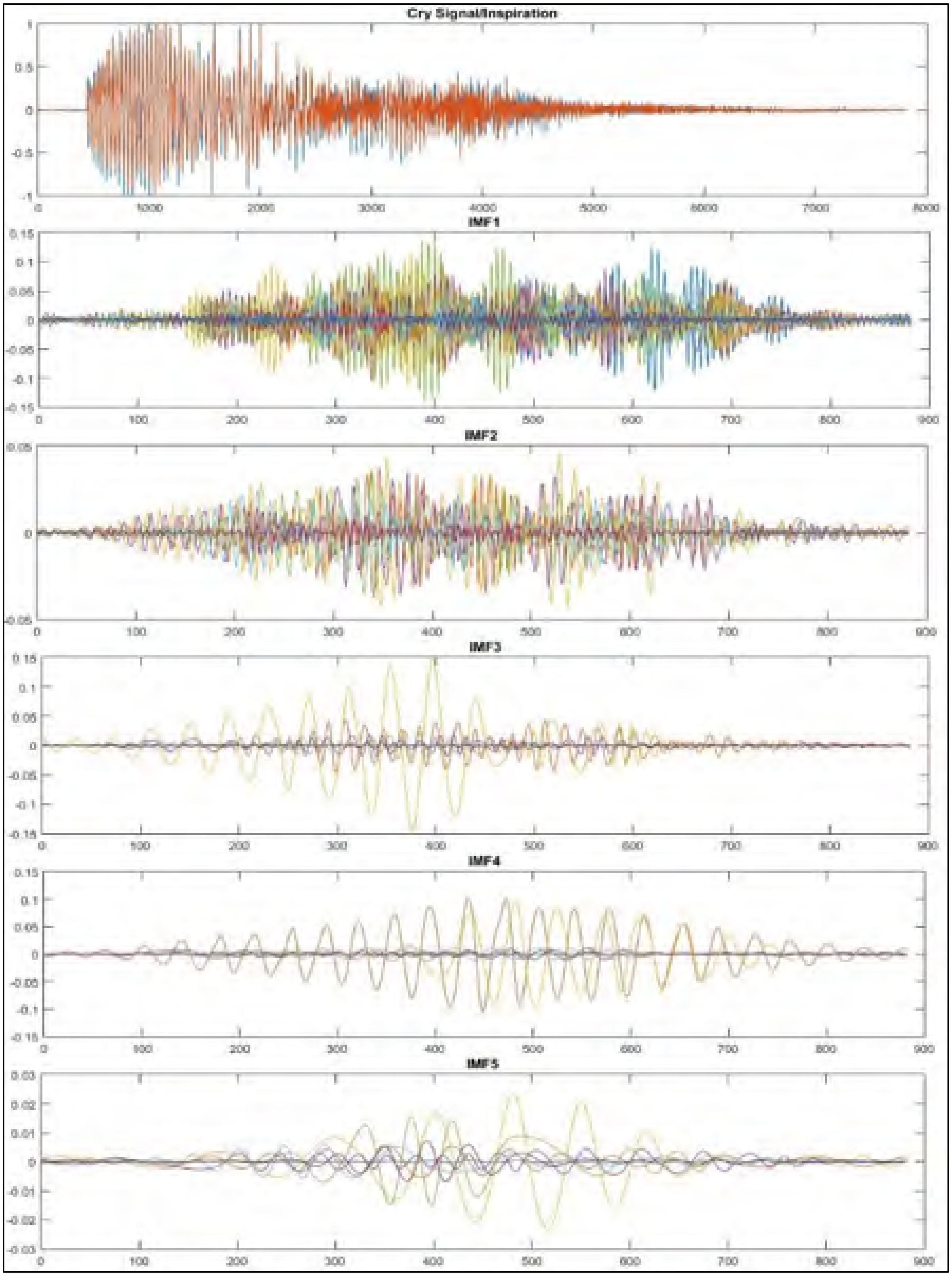


Figure 3.8 Example of IMF functions of an inspiration.

3.6 Features Extraction

Features extraction can be defined as the most prominent step in an automatic recognition system. It consists of decreasing the amount of information present in the signal under study by transforming the raw acoustic signal into a compact representation. Among several features extraction techniques that have been used in previous works, MFCC which is still one of the best methods has been chosen. It demonstrates good performance in various applications as it approximates the response of the human auditory system. Wavelet Packet based features have been also chosen due to their efficiency for segmentation proved in the previous work (Abou-Abbas, Fersaei Alaei et Tadj, 2015).

3.6.1 FFT-Based MFCC

MFCC are used to encode the signal by calculating the short term power spectrum of the acoustic signal based on the linear cosine transform of the log power spectrum on a nonlinear Mel scale of frequency. See Figure 3.11.

Mel scale frequencies are distributed in a linear space in the low frequencies (below 1000 Hz) and in a logarithmic space in the high frequencies (above 1000 Hz) (Rabiner et Juang, 1993b).

The steps from original input signal to MFC coefficients are as follows:

1. Slice signal into small segments of N samples with an overlapping between segments;

2. Reduce discontinuity between adjacent frames by deploying Hamming window which has the following form;

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \tag{3.6}$$

3. Use FFT to convert the signal into spectrum form;

4. Consider the log amplitude of the spectrum and apply it to the Mel scale filter banks. The famous formula to convert f Hz into m Mel is given in the equation below;

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{3.7}$$

5. Apply DCT on the Mel log amplitudes;

6. Perform IFFT and the resulting amplitudes of the spectrum are MFCCs and are calculated according to the equation below.

$$c_n = \sum_{k=0}^{n-1} \log(S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right], n = 1, 2, ..., K \tag{3.8}$$

.

where S_k is the output power spectrum of filters and K is chosen to be 12

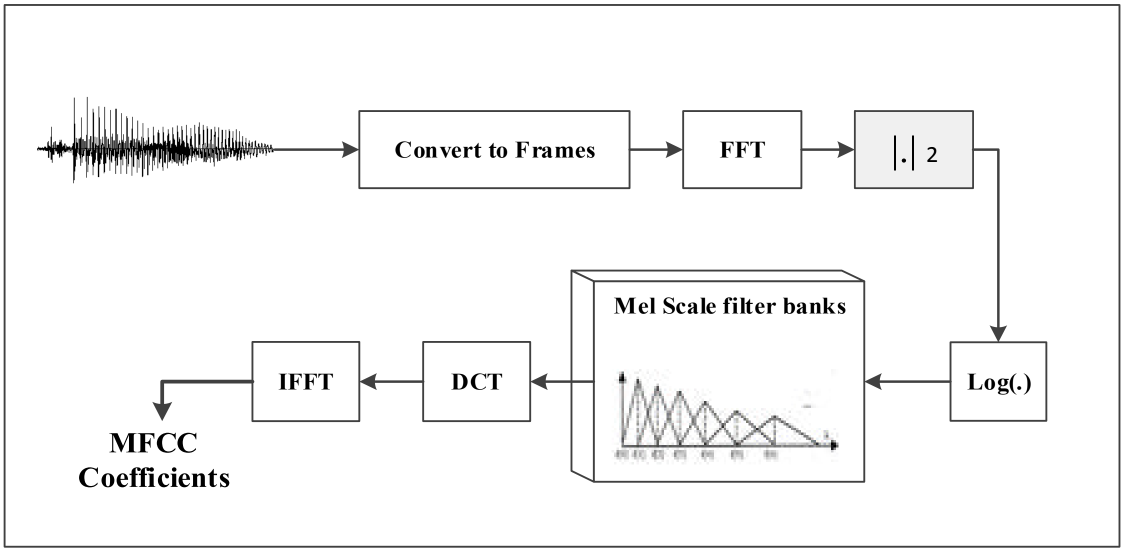


Figure 3.9 Extraction Mel Frequency Cepstral Coefficients (MFCC) from the audio recording signals

3.6.2 Wavelet Packet-Based Features

The shortcoming regarding traditional MFCCs is related to the use of FFT whose calculation is based on fixed window size. Another drawback concerning MFCCs is the assumption that the segment is stationary during the frame duration, it is however possible that this assumption could be incorrect. In order to solve this issue, wavelets have been given particular consideration due to their multi-resolution property. The extraction of features based on Wavelets similar to MFCC with higher performance has been shown in several works and in different ways (Farooq et Datta, 2001; Farooq et Datta, 2003; Farooq, Datta et Shrotriya, 2010; Sarikaya, Pellom et Hansen, 1998; Siafarikas, Ganchev et Fakotakis, 2007; Siafarikas, Ganchev et Fakotakis, 2004). In (Sarikaya, Pellom et Hansen, 1998) authors proposed two set of features called Wavelet Packet Parameters- (WPP) and sub-band based cepstral parameters (SBC) based on WPT analysis and proved that these features outperform traditional MFCCs. Authors of (Gowdy et Tufekci, 2000) proposed Mel-Frequency Discrete Wavelet Coefficients (MFDWC) by applying DWT instead of DCT to the Mel scale filters banks of the signal. MFDWC was used in many recent works and proved its performance in speech and speaker recognition (Bai, Wang et Zhang, 2013; Farhid et Tinati, 2008; Gowdy et Tufekci, 2000).

In (Farooq et Datta, 2001), authors used Admissible Wavelet Packet. The division of frequency axis is performed such that it matches closely the Mel scale bands. In (Siafarikas, Ganchev et Fakotakis) and in (Abdalla et Ali, 2010) another feature extraction technique is presented for deployment with speaker identification: same MFCCs extraction technique presented in the section 3.2.1 but applying at the input wavelet channels instead of the original signal. In this work, features based on WPT have been considered and the following steps have been taken for calculation purposes:

- the wavelet packet transform is used to decompose the raw data signal into different resolutions levels at a maximum level of j=5;

- the normalized energy in each frequency band is calculated according to the following formula;

$$E_j = \frac{1}{N_j}\sum_{m=1}^{N_j}\left[W_j^n(m)\right]^2, j = 1,2....,B \tag{3.8}$$

where W $_j$ ^{n} (m) is the mth coefficient of WPT at the specific node W$_j^n$, n is the sub-band frequency index and B is the total number of frequency bands obtained after WPT

- the Mel scale filter banks are then applied to the magnitude spectrum;

- the logarithms of the Mel energies obtained in each frequency band are then de-correlated by applying the discrete cosine transform according to the following formula.

$$WE\_DCT(n) = \sum_{p=0}^{B-1}\log_{10}\left(E_{p+1}\right)\cos\left[n\left(p+\frac{1}{2}\right)\frac{\pi}{B}\right], n = 0,1,...,B-1 \tag{3.9}$$

WE_DCT, stands for wavelet-energy based DCT, is estimated from wavelet channels and not from the original signal. See Figure 3.12.

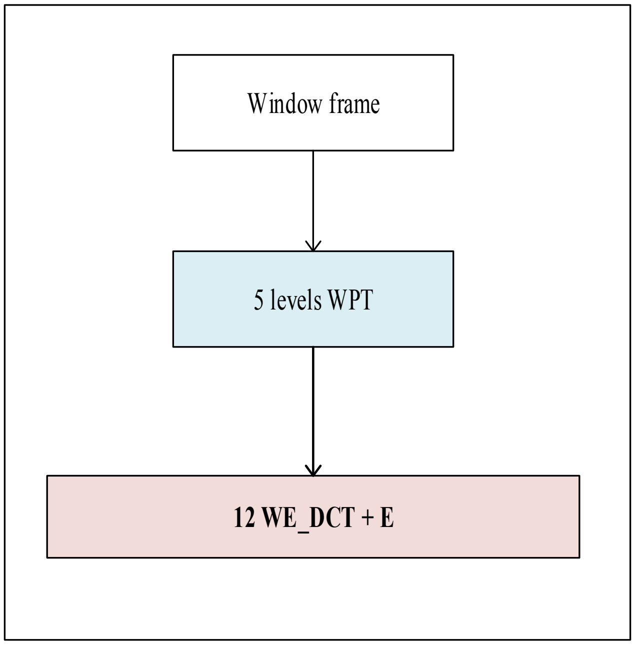


Figure 3.10 Features extraction step after WPT

3.6.3 EMD-based MFCC

These coefficients are estimated by applying MFCC extraction process on each IMF or on the sum of IMFs instead of applying it on the original signal. This technique has been successfully used in many recent works in speech and Heart signals classification (Becerra et al., 2012; Chu et al., 2013; Saïdi, Pietquin et André-Obrecht, 2010; Shi, Xiong et Chen, 2014; Tu et Yu, 2012).

EMD algorithm with resolution of 50 dB, residual energy of 40 dB has been applied to the subjected cry signals in order to decompose them into five IMFs. Next, four different combinations of two or three IMFs have been created to be used in feature extraction phase. These sets are as follows:

Set 1 IMF34=IMF3+IMF4
Set 2 IMF45=-IMF4+IMF5
Set 3 IMF234=IMF2+IMF3+IMF4
Set 4 IMF345=IMF3+IMF4+IMF5

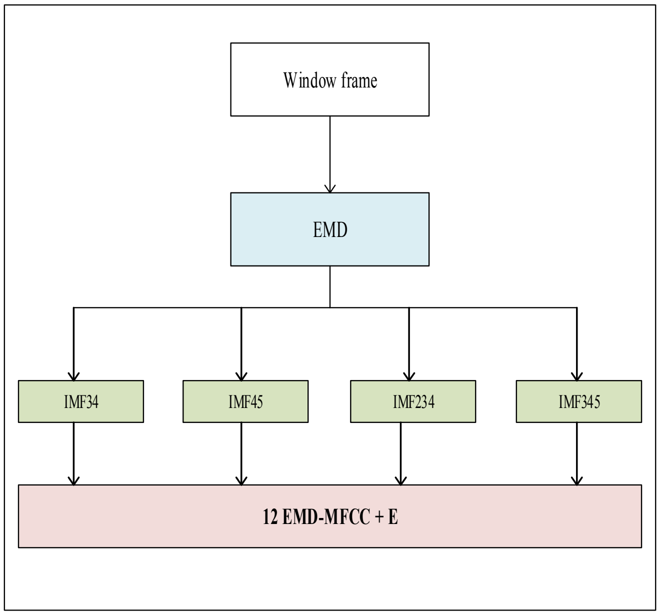


Figure 3.11 Features extraction step after EMD

12 Mel-frequency cepstral components as well as their corresponding energy have been further derived from different sets of IMFs. See Figure 3.13.

3.7 Modeling and Classification

Once the important parameters are retrieved from an input signal (train or test), these parameters are used as input to a nonlinear classifier whose role is to correctly attribute a class to an input frame under numerous conditions. For the classification stage of this research, two efficient statistical classifiers widely used in machine learning and pattern recognition over the last decades especially in speech and speaker recognition have been chosen: The Hidden Markov Models and Gaussian Mixture Models. GMM and HMM are well suited for audio recognition. GMMs are often used due to their reduced computational costs whereas HMM allows a more refined analysis while taking into consideration the variation of the signal over time. In the following sub-sections, some theoretical backgrounds of these two techniques will be discussed.

3.7.1 Gaussian Mixture Models

The Gaussian mixture model is a probabilistic model for the computation of the probability density function *p* of a set of observed variables *o* using a multivariate Gaussian mixture density. A GMM is represented as a weighted sum of Gaussian distributions and is expressed by the equation below:

$$p(o \mid \lambda) = \sum_{j=1}^{J} w_j G(o : \mu_j, \Sigma_j) \tag{3.9}$$

where :

$p(o \mid \lambda)$ is the likelihood of an input observation *o* of dimension *D*.

J is the number of mixtures.

w_j represents positive weighting factors satisfying the constraint $\sum_{j=1}^{J} w_j = 1$

$G(o_j, \mu_j, \Sigma_j)$ denotes the jth Gaussian with a mean vector μ_j and covariance matrix Σ_j. It is given by the equation below :

$$G(o;\mu_j,\Sigma_j) = (2\pi)^{-\frac{D}{2}} \left|\Sigma_i\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(o-M_i)^T \Sigma_i^{-1}(o-M_i)\right\} \tag{3.10}$$

Given a set of observations inputs $\{O_1, O_2, ..., O_n\}$, GMM has been shown to accurately compute the continuous probability density functions $P = \{p_{ij}\}$ .The parameters of each distribution w_j, μ_j *and* Σ_j are estimated by using the Expectation-Maximization (EM) algorithm. Readers seeking more details about GMM should consult the paper of Reynolds (Reynolds et Rose, 1995).

During the training stage, and for each audio class defined, the parameters of each Gaussian model are computed from some sequence of training input observations by maximizing the likelihood.

During the classification or testing stage, an observation input is attributed to a specific class for which the likelihood is maximum.

3.7.2 Hidden Markov Models

HMM are used in most modern automatic speech recognition (ASR) systems. They provides an efficient framework for modelling time-varying spectral feature vectors (Gales et Young, 2008a). Different applications of HMM in statistical signal processing and acoustic modeling can be found in literature especially in speech and audio domains (Gales et Young, 2008a; Juang et Rabiner, 1991).

An HMM is defined by different set of parameters: number of hidden states, state transition probability distribution A, observation probability distribution B and initial state distribution π.

HMM model is denoted as λ= {A, B, π}

Considering a spectral sequence of observations $O=O_1, O_2, \ldots, O_T$, one can model the sequence of spectra by using Markov Chain.

$q = (q_0, q_1, ...q_t)$ q_t as the state of the system at time t, and N the number of states of HMM.

$$A = [a_{ij}] \text{ with } a_{ij} = \Pr(q_t = j \mid q = i) \quad 1 \leq i, j \leq N \tag{3.11}$$

The probability of q being generated by the Markov chain is given by the following equation:

$$\Pr(q \mid A, \pi) = \pi_{q_{11}} a_{q_0 q_1} a_{q_1 q_{12}} ... a_{q_{t-1} q_t} \tag{3.12}$$

For more details about HMM parameters estimation readers are referred to works (Juang et Rabiner, 1991; Rabiner et Juang, 1993b).

As main training function and to initialize HMM parameters, the Viterbi algorithm is used to find the most likely state sequence for each training input. The log likelihood of the training data is calculated and the process is repeated until no further increase in likelihood can be found. By applying the so-called Baum-Welch algorithm, the re-estimation of the HMM parameters is carried out. For testing an unknown observation, the probability of the observation generated by each class is computed and a decision is then taken based on the maximum probability obtained.

3.8 System evaluation

The aim of this work is to develop an automatic segmentation system with a low error rate. Figure 3.14 depicts an overview of the adopted methodology. It is based on three essential stages: Signal Decomposition, Features extraction and Classification. In this study, the efficiency of nine differently implemented systems listed below by varying approaches in each stage are evaluated:

1. FFT+FFT-MFCC+GMM
2. FFT+FFT-MFCC+4-states HMM
3. FFT+FFT-MFCC+5-states HMM
4. WPT+WE-DCT+GMM
5. WPT+WE-DCT+4-states HMM
6. WPT+WE-DCT+5-states HMM
7. EMD+EMD-MFCC+GMM
8. EMD+EMD-MFCC+4-states HMM
9. EMD+EMD-MFCC+5-states HMM

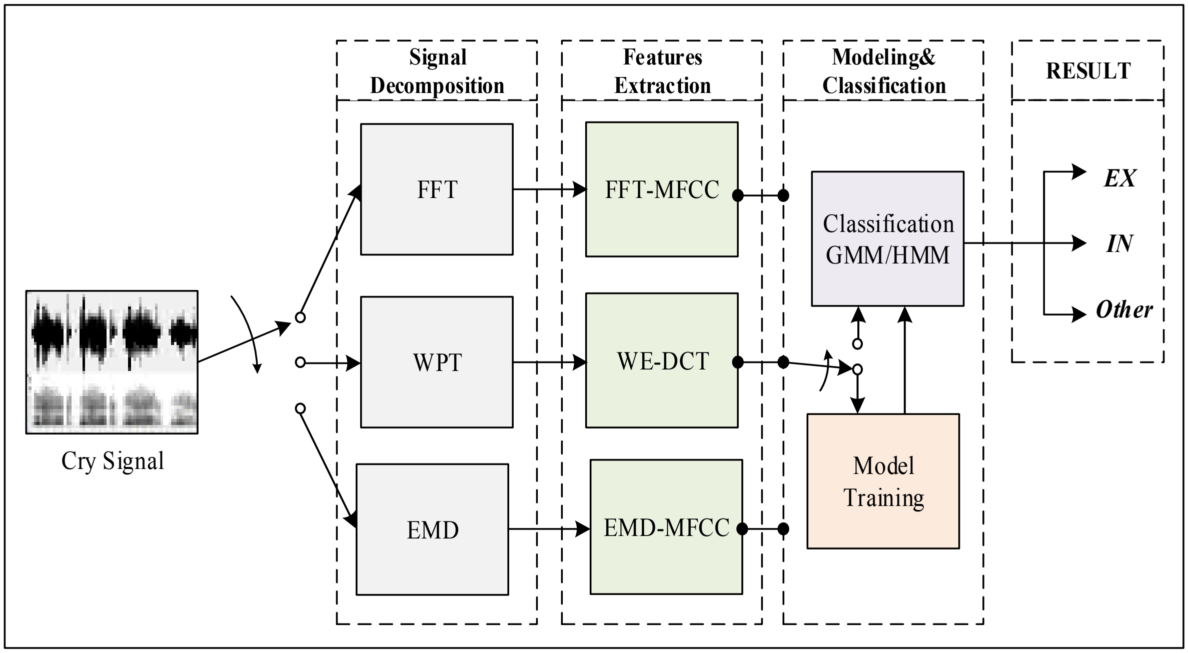


Figure 3.12 Overview of the different methodologies used in this work

GMM is compared to 4 and 5 states, left to right HMMs baseline system using multiple Gaussian mixtures with diagonal covariance matrices for each class. A varying number of mixtures per state from 16 to 64 Gaussians has been also considered. The efficiencies of the proposed systems are evaluated by comparing their performances to the FFT based system designed in the previous work (Abou-Abbas, Fersaie Alaie et Tadj, 2015).

Each frame was represented by a 13-dimensional feature vector. Two different window frame size 30ms and 50ms with an overlap of 30% are employed.

For both training and evaluation purposes, 507 cry signals used in this paper are manually labeled. The experiments were performed using the 10-fold cross validation. The whole database is divided several times into two parts: the first part has been used for training and the second for testing. The average duration of the corpuses used was shown in Table 3.2. The process of training and testing was repeated for each set of corpuses. To ensure reliable results, the average of the total classification error rate of same experiments repeated with different set of training and test corpuses was considered.

To evaluate the efficiency of the systems, the manual transcript files and the files generated at the front end of the system are compared. The performance of the designed systems is then calculated as shown below:

$$CER = 100 - \frac{Nb\ of\ Correctly\ Classified\ Segments}{Total\ number\ of\ Observation\ in\ the\ test\ Corpus} \times 100\% \qquad (3.13)$$

Where CER stands for Classification Error Rate.

Systems 2 and 3 based on FFT decomposition were considered in the previous work (Abou-Abbas, Fersaie Alaie et Tadj, 2015). Training and testing phases using the corpuses described in section 3 are re-executed. Table 3.3 summarizes the comparison between systems 1, 2 and 3 based on FFT decomposition.

Table 3.3 Classification Error Rates for a FFT-based extraction features

FFT_MFCC	30ms-21ms	50ms-35ms
GMM	8.98	15.99
4-states HMM	26.3	21.23
5-states HMM	23.4	17.29

It can be concluded that based on a FFT decomposition:

1. A GMM classifier outperforms the 4 and 5-states HMM classifiers;
2. A lower window size with GMM classifier gives better results than higher window size;
3. A HMM Classifier produces best results by increasing its number of states;
4. A Higher window size with HMM classifier presents best overall results than lower window size.

The obtained results are summarized in Figure 3.15. A GMM with 40 mixtures outperforms all experiments and gives a low classification error rate of 8.98%.

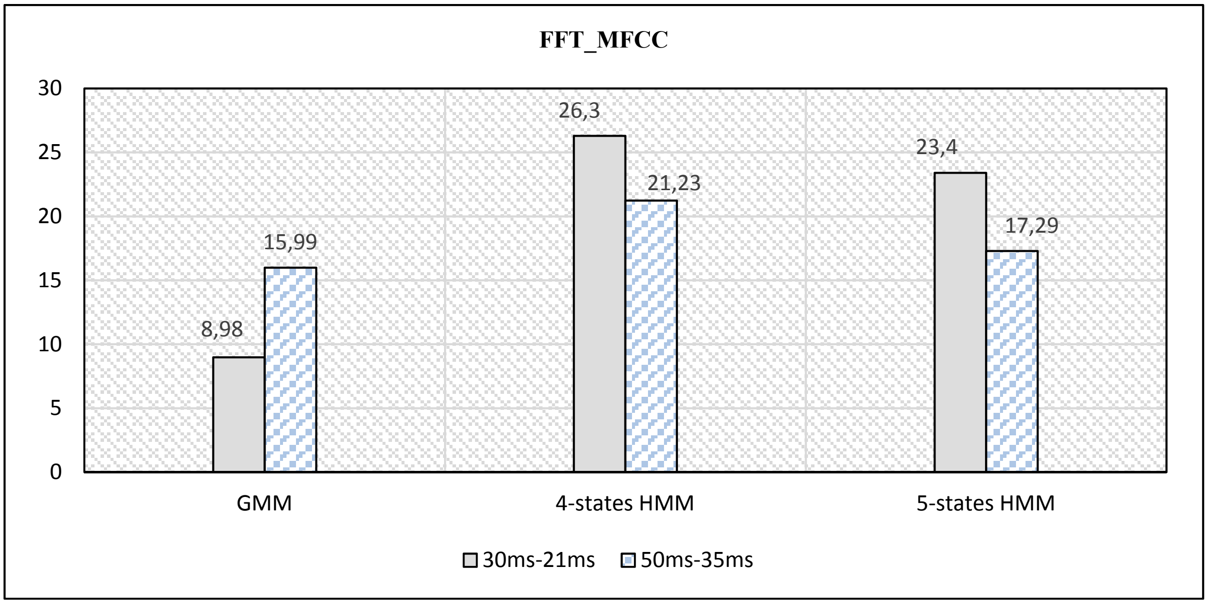


Figure 3.13 Comparison of CER between different classifiers for an FFT-based MFCC

Repeating experiments using the systems 4 to 6 yielded to the results indexed in Table 3.4 where wavelet Packet Transform was employed as decomposition method.

Different levels of decomposition such as 4, 5 and 6 are tried. The best results were obtained using 5 levels of decomposition. In this paper, therefore, only results obtained by 5-level are addressed.

From Table 3.4 and Figure 3.16, it can be concluded that using a Wavelet Packet decomposition:

1. A HMM classifier outperforms a GMM classifier;
2. A lower window size with either a GMM or HMM classifiers gives better results than higher window size;
3. A HMM classifier with 4 states outperforms a HMM classifier with 5 states.

The results obtained from the systems 4, 5 and 6 are shown in the chart in Figure 3.16. It is proved that lower classification error rate of 17.02 % is achieved using a 4-states HMM and a window size of 30ms.

Table 3.4 Classification Error Rates for a WPT-based extraction features

WE_DCT	**30ms-21ms**	**50ms-35ms**
GMM	22.2	29.75
4-states HMM	17.02	27.09
5-states HMM	21.17	27.42

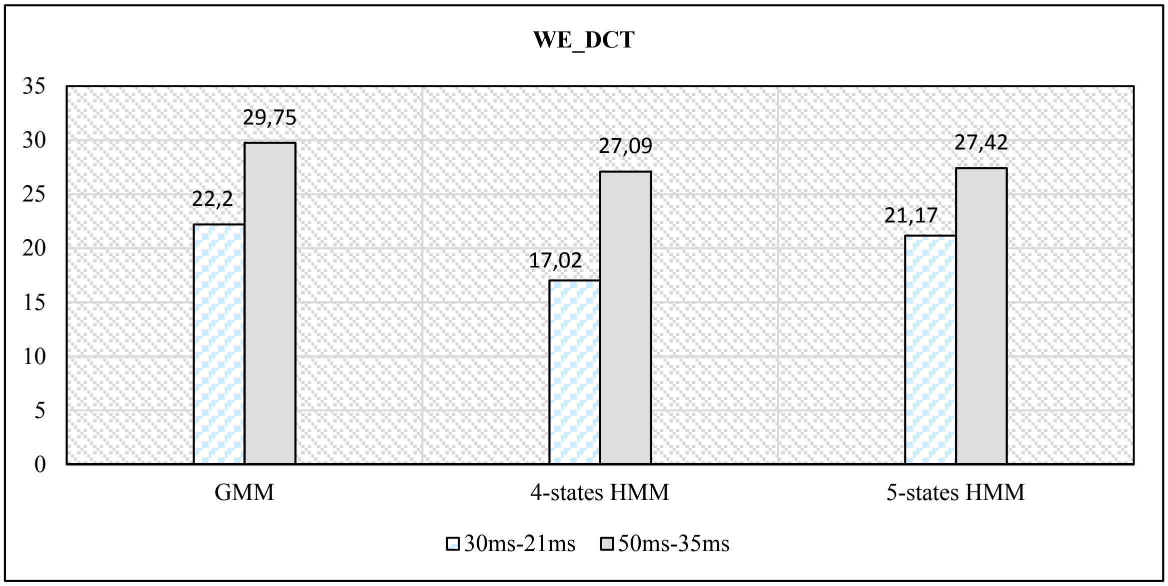


Figure 3.14 Comparison of CER between different classifiers for a WPT-based features

Using the EMD decomposition technique, 4 sets of different IMFs combinations are examined. These 4 sets are chosen based on results obtained from the experiments of the previous work (Abou-Abbas et al., 2015).

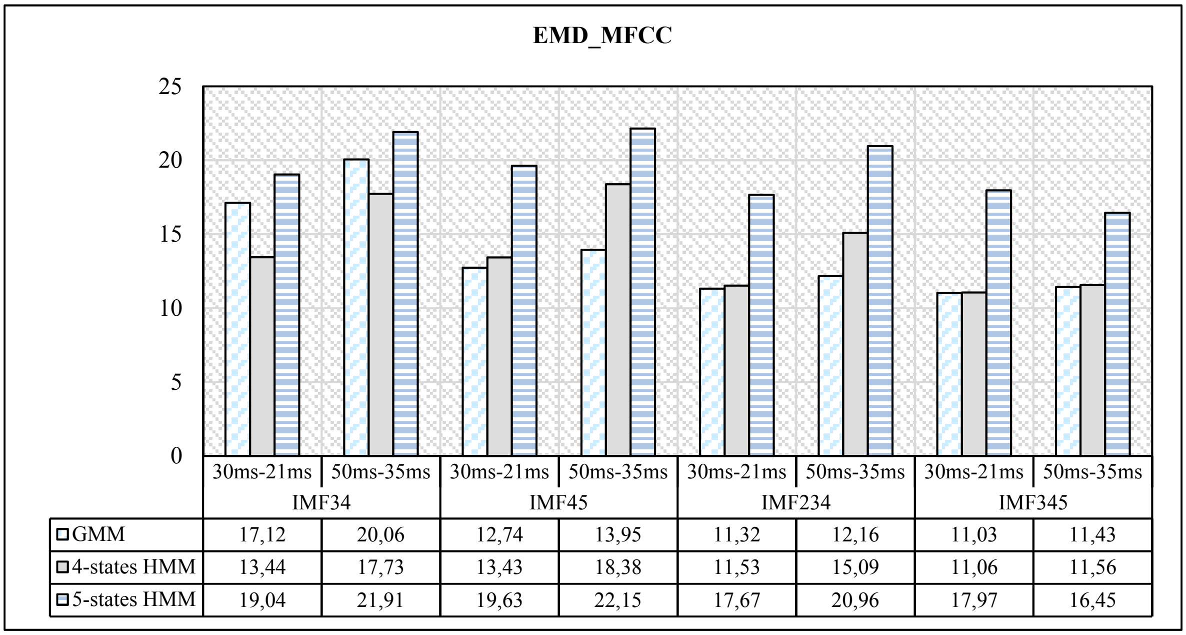
In Figure 3.17, it can be concluded that while using different combinations of intrinsic mode functions:

1. The parameters based on the combination of IMF3, IMF4 and IMF5 yielded the best results;

2. GMM classifier outperforms a HMM classifier in the set IMF45, IMF234 and IMF345;

3. 4-states HMM outperforms GMM classifier and 5-states HMM classifier while using the set IMF34;

4. Best results in the most classifiers are yielded using a lower window size.



	30ms-21ms	50ms-35ms	30ms-21ms	50ms-35ms	30ms-21ms	50ms-35ms	30ms-21ms	50ms-35ms
	IMF34		IMF45		IMF234		IMF345	
□GMM	17,12	20,06	12,74	13,95	11,32	12,16	11,03	11,43
□4-states HMM	13,44	17,73	13,43	18,38	11,53	15,09	11,06	11,56
□5-states HMM	19,04	21,91	19,63	22,15	17,67	20,96	17,97	16,45

Figure 3.15 CER of different classifiers used and different window sizes for EMD-based features

Table 3.5 CER of different features extracted and different classifiers for a win size of 30 ms

Features/Classifier-30ms	GMM	4-states HMM	5-states HMM
FFT_MFCC	8.98	26.3	23.4
WE_DCT	22.2	17.02	21.17
IMF34	17.12	13.44	17.73
IMF45	12.74	13.43	18.38
IMF234	11.32	11.53	15.09
IMF345	11.03	11.06	11.56

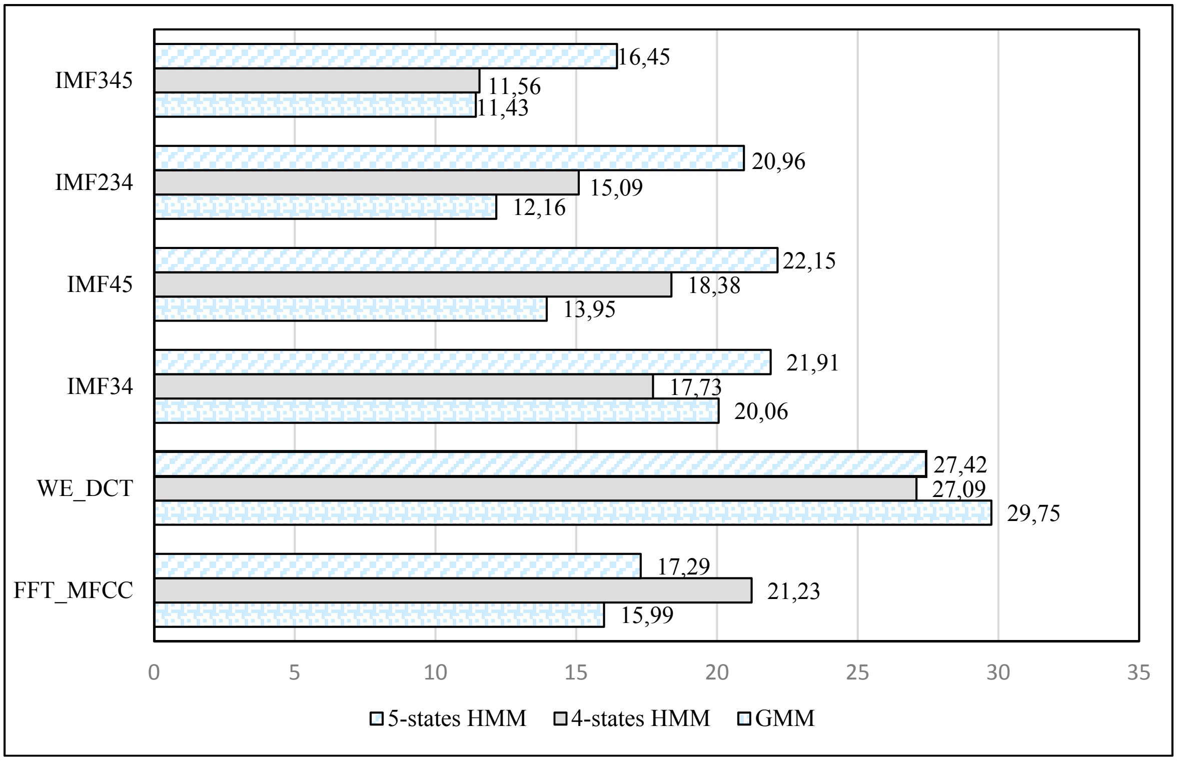


Figure 3.16 Comparison between CER of different features extracted and different classifiers for a window size of 30 ms

It can also be seen from Figure 3.17 that the features represented by the so-called IMF345 which is the combination of IMF3, IMF4, and IMF5 yielded the lowest error rate of 11.03% using again a GMM classifier and a window size of 30ms. Table 3.5 and Figure 3.18 compare the proposed systems in terms of features and classifiers.

It can be seen in Table 3.5 that best results are yielded using the features obtained based on FFT decomposition and using a GMM classifier while using a window size of 30ms. The minimum obtained classification error rates while employing a 50ms of window size are marked by using EMD decomposition combining IMF3, 4 and 5 and GMM classifier to reach a classification error rate of 11.43%. The results are demonstrated in Table 3.6 and Figure 3.19.

Table 3.6 CER of different features extracted and different classifiers for a win size of 30ms

Features/4-states HMM-50ms	GMM	4-states HMM	5-states HMM
FFT_MFCC	15.99	21.23	17.29
WE_DCT	29.75	27.09	27.42
IMF34	20.06	17.73	21.91
IMF45	13.95	18.38	22.15
IMF234	12.16	15.09	20.96
IMF345	11.43	11.56	16.45

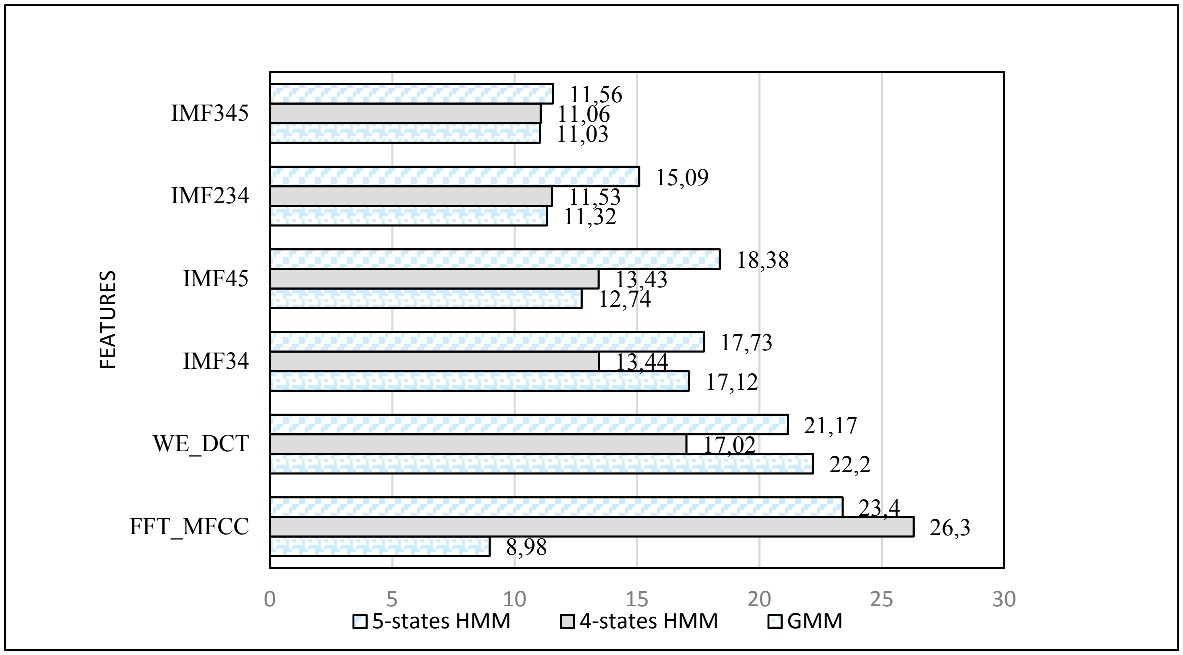


Figure 3.17 Comparison between CER of different features extracted and different classifiers for a window size of 50 ms

In order to compare the performance of all examined systems in this paper, Table 3.7

summarizes the best error rate obtained by varying different parameters. Analyzing these results, the following conclusions can be outlined:

1. The system number 1: FFT+FFT-MFCC+GMM performed the best among the nine proposed systems, by giving an average Class Error rate of 8.98% for various training and testing datasets;

2. The next is the system number 7: EMD+EMD-MFCC+GMM that achieved an error rate of 11.03% by using a combination of IMF345;

3. It can also be observed that results are the best for the GMM-based classification method in the case of FFT and EMD decompositions and for the 4-states HMM in the case of WPT decomposition;

4. For the FFT decomposition with HMM, best results are reached by increasing the number of states and the window size.

Table 3.7 The best CER obtained for the different systems implemented

System	**Decomposition technique**	**Features Extraction**	**Classification method**	**Best Error Rate %**
1	FFT	FFT-MFCC	GMM	8.98
2	FFT	FFT-MFCC	4-states HMM	21.23
3	FFT	FFT-MFCC	5-states HMM	17.29
4	WPT	WE-DCT	GMM	22.2
5	WPT	WE-DCT	4-states HMM	17.02
6	WPT	WE-DCT	5-states HMM	21.17
7	EMD	EMD-MFCC	GMM	11.03
8	EMD	EMD-MFCC	4-states HMM	11.06
9	EMD	EMD-MFCC	5-states HMM	11.56

3.9 Conclusion

Newborn cry signals provide valuable diagnostic information concerning their physiological and psychological states. In this paper, EMD-based and wavelet-based architectures have been examined for the purpose of automatic expiratory and inspiratory episodes detection under the scope of designing a complete automatic Newborn Cry-based Diagnostic System. The methodology employed in this research is based on three phases: Signal Decomposition, Features extraction as well as modeling and Classification. Different approaches at each phase have been addressed to implement in total 9 different segmentation systems. Three signal decomposition approaches were compared: Fast Fourier Transform, Wavelet Packet Decomposition, and Empirical Mode Decomposition. WPT is applied to capture the more prominent features in high and intermediate frequency bands for the segmentation purpose and is compared to intrinsic mode functions resulted from EMD decomposition. GMM classifier is also compared to 4 and 5 states, left to right HMMs baseline system using multiple Gaussian mixtures with diagonal covariance matrices for each class. For training and evaluating the proposed systems, cry signals recorded in various environments are used: this dataset includes 507 cry signals with average duration of 90 seconds from 207 babies. To insure the liability of results, the ten-fold technique is carried out: 90 % of the data corpus was randomly chosen for the training stage and the rest 10% for the testing stage while repeating experiments for several times. The effect of different window sizes and different extracted features have been examined. The main goal of this research was to measure the ability of the system in classifying audible cries: expiration and inspiration. Results presented in this study show that best results were obtained by using GMM classifier with the low error rate of 8.9%. Future direction of research may include a post-processing step in the systems designed based on some spectral and temporal approaches in order to reduce the error rates and increase the performance of the system.

Acknowledgment

This work is supported by Bill and Melinda gates foundation. The authors thank the personnel at the Department of neonatology at Sainte Justine hospital and Sahel Hospital for their cooperation in the data collection process.

CHAPITRE 4

A FULLY AUTOMATED APPROACH FOR BABY CRY SOUNDS SEGMENTATION IN A REALISTIC CLINICAL ENVIRONMENT AND BOUNDARY DETECTION OF CORRESPONDING EXPIRATORY AND INSPIRATORY EPISODES

L. Abou-Abbas [a], C Tadj [b] and H. Fersaie Alaie [c]

[a,b,c] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montréal, QC, Canada H3C1K3

This paper has been submitted for publication
in *The Journal of Acoustical Society of America*, June 2016.

4.1 Abstract

The detection of cry sounds is generally an important pre-processing step for various applications involving cry analysis, such as diagnostic systems, electronic monitoring systems, emotion detection, and robotics for baby caregivers. Given its complexity, an automatic cry segmentation system is a rather challenging topic. In this paper, a new framework for automatic cry sound segmentation for application in a cry-based diagnostic system has been proposed. We studied the contribution of various additional time- and frequency-domain features to increase the robustness of a GMM/HMM-based cry segmentation system in noisy environments. We introduced a fully automated segmentation algorithm to extract cry sound components, namely, audible expiration and inspiration, based on two approaches: statistical analysis based on Gaussian Mixture Models (GMM) or Hidden Markov Models (HMM) classifiers and a post-processing method based on intensity, zero crossing rate, and fundamental frequency feature extraction. The main focus of this paper is to extend the systems developed in our previous works to include a post-processing stage with a set of corrective and enhancing tools to improve the classification performance. This full approach allows us to precisely determine the start and end points of the expiratory and inspiratory components of a cry signal, EXP and INSV in any given sound signal. Experimental results have indicated the effectiveness of the proposed solution. EXP and INSV detection rates of approximately 94.29% and 92.16%, respectively, were achieved by applying a 10-fold cross-validation technique to avoid over-fitting.

Keywords: babies' cries segmentation, time domain features, fundamental frequency, MFCC, GMM/HMM classification, Empirical Mode decomposition.

4.2 Background

Cry signals have been the object of research and analysis for many years. Researchers have found sufficient evidence that cry signals can provide relevant information about the physical and psychological states of newborns (Barr, 2006; Farsaie Alaie, Abou-Abbas et Tadj, 2016; Golub, 1979; Soltis, 2004). Over the years, a considerable number of works have been conducted in the field of infant cry analysis. (Orozco-García et Reyes-García, 2003a), (Várallyay, 2006), (Hariharan, Sindhu et Yaacob, 2012), and (Reyes-Galaviz, Cano-Ortiz et Reyes-Garcia, 2008) described efficient classification algorithms for distinguishing cries of normal infants from those of Hypo acoustic infants, and as a result accuracy rates ranging from 88 to 100% were obtained. The accuracy rates for classification tasks between healthy infants and infants with asphyxia, as performed by (Sahak et al., 2010) and (Zabidi et al., 2010), were reported to be 93.16% and 94%, respectively. Moreover, anger, pain, and fear detection from cry signals were carried out by (Petroni et al., 1995) yielding a recognition rate of 90.4%.

A Newborn Cry-based Diagnostic System (NCDS) aims to achieve preliminary screening of newborn pathologies by analyzing the features of audible cry components detected in a realistic clinical environment.

According to the World Health Organization*, "every year, approximately 40% of child deaths are deaths of newborn infants in their first 28 days of life, 75 % of newborn deaths occur in the first week of life, and up to two-thirds of newborn deaths can be prevented if known."*

Therefore any technique that can contribute in identifying the very first signs of newborn diseases could have a great influence on decreasing infant mortality. Specifically, this is the main goal of our project: to develop a fully automatic noninvasive system able to diagnose diseases based solely on cry sound analysis. The implementation of such a diagnostic system first addresses the issue of finding the useful cry components in an input waveform. The NCDS may suffer in terms of intelligibility if the input audio file contains acoustical activities other than crying. Therefore, one of the challenges in implementing such a system is to create an automatic segmentation system to correctly locate the expiratory and inspiratory phases of a cry sequence. Despite the significant amount of research conducted on pathological cry signal classification, little has been done to address the problem of automatic segmentation of audible expiratory and inspiratory phases of crying. If we could automatically segment and identify important parts of a given recorded signal, it would be easier to develop a fully automatic diagnostic system. Such a system could hopefully be used as a real-time clinical decision support tool, and in the case of early detection of a symptom, the necessary treatment could be provided easily and cheaply.

This paper is organized as follows. Section 3 describes the corpus. Section 4 details our proposed solution and describes the post-processing stage. The results and discussion are presented in Section 5, and finally, Section 6 concludes the paper by summarizing the main contribution of this work and describing future work.

4.3 Review of literature

The main components of a cry sound are expiration and inspiration segments with vocalization, audible expiration and inspiration (EXP and INSV, respectively). The main challenge of this work is to develop a method able to locate EXP and INSV correctly within a given audio signal. The problem of cry segmentation/detection cannot be considered a problem of voiced/unvoiced detection because a single typical audible cry can contain both voiced and unvoiced segments. Alternately, the problem of cry detection in a corpus recorded in a very noisy clinical environment cannot be solved simply by traditional voice activity detection (VAD) modules, which are common in previous cry analysis systems proposed in the literature (Kuo, 2010; Rui et al., 2010; Várallyay, 2006; Zabidi et al., 2009b) . VAD refers to the issue of locating speech regions from other acoustic activity regions in a given audio signal. The other acoustic activity regions can be any type, such as noise, silence or an alarm warning. However, the signal-to-noise ratio (SNR) represents a crucial parameter and may result in major errors. VAD is an essential part of various audio communication systems, such as automatic speech recognition, mobile phones, personal digital assistants, and real-time speech transmission. Common VAD methods are composed of two main modules: (1) feature extraction and (2) decision rules.

Common features are those that depend on energy calculation of the signal (Kyoung-Ho et al., 2000), cepstral coefficients, zero-crossing rate (ITU, 1996), spectrum analysis (Marzinzik et Kollmeier, 2002), entropy and wavelet transforms. The decision rule is often formulated on a frame-by-frame basis and simple thresholding rules. In general, no specific parameter has been proven to perform well under varying background conditions.

We applied well-known VAD algorithms used in G.729b and the Rabiner-Sambur method to detect cry segments (Benyassine et al., 1997; Rabiner et Sambur, 1975).

We found that:

1. Threshold settings were not easy to select in a variable and noisy environment;

2. Traditional VAD could not distinguish between important cry segments (EXP and INSV) and speech segments recorded during the data acquisition;

3. Traditional VAD failed in distinguishing expiration phases from inspiration phases of cry signals.

To solve the issue of adjusting thresholds, statistical approaches seem to be a good solution. That is why we have given due consideration in our recent and present works to statistical model-based approaches (Abou-Abbas et al., 2015; Abou-Abbas, Fersaei Alaei et Tadj, 2015; Abou-Abbas, Fersaie Alaie et Tadj, 2015).

Like any audio recognition/detection system, a cry segmentation system can be briefly analyzed using three main modules:

1. Feature selection and extraction, where suitable information is estimated in a relevant form and size from a cry signal to represent it in a different and more convenient domain;

2. Classification, where representative models are created and adapted using extracted feature vectors for each available pathology class;

3. Decision making.

Selecting adequate features is the key to guaranteeing robust system performance estimation. Cry signals can be described by their features within two common domains: (1) the time domain and (2) frequency domain.

From either of the mentioned domains, a number of significant characteristics can be extracted (Scherer, 1982).

Compared to other audio-related fields, such as speech and music, investigation of cry segmentation has seen low consideration. We refer readers to the state of the art discussed in our recent works (Abou-Abbas et al., 2015; Abou-Abbas, Fersaei Alaei et Tadj, 2015).

Thus far, existing cry segmentation algorithms have mainly been able to separate cries from silent pauses or respiratory phases. (Várallyay, Illényi et Benyó, 2009; Várallyay Jr, Illényi et Benyó, 2008). Noisy backgrounds or other acoustic activities have not been considered in these works because the database used was collected in a laboratory environment and not in a real clinical environment.

Cry segmentation or detection has also been studied in (Kim et al., 2013) (Yamamoto et al., 2010) (Yamamoto et al., 2013).The work of Jong Kim et al. investigated a new feature called segmental two-dimensional linear frequency cepstral coefficients (STDLFCC), which is based on linear frequency cepstral coefficients (LFCC). The idea behind it was to capture the lower frequency as well as the higher frequency within long-range crying segments and provide better discrimination between crying and non-crying segments. An average equal error rate of 4.42% was reported in their article.

In (Yamamoto et al., 2010; 2013), a method for detecting a baby's voice using a well-known speech recognition system called JULIUS and fundamental frequency analysis was introduced achieving a detection rate of 69.4%.

We recently developed different approaches for cry segmentation (Abou-Abbas et al., 2015; Abou-Abbas, Fersaie Alaie et Tadj, 2015). The general structure of the systems designed can be described with a block diagram, as shown in Figure 4.1.

We used machine-learning methods, such as Hidden Markov Models and Gaussian Mixture Models, which have been proven to work well in the acoustic domain. Different feature extraction techniques were employed and compared, reaching an accuracy rate of up to 91%. These methods were validated on our available signals, which were recorded in a clinical environment through a 10-fold cross-validation technique to perform training and testing operations of dissimilar sets of the total data.

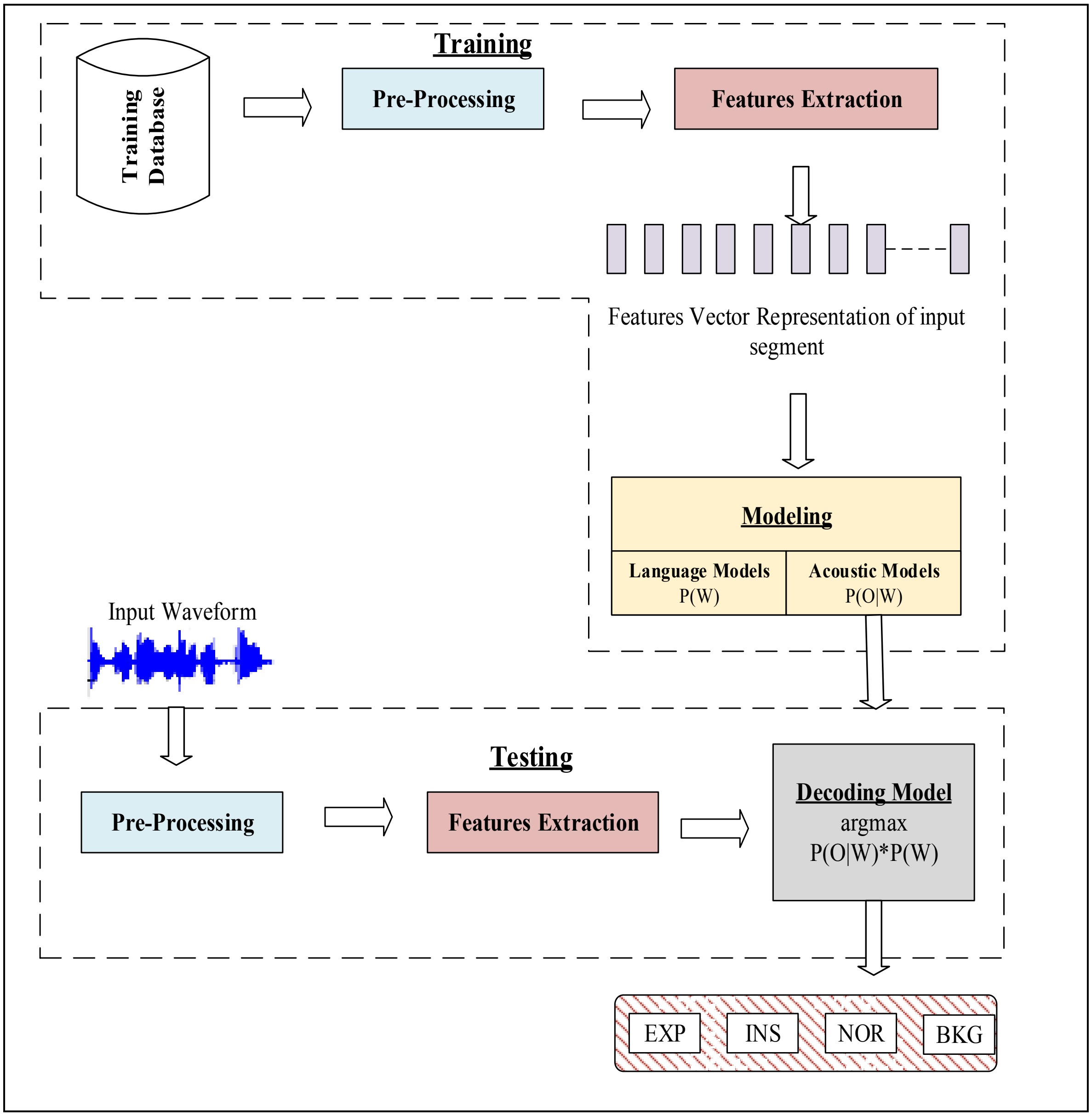


Figure 4.1 Architecture of a supervised cry segmentation system

However, there was a problem in precise boundary detection of the classified segments. This misdetection affects the accuracy rate of the overall system because the results of classification were compared to our manually segmented signals, which were labeled by trained colleagues. The exact cry length and pause length between cries, although commonly overlooked by cry sound classification approaches, represent important parameters containing useful related information. For example, researchers in (Várallyay, 2006) used cry duration as an additional feature to detect the gender of the baby, and in (Aucouturier et al., 2011), the duration of the expiration between contexts of cries, such as hunger, peeing, and sleepiness were investigated. Significant differences between durations have been observed. (Lind et al., 1970) proved that cry durations are higher than usual in the cries of infants with Down's syndrome.

In this paper, we developed a method for improving the accuracy rate of our previous systems based on a statistical approach via a post-processing step based on both temporal and frequency-domain features.

Our method aimed to obtain more reasonable and accurate segmentation results by automatically applying additional steps. Experiments were conducted to evaluate the utility of the post-processing stage and the use of additional temporal and frequency-domain information as acoustic features. The obtained results show that the use of our proposed approach for automatic segmentation was able to provide a major contribution to the performance of a cry segmentation system.

Our system uses a cascaded architecture to create an efficient cry segmentation system with a low false-positive rate while keeping the true-positive rate as high as possible.

4.4 Corpus of infants' cries

Most of the subjects recruited for this study were newborns that were just a few days old. The size of our database and the diversity presented in our collected data distinguish our work from other aforementioned research studies:

- signals collected in hospital environments at different times and in different situations, such as: after birth, first bath, medical care, neonatal intensive care unit, and private or public maternity rooms with parents or visitors;
- signals from crying babies with different contexts/causes, such as pain, hunger, fear, and birth;

- cry signals of babies with different pathological conditions, such as normal infants and babies suffering from respiratory diseases, blood diseases, neurological diseases, or heart diseases.

In Table 4.1, you can find a brief description of our cry database:

Table 4.1 Description of our cry database

Gender	Both male and female
Prematurity	Both preterm and full term
Babies Ages	1 to 53 days old
Weight	0.98 to 5.2 Kg
APGAR Result	0 to 10 measured 2- 3 times: once at the birth, then 1, 5, 10 min after of the baby's birth
Gestational age	27 weeks and 2 days up to 41 weeks and 4 days
Origin	Canada, Haiti, Portugal, Syria, Lebanon, Algeria, Palestine, Bangladesh, Turkey
Race	Caucasian, Arabic, Asian, Latino, African, Native Hawaiian, Quebec
Reason of crying	Birth cry, hunger, dirty diaper, discomfort, needs to sleep, cold, pain, tummy troubles (colic, reflux,etc)
Health condition	Healthy, heart diseases, respiratory diseases, neurological diseases, blood diseases, and other.

The most accurate reason for crying was determined with the help of nurses and newborns' parents according to the situations that caused the infants to cry. The health condition was established by the doctor who examined the newborn and was based on different tests that had been performed in the hospital after the birth.

For recording purposes, an Olympus hand-held digital 2-channel recorder was used and placed 10 to 30 cm from the newborn's mouth to be effective at a sampling frequency of 44.1 kHz and a sample resolution of 16 bits. There was no well-defined procedure during the acquisition of the cry sounds. Therefore, unwanted noises, cross-talk and clinical environment sounds were also recorded during the data collection process. For this reason, we consider our database a real corpus recorded in a real clinical environment.

Using the Wave Surfer software, we manually annotated recordings of our corpus. Annotations identified the start and end points of each vocalization. The boundaries of each annotation were fixed by the point where the sound cannot be heard anymore. A newborn cry can comprise typical cry sounds, glottal sounds, hiccups, short pause segments between cries and faint cries. The labels were chosen according to the different types of acoustic activities available in our corpora. The labels were defined as follows:

1. Expiration (EXP), which represents the total vocalization occurring during one expiration of a cry, as well as the sound between two inspirations (Lewis, 2007). It is composed of voiced or/and unvoiced segments.

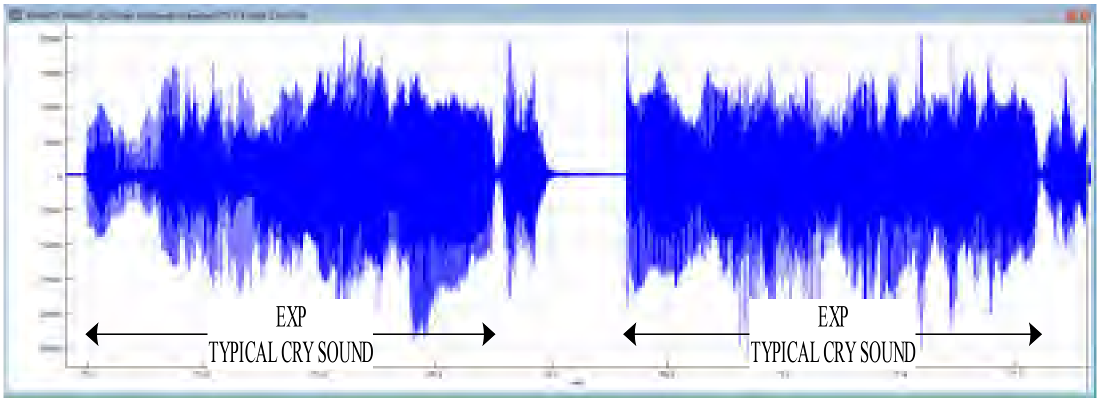


Figure 4.2 A waveform of a typical cry sound

It is represented by three types: a typical cry sound (see Figure 4.2), a glottal sound, i.e., a cough sound (see Figure 4.3) or a spasmodic sound (see Figure 4.4). We can easily distinguish typical expiration sounds from other expiration sounds by comparing their durations;

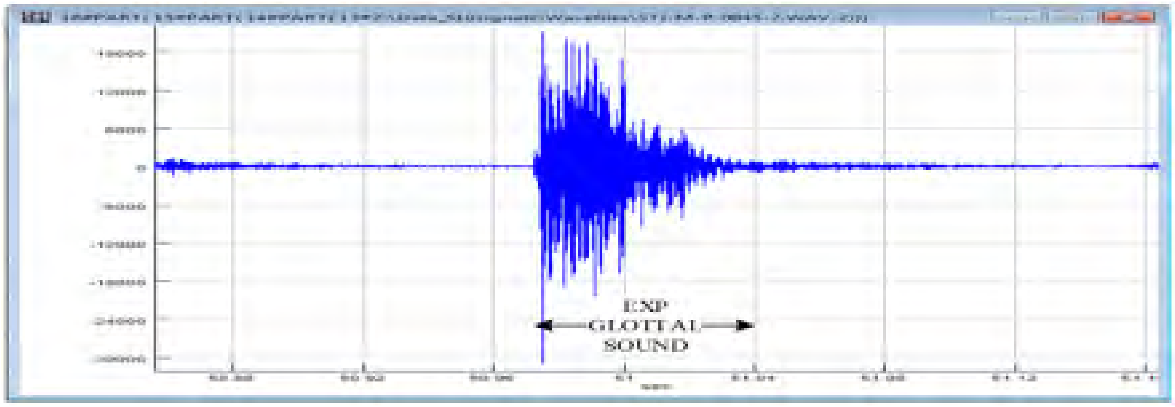


Figure 4.4 A waveform of a spasmodic cry

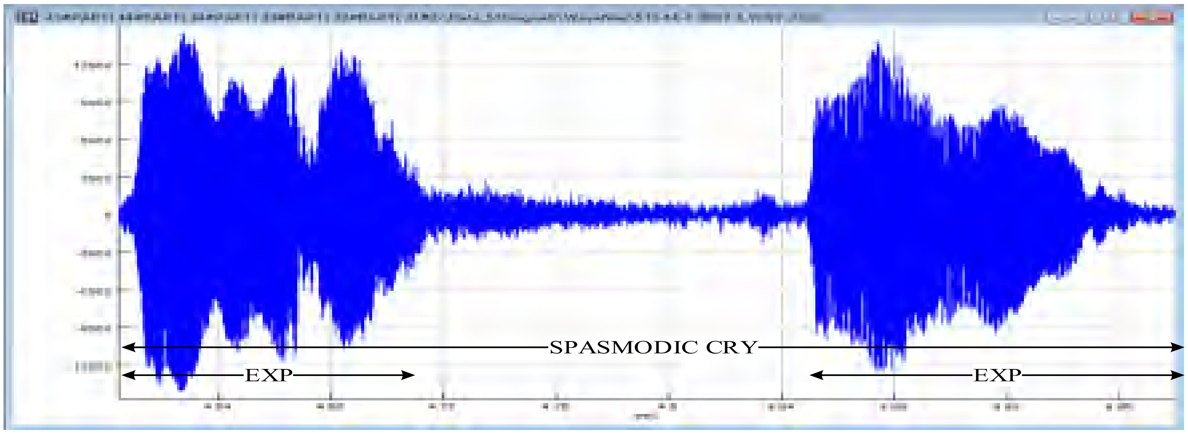


Figure 4.3 A waveform of a glottal sound during cry

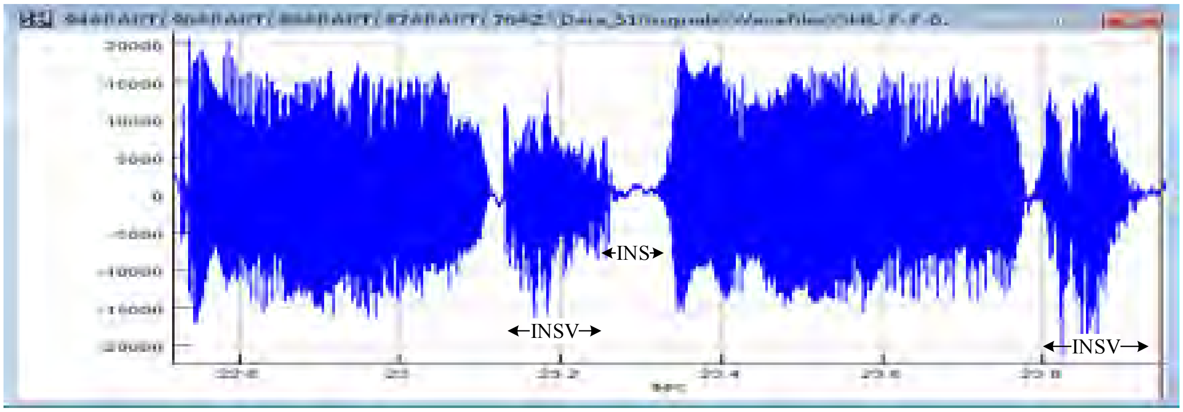


Figure 4.5 Waveform representing INSV and INS labels

2. Inspiration (INSV) with vocalization also called Hiccups. It is the total vocalization occurring during one inspiration, and it usually follows a long Expiration and is followed by a short pause segment. We noticed the presence of INSV in sick babies' cries more than those of healthy babies (see Figure 4.5);

3. Silent inspiration (INS), also called a short pause segment. It commonly happens between one INSV and the next EXP or directly after one EXP to be followed by another EXP (see Figure 4.5 and Figure 4.6) ;

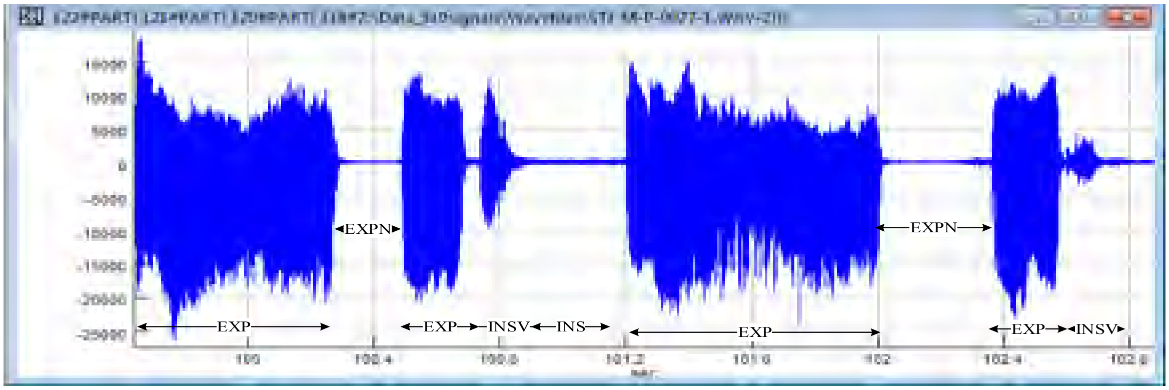


Figure 4.6 A cycle of a cry that contains EXP, INSV, INS, and EXPN

4. Expiration with non-vocalization (EXPN), also called a faint cry. It happens commonly after a long cry and always between two EXP or EXP and INSV (See Figure 4.6) ;

5. EXP2 and INS2 represent components of vocalization during expiration and inspiration, respectively, and are produced by infants during babbling, cooing, etc. We could easily distinguish them from the EXP and INSV of the cry by their short durations (see Figure 4.7) ;

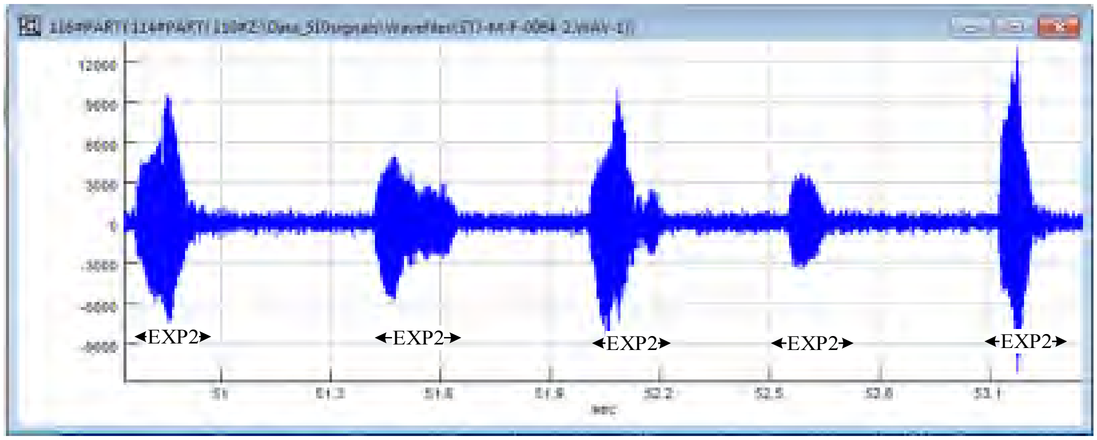


Figure 4.7 Babbling period after or before a cry

6. Speech (SP) of the medical staff or the parents around the baby. In general, speech signals have a low pitch range of 100 to 300 Hz (see Figure 4.8);

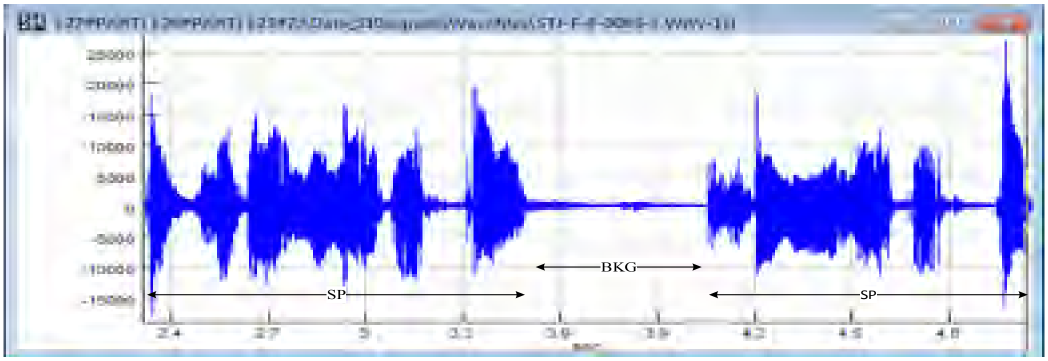


Figure 4.8 A period of human speech in a cry signal

7. BIP is the sound of medical machines around the baby. It is characterized by a constant fundamental frequency;

8. NOR or Noises represent different types of sounds that can be either outside the cry or during the cry ;

9. BKG or background silence between cries.

Our database in this work consists of a total of 507 waveforms of cry sounds interspersed by different unwanted acoustic activities. A summary of the database is given in Table 4.2 and Table 4.3, and it is the same database as used in our previous work (Abou-Abbas et al., 2016) for comparison purposes.

Table 4.2 Corpus statistics

			Number of babies	**Number of signals**
Female	Full term	Healthy	56	141
		Pathological	34	94
	Preterm	Healthy	20	23
		Pathological	17	49
Male	Full term	Healthy	4	11
		Pathological	54	146
	Preterm	Healthy	5	11
		Pathological	13	32
Total			**203**	**507**

To divide the dataset into training and testing corpuses, the 10-fold cross-validation technique was applied.

Table 4.3 Data used for training and testing corpuses

Classes	**Time in sec**	**Average time for training corpus/sec**	**Average time for testing corpus/sec**
Expiration	21414	19348	2066
Inspiration	2154.8	1930	224.8
Background	5683.4	5228.1	455.3

4.5 Proposed approach

In this paper, a new approach for cry segmentation was proposed to correctly and precisely classify cry components. In brief, before the final decision, an initial classification was performed. Then, a post-processing stage was added for two reasons:

1. To minimize switching errors between the two predefined classes EXP and INSV;

2. To adjust the start and end points of each utterance.

In any acoustic recognition problem addressed in the literature, the following two concepts of signal processing were applied:

1. Segmentation: first, a detector is used to segment the audio signal;

2. Recognition: then, features are extracted and used to create a model to provide information about the active segments and take the best decision defined in the corresponding application.

The problem that arises from applying this traditional methodology concerns the adjustment of a static threshold, which should depend on the environmental conditions. Unlike this traditional methodology, we relied on frame-by-frame results obtained from standard classifiers to achieve a better segmentation by finding the exact boundaries of useful acoustic components.

A number of features have been extracted, including MFCC, energy, intensity, zero crossing rate, and fundamental frequency. We divided the proposed segmentation approach into two separate blocks. In the first block, we chose HMM and GMM classifiers to model and classify each label (EXP/INSV/BKG and NOR, which represents all other acoustic activities, including noise). In the second block, a post-processing step followed the classification task. In this section, we give a general overview of the cry segmentation system and describe its different elements. Our goal in this work was to automatically detect cries, expiration, and audible inspiration segments for a given recorded signal containing different acoustical activities, such as crying, speech, silence, and different types of noises from common hospital environments. A block diagram of the main algorithm is depicted in Figure 4.9.

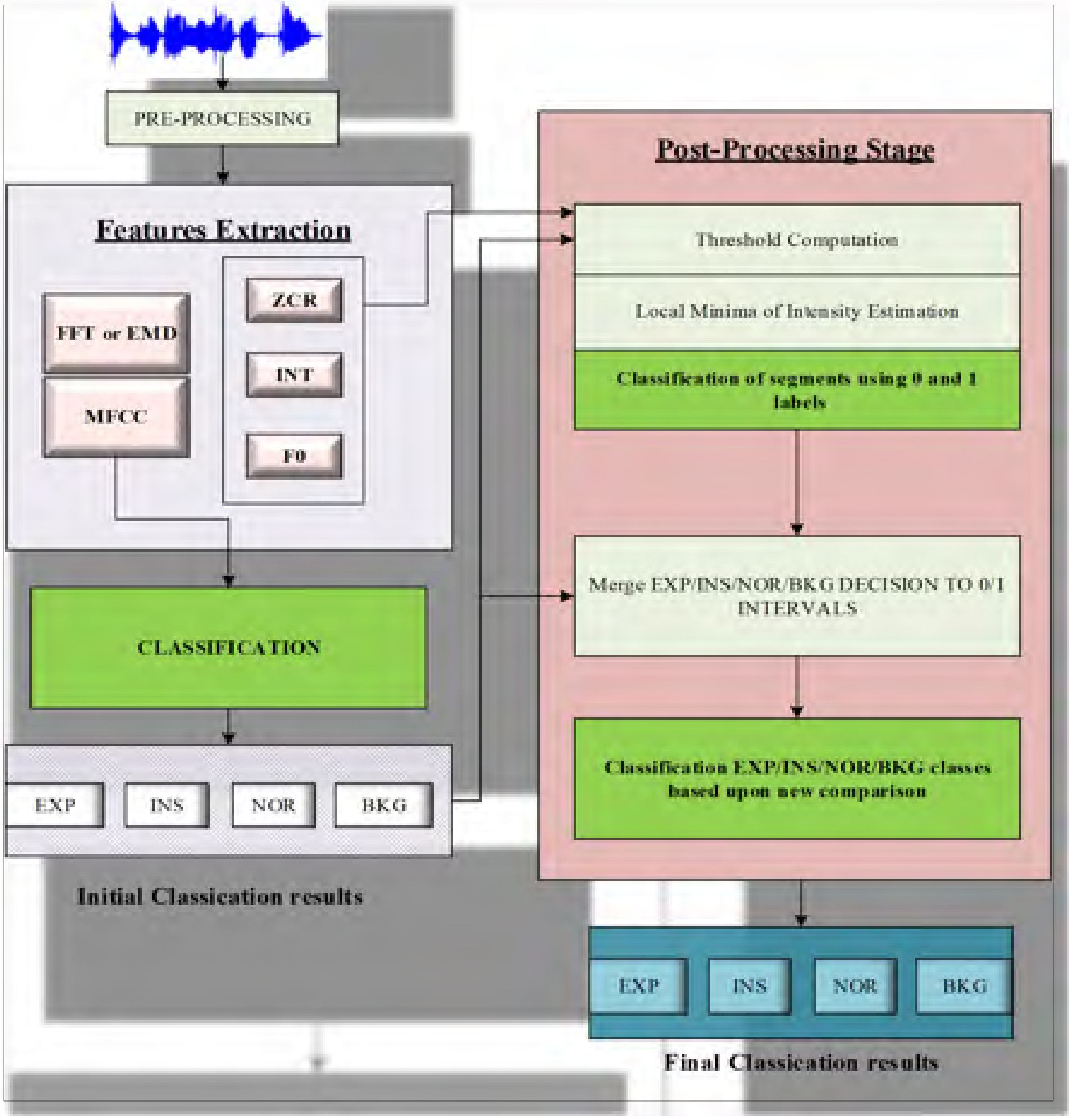

PRE-PROCESSING
Features Extraction
FFT or EMD
MFCC
ZCR
INT
F0
CLASSIFICATION
EXP
INS
NOR
BKG
Initial Classication results
Post-Processing Stage
Threshold Computation
Local Minima of Intensity Estimation
Classification of segments using 0 and 1 labels
Merge EXP/INS/NOR/BKG DECISION TO 0/1 INTERVALS
Classification EXP/INS/NOR/BKG classes based upon new comparison
EXP
INS
NOR
BKG
Final Classication results

Figure 4.9 Architecture of the proposed method

4.5.1 Pre-processing stage

It is essential to have a Pre-Process step as the first step of any audio analysis system. The block diagram of our designed pre-processing stage is depicted in Figure 4.10 and contains four modules:

1. The input audio signal was converted to mono by calculating the average of both channels' audio files;

2. A high-pass filter was applied to emphasize the signal. The most common filter is $1-0.97z^{-1}$;

3. A framing module converted the continuous audio signal into overlapped frames based on frame size and overlap percentage;

4. A hamming window was used to avoid the aliasing effect.

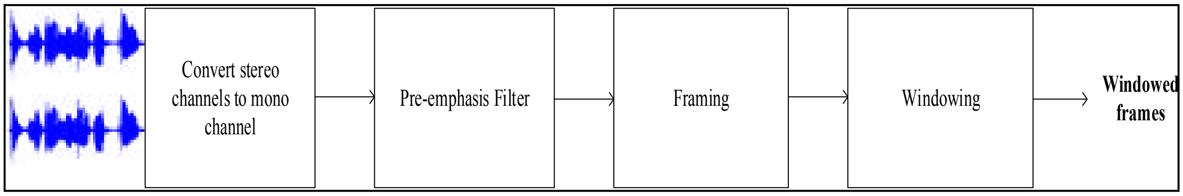


Figure 4.10 Scheme of pre-processing stage

4.5.2 Feature extraction procedure

A wide range of features can be extracted from the cry signals, but they do not necessarily correlate with a newborn's health condition. The selected feature set used in this work consisted of two parts: 39 features were used in the first classification stage (explained in section 4.5) to provide initial results, and another five features (zero crossing rate, intensity, and fundamental frequency (minimum, maximum, and mean)) were used in the post-processing stage to give the final results. In this section, we describe different domain features applied at different levels. The time / frequency-domain features that we employed are listed in Table 4.4.

Table 4.4 Set of features employed in the proposed method

Domain	**Features**
Time	Zero crossing rate (ZCR) Sound Intensity level (I)
Frequency	Min f0 Max f0 Mean f0
Time- Frequency	FFT-MFCC EMD-MFCC

4.5.2.1 Time-domain features

Intensity

The intensity of a sound, which is also called the loudness, is related to the amplitude of the signal. It represents the amount of energy a sound has per unit area. The intensity is defined by the logarithmic measure of a signal section of length N in decibels as follows:

$$I = 10\log\left(\sum_{n=1}^{N} s^2(n)w(n)\right) \tag{4.1}$$

where w is a window function.

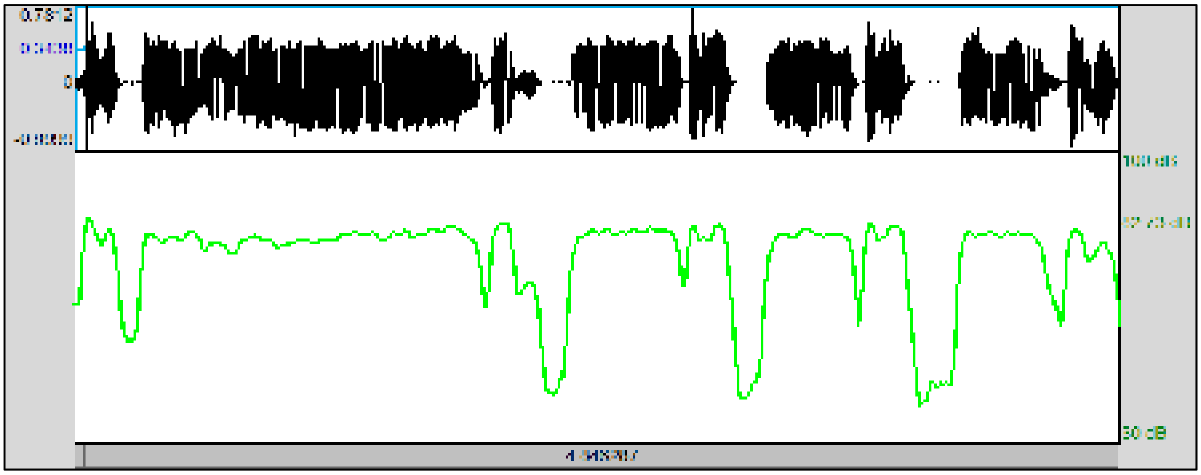


Figure 4.11 Example of an intensity contour in Praat software which is calculated on a consecutive expiration and inspiration with vocalization followed by a pause period

Intensity is an essential feature widely used in different applications, such as music mood detection (Liu, Lu et Zhang, 2003; Lu, Liu et Zhang, 2006), and an accuracy rate of 99% is achieved, proving the good performance of intensity features.

Considering Figure 4.11, one can note that the intensity was increased considerably during the period of cry segments.

Zero crossing rate (ZCR)

It is the one of the most widely used time-domain features in voice activity detection algorithms (Bachu et al., 2010; Rabiner et Sambur, 1975). Zero crossing occurs when consecutive samples have different algebraic signs as defined by (Shete, Patil et Patil, 2014). Therefore, ZCR represents the number of times in a specific frame the amplitude of the signal crosses the zero axis.

$$ZCR(i) = \frac{1}{2N}\sum_{n=1}^{N} \operatorname{sgn}[x_n(i)] - \operatorname{sgn}[x_{n-1}(i)] \tag{4.2}$$

$$\text{Avec } \operatorname{sgn}[x(i)] = \begin{cases} 1, si\ x(n) \geq 0 \\ -1, si\ x(n) < 0 \end{cases} \tag{4.3}$$

4.5.2.2 Frequency-domain features

Fundamental frequency or pitch

The time between successive vocal fold openings is called the fundamental period, T_0, while the rate of vibration is called the fundamental frequency of the phonation, F_0= l/T_0. The term pitch is often used interchangeably with fundamental frequency. However, there is a subtle difference. Psycho-acousticians use the term pitch to refer to the perceived fundamental frequency of a sound, whether or not that sound is actually present in the waveform (John R. Deller, Proakis et Hansen, 1993).

It is one of the most used features in many applications, such as vowel classification, emotion classification, and music recognition. We have noticed that there are differences between expiratory and inspiratory phases regarding their frequency range (see Figure 4.12).

There are several ways to estimate the fundamental frequency, such as cross-correlation and auto-correlation. In this work, we employed the classic auto-correlation method by using PRAAT software. Traditional statistical features, such as maximum, minimum, and mean of the fundamental frequency, were calculated over the entire corresponding utterance.

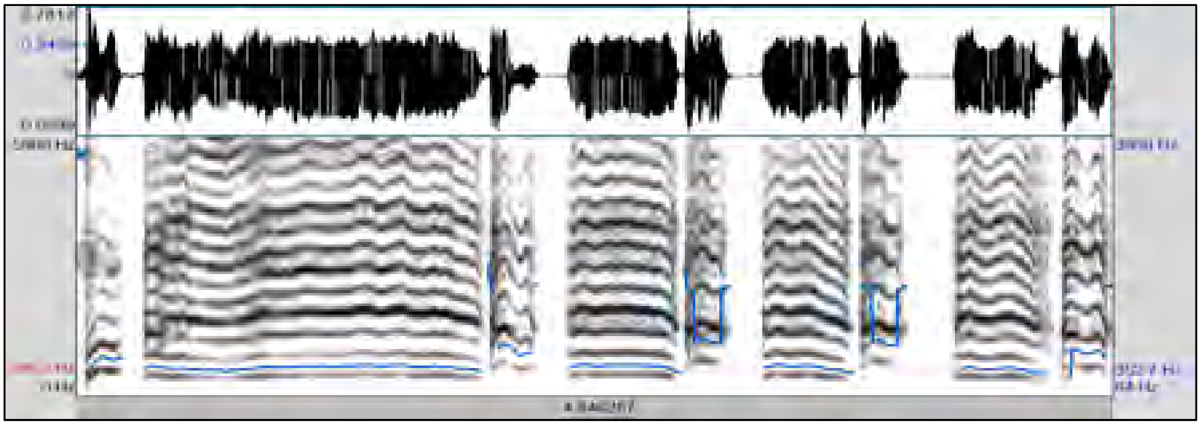


Figure 4.12 Example of the fundamental frequency contour in Praat software which is calculated on a consecutive expiration and inspiration with vocalization followed by a pause period

4.5.2.3 Time-frequency features

FFT-based MFCC

Mel frequency cepstrum coefficients (MFCC) are introduced in (Davis et Mermelstein, 1980) as the discrete cosine transform of the log-energy output of the triangular band pass filters. MFCCs are used to represent the human speech signal by calculating the short term power spectrum of the acoustic signal based on the linear cosine transform of the log power spectrum on a nonlinear Mel scale of frequency. Mel scale frequencies are distributed in a linear space in the low frequencies (below 1000 Hz) and in a logarithmic space in the high frequencies (above 1000 Hz) (Rabiner et Juang, 1993b).

The steps from the original input signal to MFCCs are as follows:

1. Split the signal into small overlapped frames of N samples;

2. Decrease discontinuity between consecutive frames by employing the Hamming window defined as follows;

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \tag{4.4}$$

3. Apply the fast Fourier transform (FFT);

4. Apply the log amplitude of the spectrum to the Mel scale filter banks:

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{4.5}$$

5. Apply the discrete cosine transform (DCT);

6. Perform the Inverse of FFT, and the resulting amplitudes of the spectrum are MFCCs according to the following equation;

$$c_n = \sum_{k=0}^{n-1} \log(S_k)\cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right], n = 1, 2, ..., K \tag{4.6}$$

where S_k is the output power spectrum of filters and K is chosen to be 12.

EMD-based MFCC

Empirical mode decomposition (EMD) is used for analyzing nonlinear and non-stationary signals and was proposed by Huang et al. in 1998. EMD breaks down a given signal into intrinsic mode functions (IMFs) and a residual function during a sifting process. The main advantage of EMD is that basic functions are obtained directly from the signal. An IMF represents a simple oscillatory mode of the signal and is defined to draw the position of the signal in time. Each IMF must fulfill the following two basic criteria (Huang et al., 1998):

1. The number of zero crossing rates and extremes in the entire sequence of data must be equal or differ by only one;

2. The mean value of the envelope defined by the local maxima and the envelope defined by local minima must be zero at any point.

To extract an IMF from a given signal, the following steps are necessary (Huang et al., 1998):

1. Identify the extrema of the signal x(t) separately;

2. Using the cubic splines interpolation method, interpolate the local maxima to form the upper envelope u(t);
3. Using cubic splines interpolation method, interpolate the local minima to form the lower envelope l(t);

4. Consider the local mean value of the upper and lower envelopes;

$$m(t) = \frac{u(t) + l(t)}{2} \tag{4.7}$$

5. Consider the local mean value from the original signal;

$$h(t) = x(t) - m(t) \tag{4.8}$$

6. Repeat the sifting approach until h (t) satisfies the basic criteria to be an IMF;

7. Finally, the residue $r(t) = x(t) - h(t)$ is regarded as the new signal to repeat the sifting process for the extraction of the second IMF and so on.

The EMD-based MFCC features are calculated by applying the MFCC extraction technique to the IMFs instead of the original signal. Based on the results obtained from our previous works (Abou-Abbas et al., 2015; Abou-Abbas et al., 2016) regarding finding the best IMF combination, we have found that the parameters extracted from the sum of IMF3, IMF4, and

IMF5 yielded the best cry segmentation results. Therefore, this combination was employed in this work: IMF345=IMF3+IMF4+IMF5.

4.5.3 Differential and acceleration coefficients

The features extracted from the decomposed frame describe only the static features of the corresponding segment. To add dynamic information and extract both linear and nonlinear properties, delta and delta-delta coefficients were used.

Delta coefficients or first-order regression coefficients are calculated according to the following equation:

$$d_i = \frac{\sum_{n=1}^{N} n(c_{n+i} - c_{n-i})}{2\sum_{n=1}^{N} n^2} \qquad (4.9)$$

where "d_i" is the delta or differential coefficient for ith frame and c_{n-i} and c_{n+i} represent the MFCC feature vectors ranging from n-i to n+i.

To compute acceleration or delta-delta coefficients, the same equation could be used but the static coefficients "c" should be replaced by the delta coefficients obtained in the previous step.

The dynamic information obtained by delta and delta-delta can be merged with static information to form one feature vector.

4.5.4 Classification

Each frame should be classified as either EXP, INS, BKG, or Other. Extracted features were used to classify each frame by comparison to the results obtained from the training data or predefined threshold. In this work, the classification part was performed three times. The first classification was based on a supervised classifier (a mixture of several weighted normal distributions, GMM, and a hidden Markov model, HMM), and the other two used dynamic thresholds, as will be discussed in section 4.6.

4.5.4.1 GMM

Gaussian Mixture Models (GMMs) are commonly used for different applications, mostly speech recognition, speaker verification and emotion recognition systems (Reynolds et Rose, 1995). It can be represented as a weighted sum of Gaussian distributions as follows:

$$p(o \mid \lambda) = \sum_{j=1}^{J} w_j G(o : \mu_j, \Sigma_j) \tag{4.10}$$

$$G(o; \mu_j, \Sigma_j) = (2\pi)^{-\frac{D}{2}} \left|\Sigma_i\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(o - M_i)^T \Sigma_i^{-1} (o - M_i)\right\} \tag{4.11}$$

where:

$p(o \mid \lambda)$ is the likelihood of the input observation *o* with dimensionality of *D*.

J is the number of mixtures.

w $_j$ are the weighting coefficients satisfying the constraint. $\sum_{j=1}^{J} w_j = 1$

$G(o_j, \mu_j, \Sigma_j)$

denotes the jth Gaussian with mean vector and covariance matrix. $\mu_j \; and \; \Sigma_j$

4.5.5 HMM

HMM-based methods have many potential applications in ASR systems, statistical signal processing and acoustic modeling, including the segmentation of recorded signals (Young et Evermann, 1996). The basic principles of any ASR system involve constructing and manipulating a series of statistical models that represent the various acoustic activities of the sounds to be recognized (Young et Evermann, 1996). Many studies have shown that speech, music, newborn cries, and other sounds can be represented as a sequence of feature vectors (temporal, spectral, or both), and HMMs could provide a very important and effective framework for building and implementing time-varying spectral vector sequences (Gales et Young, 2008b).

An HMM generates a sequence of observations $O=O_1, O_2, \ldots, O_T$ and is defined by the following parameters: the number of hidden states, state transition probability distribution A, observation probability distribution B, and initial state distribution π. We denote the model parameters of the HMM as $\lambda= \{A, B, \pi\}$ (Kuo, 2010; Várallyay, 2005).

To build and manipulate an HMM, three problems must be solved: the evaluation problem, the decoding problem, and the training problem. HMM theory, the aforementioned problems, and proposed solutions are widely explained in the literature, especially in the well-known Rabiner tutorial.(Rabiner, 1989) . Moreover, the Viterbi algorithm was proposed as a decoding solution to find the most probable future state of the system based on its current state (Rabiner, 1989). The Baum Welch algorithm is an iterative procedure used to estimate the HMM parameters.

4.5.6 Supervised cry segmentation system-initial classification

A supervised cry segmentation system is a chain of complex processing units that aims to locate cry sounds in audio signals collected in noisy clinical environments and identify its type as expiration, inspiration, background or other classes based on trained models for all available classes in the training phase. A cry segmentation system consists of the extraction of appropriate features from the signal in both training and testing phases, followed by a classifier. In the testing phase, the probability that a segment belongs to each one of the classes is calculated using the Viterbi algorithm, and a decision is taken to produce a label for the segment.

In the works conducted previously (Abou-Abbas et al., 2015; Abou-Abbas et al., 2016; Abou-Abbas, Fersaei Alaei et Tadj, 2015; Abou-Abbas, Fersaie Alaie et Tadj, 2015), studies were carried out to compare the accuracy rate of different feature vectors and classifiers. It has been shown that all of the proposed approaches mentioned in Table 4.5 allow for reliable discrimination of clean expiration from any other sounds.

Table 4.5 Set of experiments yielded using different decomposition techniques, different features extraction, and different classification methods.

Decomposition technique	**Features Extraction**	**Classification method**
FFT	FFT-MFCC	GMM
FFT	FFT-MFCC	4-states HMM
FFT	FFT-MFCC	5-states HMM
EMD	EMD-MFCC	GMM
EMD	EMD-MFCC	4-states HMM
EMD	EMD-MFCC	5-states HMM

The first step of the new proposed cry segmentation scheme was to make the first classification decision based on a supervised cry segmentation system to discriminate between four different classes: EXP, INS, BKG, and other. We choose two previously designed systems to test alongside the new approach due to their robust performance.

1. Supervised cry segmentation system using a GMM classifier based on FFT-MFCC features;

2. Supervised cry segmentation system designed using a 4-state HMM classifier based on MFCC extracted from a combination of three intrinsic mode functions: IMF3, IMF4, and IMF5 resulted from EMD.

We also added delta and acceleration features to have a feature vector of 39 parameters per frame. We trained four different classes based on their available training data**.**

4.5.7 Post-Processing stage

To further improve the obtained classification results from the previous step, a post-processing stage depicted in Figure 4.13 is proposed.

The main idea behind this stage was to refine the initially obtained results to reduce switching errors and adjust the boundaries of the expiratory and inspiratory phases. Then, to reline the classification results and make the final decision, we propose the second step based on temporal features and frequency features separately.

In particular, the following steps were taken in this stage:

1- First, the results obtained from the initial classification step (in label files) were further taken as input to the post-processing phase;

2- Different features were extracted for each frame from the corresponding input wave file, such as zero crossing rate, intensity, and fundamental frequency;

3- Two thresholds were calculated: ZC_BKG and INT_BKG. It was based on the Rabiner and Sambur rules in the endpoint detection algorithm (Rabiner et Sambur, 1975). The threshold computation module explained in Figure 4.13 is an essential step in estimating measures of background silence;

4- Each frame was labeled as “0” and “1” based on the F0 results. If F0 exists, the frame was indexed by “1”, and if F0 does not exist for the frame, it was indexed by “0”. Table 4.6 shows an example of indexing frames;

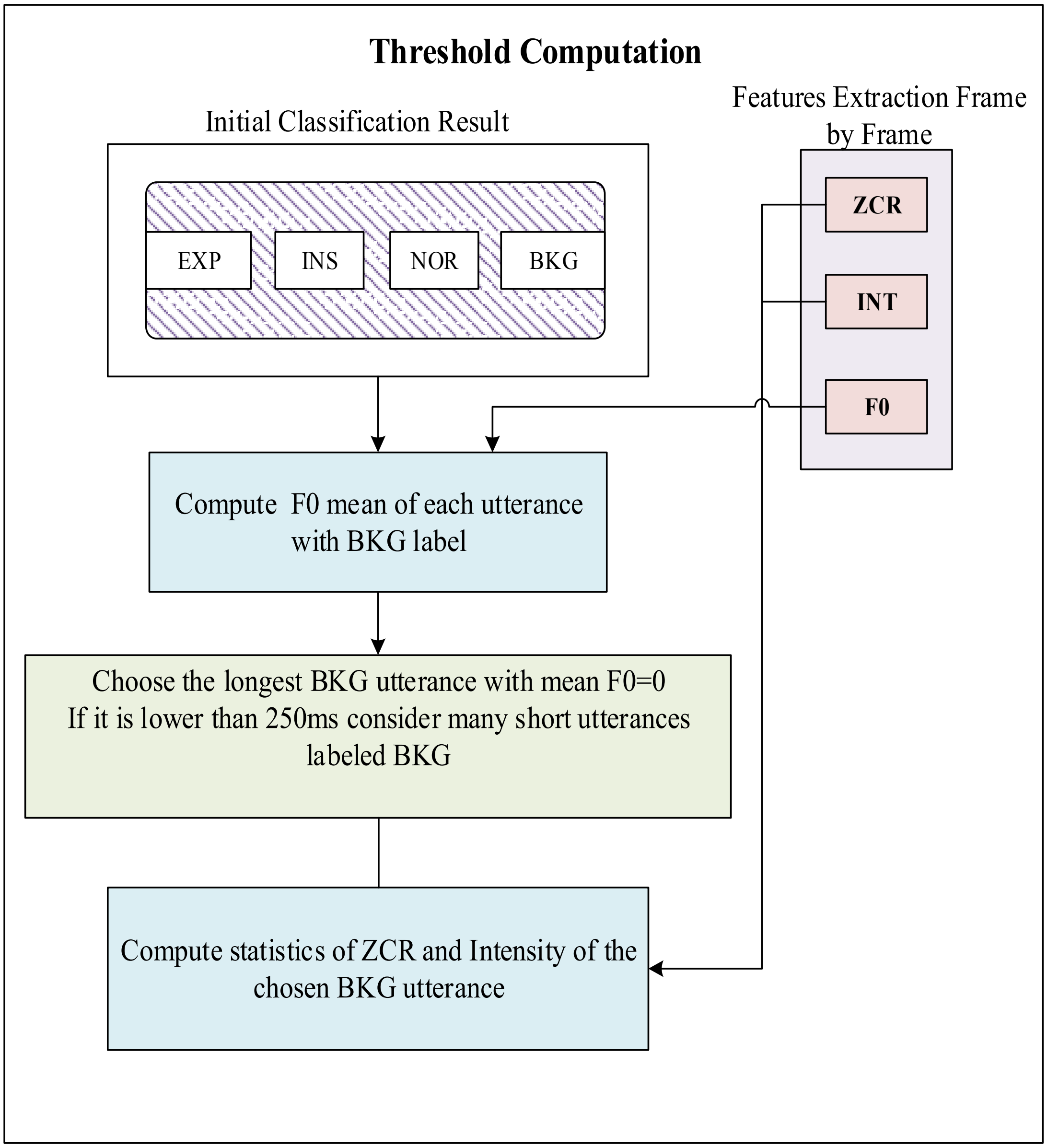

Threshold Computation
Initial Classification Result
Features Extraction Frame by Frame
EXP
INS
NOR
BKG
ZCR
INT
F0
Compute F0 mean of each utterance with BKG label
Choose the longest BKG utterance with mean F0=0
If it is lower than 250ms consider many short utterances labeled BKG
Compute statistics of ZCR and Intensity of the chosen BKG utterance

Figure 4.13 Dynamic Threshold Computation Module

5- The localized points were predicted from the transition between 0 and 1. These frames were indexed by “2” to determine the transitions within the localized points. See columns 2 and 3 in Table 4.6;

6- Then, all local Minima positions of intensity were indexed separately from the localized points. See column 4 in Table 4.6;

7- Three different indexes were gathered to create new intervals. The start and end of each interval were marked by the index “2.” See column 5 in Table 4.6;

Table 4.6- Process of threshold computation –step 3 to 7

Frame number	**First labeling**	**Marking Transition**	**Marking Local Minima**	**Combination of three columns indexes**
1	1	x	2	2
2	1	x	x	1
3	1	x	x	1
4	1	x	x	1
5	1	2	x	2
6	0	x	x	0
7	0	x	2	2
8	0	x	x	0
9	0	2	x	2
10	1	x	x	1
11	1	x	2	2
12	1	x	x	1
13	1	2	2	2
14	0	x	x	0
15	0	x	x	0
16	0	x	x	2

8- Calculate the mean values of the intensity and zero crossing rate;

9- Compare the intensity values and ZCR values with the thresholds (INT_BKG and ZC_BKG) and give new index labels of 0 and 1 to the intervals. See Table 4.7;

Table 4.7- Process of Threshold computation – step 8 and 9

Intervals from frame i to frame j	**Intensity**	**ZCR**	**New Index**
1-5	x	x	0
6-7	x	x	0
8-9	x	x	1
10-11	x	x	1
12-13	x	x	0
14-16	x	x	1

10- Merge the decisions of the initial classification with the new indexes to create new improved intervals. See Table 4.8;

11- Calculate the mean of the intensity to create new intervals, and add F0 (mean, max, min) values for each new interval;

12- Re-compare the mean values of the intensity and fundamental frequency statistics to make the final decision on utterances.

Table 4.8- Process of Threshold Computation – step 10

Frame number	**Initial Classification results**	**New Index**	**Combination**
1	BKG	0	BKG0
2	BKG	0	BKG0
3	BKG	0	BKG0
4	BKG	0	BKG0
5	BKG	0	BKG0
6	EXP	0	EXP0
7	EXP	0	EXP0
8	EXP	1	EXP1
9	EXP	1	EXP1
10	INS	1	INS1
11	INS	1	INS1
12	INS	0	INS0
13	INS	0	INS0
14	NOR	1	NOR1
15	NOR	1	NOR1
16	NOR	1	NOR1

4.6 Results and Discussion

The proposed method in this work was tested on a set of 507 cry signals collected from 203 babies. The performance was estimated at the frame level by comparing each classified frame to the reference frame (the results of manual segmentation performed by experts). Because the main advantage of this work is robustness, the systems were trained and tested using recorded cry signals under varying conditions, as mentioned in section 4.3.

Four different statistical measures were derived for each class: True positive (TP), False positive (FP), True Negative (TN), and False Negative (FN).

1. TP is the number of correctly detected sounds;
2. FP is all sounds detected erroneously;
3. TN represents the number of correctly rejected sounds;
4. FN is all missed sounds.

The performance of the proposed approach was evaluated using metrics such as:

$$True\ Positive\ Rate : TPR = \frac{TP}{TP + FN} \tag{4.12}$$

$$False\ Positive\ Rate : FPR = \frac{FP}{FP + TN} \tag{4.13}$$

$$False\ Negative\ Rate : FNR = \frac{FN}{FN + TP} \tag{4.14}$$

$$Accuracy : ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.15}$$

The first set of experiments was conducted using two supervised segmentation systems using GMM and HMM classifiers, respectively. Ten-fold cross-validation was used to evaluate the performance of the system in a manner such that one fold was reserved for validation while the remaining nine folds constitute the training set. This procedure was performed five times, and the average accuracies depicted in Table 4.9 were obtained.

In our experiments, the GMM-based and HMM-based classifiers were trained, respectively, by FFT-MFCC features and MFCC features extracted from the combination of three intrinsic Mode Functions, IMF3, IMF4, and IMF5 resulted from EMD.

The second set of experiments, which was designed to evaluate the post-processing stage, was further conducted using results obtained from the first set of experiments. The results obtained before and after the post-processing stage are shown in Table 4.9.

Table 4.9 shows the average TPR, FPR, FNR values of the proposed approach and the previous designed systems. Based on the results, the overall accuracies of FFT-GMM- and EMD-HMM-based systems improved by approximately 3.18% and 3.22 % by applying the post-processing stage (Figure 4.14).

Table 4.9 Results of different experiments

	EXP			**INS**			**BKG**			
%	**TPR**	**FPR**	**FNR**	**TPR**	**FPR**	**FNR**	**TPR**	**FPR**	**FNR**	**Overall Accuracy**
FFT-GMM	93.03	18.52	6.97	84.56	29.65	15.44	95.63	10.36	3.37	91.01
EMD-HMM	88.32	23.6	11.68	90.35	27.74	9.65	92.85	19.6	7.15	88.94
FFT-GMM+post processing	**95.6**	**10.32**	4.4	89.52	20.46	10.48	97.26	3.23	2.74	**94.29**
EMD-HMM+post processing	93.7	15.08	6.3	**92.47**	**18.21**	7.53	95.07	7.84	4.93	92.16

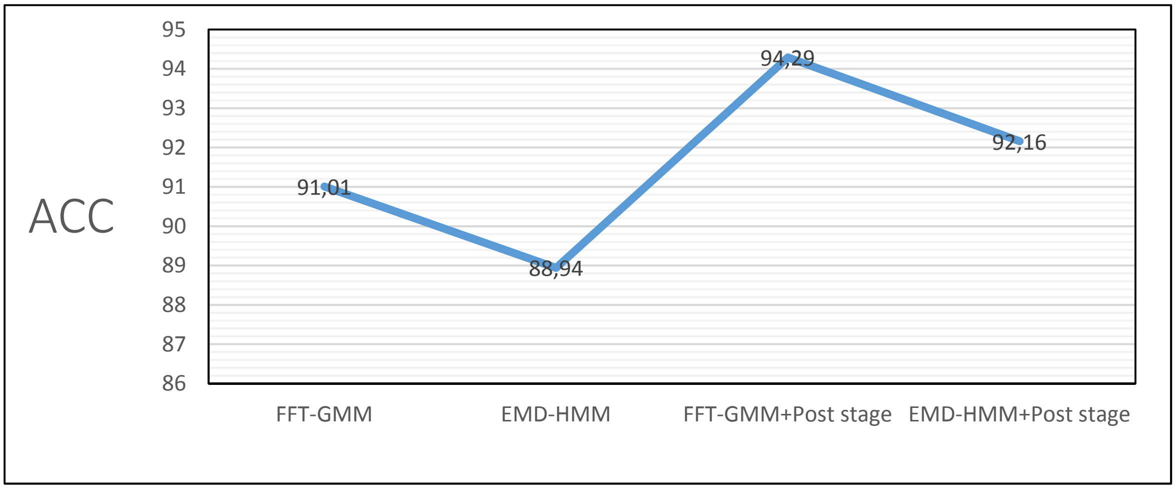


Figure 4.14 Comparison of the overall accuracy rate obtained

The proposed method achieved a high average TPR of approximately 95.6 % for the EXP class and 92.47% for the INS class by using GMM and HMM classifiers, respectively. Figure 4.15 shows a comparison of the TPR results obtained for the 4 experiments mentioned in Table 4.9. The false-positive rate was considered the most important error rate in our case because it is very important not to add misleading segments as input to a cry classification system. As we can see from Table 4.9 and Figure 4.16, the post-processing stage reduced the false-positive rate for the EXP and INS classes by 8.2 and 11.44%, respectively.

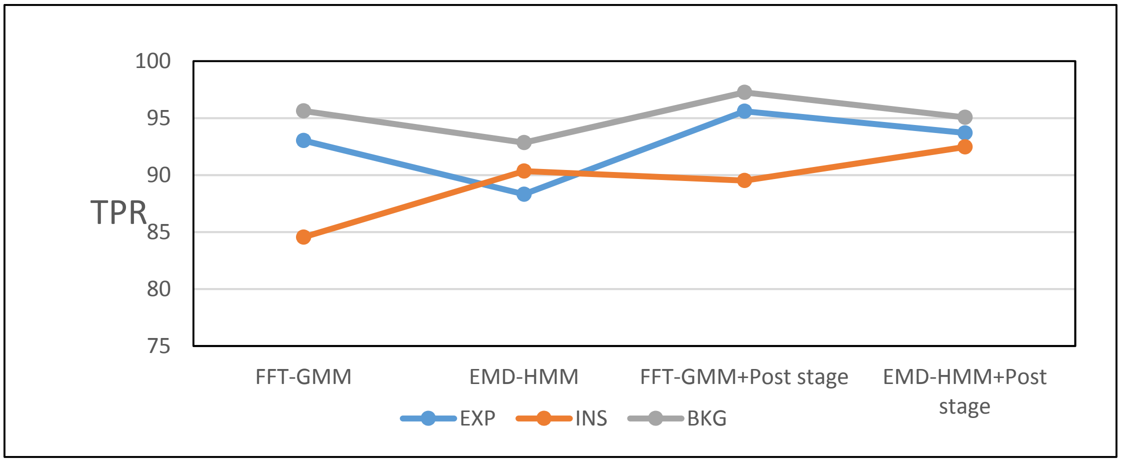


Figure 4.15 Comparison of TPR between different experiments for each class: EXP, INS, and BKG

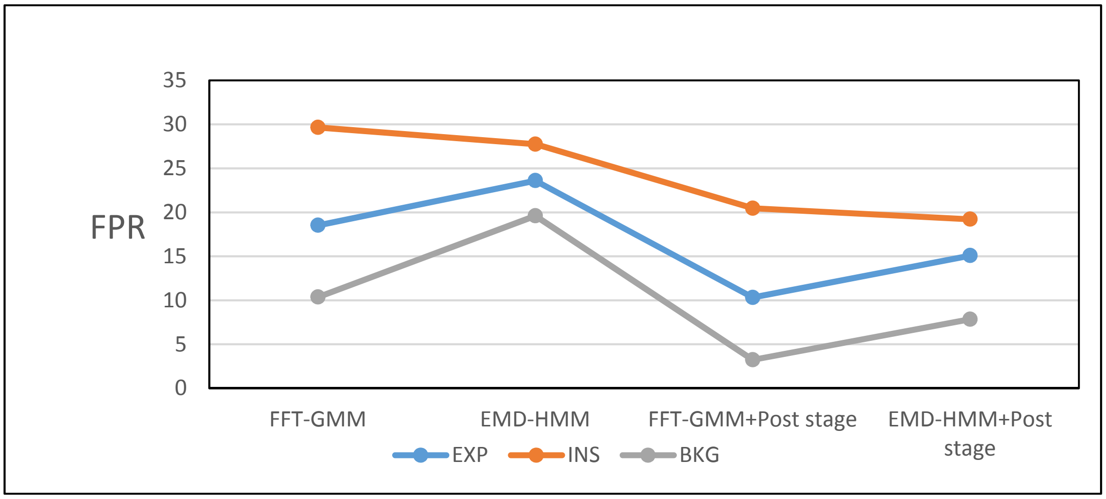


Figure 4.16 Comparison of FPR between different systems

Moreover, an average FPR rate of approximately 10.32% was obtained for the EXP class using the GMM-based classifier and 18.21 % for the INS class with the HMM-based classifier.

From the experiments, it can be seen that the proposed method achieved an average performance of 94.29%.

Based on the results, we can see that an improvement of 3.18% in the average classification rates was achieved by adding the post-processing stage, which was quite significant.

4.7 Conclusion

Cry segmentation is the process of segmenting a cry signal recorded in a real clinical environment into homogenous regions. When used in conjunction with an NCDS, an automatic cry segmentation system is able to detect useful cry units with vocalizations from non-cry audio regions (such as speech, machine sounds, speech, and other types of noises) and significantly improve the intelligibility of the NCDS.

Our main goal was to precisely locate important audible cry component boundaries of continuous recordings collected in a very noisy environment as a step towards building a cry database that contributes to developing different cry-based applications, such as the following:

- distinction between different types of cries: birth cry, pain cry, normal cry, pleasure cry, etc;
- classification between healthy and deaf infants;
- distinction between healthy infants and infants with asphyxia;
- discrimination of babies with cleft palates with palate plates from babies with cleft palates but without palate plates;

- classification of pathology from cries: asphyxia, meningitis;
- recognition of newborns' native language from cry melodies.

None of the aforementioned applications would operate well if the input signals were other than cry segments. The main goal of the proposed work was to design an automatic cry segmentation system that can be incorporated in the early stage of our NCDS system or any other cry classifier system.

In this work, expiratory and inspiratory cries with vocalization were detected from audio signals containing different acoustic activities other than cries. A post-processing stage was combined with a supervised cry segmentation system that has already been designed in our previous work. The proposed approach achieved better performance in terms of accuracy rate, true-positive rate, and false-positive rate.

Future work will include a combination of GMM- and HMM-based classifiers because the GMM-based classifier gave fewer false-positive alarms for the EXP class, while the HMM-based classifier gave fewer false-positive alarms for the INS class. Moreover, the final boundaries of EXP and INS phases could be used to extract temporal features, such as the duration of the expiratory and inspiratory phases, duration of pauses between cries, and onset of crying, which may result in a more accurate pathology classification system.

Acknowledgments

The authors would like to convey acknowledgment towards the Bill and Melinda Gates Foundation for funding this project under Grand Challenges Explorations and also the medical staff in Sainte-Justine hospital and Sahel hospital for their collaboration and assistance in acquisition of cries signals.

CONCLUSION

Ce travail de thèse entre dans le cadre global du développement d'un système automatique de diagnostic précoce qui permet d'identifier certaines pathologies chez les nouveau-nés à partir de leurs cris. Ce système vise à améliorer le diagnostic en soutenant la prise de décision clinique afin d'améliorer l'efficacité des traitements et d'éviter l'élaboration de pathologies.

Le déploiement d'un tel système à grande échelle pourrait engendrer d'importants progrès en matière de santé dans le monde en l'intégrant dans les départements de pédiatrie ou de néonatologie dans les hôpitaux.

Cette thèse s'intéresse plus particulièrement à la segmentation des signaux de cris. Cette tâche est essentielle dès que l'on veut traiter des signaux de cris.

Notre premier objectif a été la mise en place d'une base de données de cris des nouveau-nés sains et pathologiques destinée à être publique, accessible en permanence aux fins de recherches multiples relatives spécifiquement à la santé des nourrissons. 769 bébés, dont 372 sont atteints de différentes maladies, ont participé à la collecte des données pour avoir un total de 2073 signaux de cris.

Cette étude a été principalement suscitée par la résolution de la problématique principale de segmentation automatique des signaux de cris des nouveau-nés ayant pour but principal de réduire le plus possible l'intervention humaine. A cette fin, les travaux réalisés ont porté sur la détection des segments de cris importants ainsi que sur la localisation précise des bornes de début et fin des segments détectés.

La plus grande partie du travail dans cette thèse a été consacrée au développement d'un outil de segmentation complètement automatique robuste vis-à-vis le bruit et capable de détecter et bien localiser les zones utiles de cris. Nous avons exploré différentes approches pour la segmentation des signaux de cris enregistrés dans des conditions cliniques réelles en des classes bien prédéfinies qui sont les composantes primaires d'un cri, expiration et inspiration audibles ou en d'autres termes avec vocalisation. Les résultats des tests réalisés ont montré le comportement efficace des algorithmes proposés.

Dans un premier temps, une partie de travail a été dévouée à exploiter les différents modes de décomposition des signaux non-stationnaires : 1) Transformée de Fourier rapide, 2) Paquets d'ondelettes, 3) Mode de décomposition empirique. Cette partie de la recherche a fait l'objet d'une publication dans une revue (Chapitre 2) ainsi que de deux articles de conférences (Annexes I et II).

La première étape (Chapitre 2) nous a permis de mettre en œuvre un premier système de segmentation des signaux de cris à base de modèles de Markov cachés HMM, qu'on a appelé système de référence. Celui-ci a été conçu à partir de la boîte des outils HTK (Hidden Markov Model Toolkit). Nous avons choisi de construire six modèles HMM formant les différentes activités acoustiques qui composent nos enregistrements : EXP, INS, BIP, Bruit, Parole et Silence. Nous avons utilisé, pour l'étape de paramétrisation, les caractéristiques les plus communes en reconnaissance de la parole qui sont les MFCC et leurs dérivées temporelles premières et secondes. Les performances de ce système ont été évaluées à l'aide du corpus annoté et segmenté manuellement. Nous avons constaté que la taille du corpus d'apprentissage ainsi que l'utilisation des coefficients MFCC avec ou sans leurs dérivées premières et secondes, la taille de la fenêtre d'analyse, le nombre d'états et le nombre de mixtures gaussiennes ont un impact sur les performances du système de segmentation proposé.

Dans la deuxième étape (Annexe I), nous avons présenté une autre technique de paramétrisation basée sur les paquets d'ondelettes en raison de sa richesse dans la résolution fréquentielle. Nous avons comparé les performances des deux paramètres : l'énergie traditionnelle calculée à partir des coefficients d'ondelettes et l'opérateur d'énergie Teager-Kaiser (TKEO) qui est calculé en prenant en compte des échantillons voisins. Les résultats obtenus ont indiqué une amélioration de la segmentation en utilisant le paramètre TKEO.

Dans la troisième étape (Annexe II), une technique plus récente de décomposition des signaux non-stationnaires a été exploitée. Le signal d'entrée a été décomposé en cinq IMFs. Par la suite, différentes fusions des IMFs ont été testées. Les meilleurs résultats ont été obtenus en utilisant la somme des IMF3, IMF4 et IMF5.

Les résultats acquis aux trois premières étapes de travail nous ont encouragés à nous concentrer par la suite sur ces trois modes de décomposition en variant les paramètres extraits d'une part et d'autre part l'approche de modélisation. Nous avons constaté que les performances des systèmes s'améliorent au fur et à mesure qu'on augmentait la taille du corpus d'apprentissage. Nous avons choisi de déployer GMM et HMM avec chaque système développé et de comparer les performances en utilisant deux tailles de fenêtre d'analyse différentes. Neuf systèmes de segmentation des cris des nouveaux nés en parties EXP, INS ou autres, ont été manipulés dans ce travail. Les résultats ont montré que l'approche GMM a surpassé l'approche HMM en réduisant le taux d'erreur à 8.98%. Cette partie de recherche a fait l'objet d'une publication dans une revue (Chapitre 3).

La dernière partie du travail, présentée dans le Chapitre 4 de la thèse, se veut une nouvelle approche complète pour remédier aux limites des approches proposées, basée sur les approches purement markoviennes. L'objectif principal de cette partie était d'étendre les méthodes proposées dans les travaux précédents pour inclure une partie de post-traitement afin d'améliorer les performances de classification. Cette approche complète a permis de corriger les erreurs de la classification supervisée en ajoutant une étape qui permet de localiser précisément les segments expirations et inspirations audibles. Différentes techniques de calcul d'erreurs ont été ajouté afin d'estimer plus efficacement les performances de la nouvelle approche. Des taux de détection des segments EXP et INSV ont atteints environ 94,29 et 92,16%, respectivement, en appliquant une technique de validation croisée dix fois. Cette partie a fait l'objet d'une soumission de notre troisième article dans une revue.

Avec ces méthodes, nous avons réussi à atteindre nos objectifs principaux. Nous considérons que les résultats acquis sont satisfaisants.

Notre outil réalisé pourra faire partie de la phase du prétraitement des signaux audio des cris des nouveau-nés et ce en amont du système automatique de classification des pathologies chez les nouveau-nés.

RECOMMANDATIONS

Compte tenu de l'état actuel du processus de segmentation des signaux de cris des nouveau-nés proposé dans cette thèse, nous pouvons formuler des recommandations au sujet de ce qui peut être fait pour améliorer les performances.

Les approches de classification n'ont pas été suffisamment investiguées dans cette thèse. Il serait intéressant d'exploiter d'autres classifieurs existants tels que les réseaux de neurones probabilistes, la machine à vecteurs de support et la régression linéaire. D'une part, le choix du classifieur le plus adapté au problème de la segmentation des cris sera indispensable. D'autre part, les résultats obtenus au terme de ce travail nous ont montrés que les erreurs de détection des segments EXP ont été minimes avec un classifieur GMM alors que celles des segments INS ont été minimes avec un classifieur HMM. Ainsi, la prise en compte d'une classification hybride ou d'un classifieur par classe semble être intéressante.

Au niveau du modèle de langage, différents modèles spécifiques à d'autres activités acoustiques (parole, claquement de la porte, sons des machines médicaux) pourraient être considérés pour améliorer la classification. Nous n'avons néanmoins pas réussi à considérer ces modèles en raison des données d'apprentissage limitées.

Les segments correspondants aux cris sur fond de la parole et encore aux cris de plusieurs nouveau-nés en même temps, restent parfois difficiles de discriminer du cri. Il semble intéressant d'utiliser des méthodes pour séparer les différentes activités acoustiques telles que les méthodes de séparation aveugle des sources.

Un des objectifs de notre projet NCDS global est de développer un outil d'aide au diagnostic visé à être utilisé en temps réel en l'intégrant dans les départements de pédiatrie ou de néonatologie dans les hôpitaux. Ainsi, une optimisation du temps du calcul durant les étapes d'extraction des caractéristiques et de classification s'avère nécessaire.

Dans la partie d'extraction des caractéristiques, qui sert à former le vecteur d'entrée des classfieurs, nous n'avons néanmoins pas pu exploiter différentes caractéristiques. De plus, la fusion des différentes techniques de décomposition utilisées n'a pas été prise en compte. La recherche dans cette direction mérite d'être poursuivie, puisque chacune de ces techniques présente des aspects intéressants surtout par rapport à un signal non-stationnaire.

La distinction entre les expirations et les inspirations parmi les autres activités acoustiques était l'issue principale de cette thèse. Ainsi l'interprétation physique des paramètres dépendant des durées des segments d'expiration et d'inspiration détectés pourraient être une nouvelle information pertinente en permettant d'établir des corrélations entre ces paramètres et celles des différentes pathologies.

Et enfin, une autre piste de recherche, concernant la phase de prétraitement des signaux de cris, qui consiste à déployer des approches de débruitage du signal, n'a pas toujours été investiguée. L'intégration d'un module de débruitage va peut-être permettre d'augmenter les performances de la segmentation et de la classification des pathologies.

ANNEXE I

SEGMENTATION OF VOICED NEWBORNS CRY SOUNDS USING WAVELET PACKET BASED FEATURES

L. Abou-Abbas [a], H. Fersaie Alaie [b] and C Tadj [c]

[a,b,c] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Notre-Dame West, Montréal, QC, Canada H3C1K3

This paper is published
In *The 28th Canadian Conference on electrical and computer engineering* (CCECE)

Abstract

This paper proposes a method for the segmentation of newborn's cry signals recorded in real conditions using the Teager-Kaiser energy operator (TKEO). Based on the wavelet packet analysis, the audio signals are divided into different frequency channels, and then the TKEO and the energy are estimated within each band. The Hidden Markov Models have been used as a classification tool to distinguish the voiced cry parts from the irrelevant acoustic activities that compose the audio signals. The proposed method divided the audio signal containing newborns' cry sounds into different periods showing the audible Expiration and Inspiration of the cry. Different levels of wavelet packet transform have been used to verify the performance of the proposed method on crying signals segmentation and have shown that based on wavelet packet decomposition, the TKEO measure is more effective than the traditional energy measure in detecting important parts of cry signal in a very noisy environment. The proposed features have shown to achieve an accuracy rate of 84.08 %.

A I-1 Introduction

Plusieurs études ont établi l'existence d'une information importante dans le signal du cri d'un nouveau-né (Golub, 1979; Orlikoff, Baken et Kraus, 1997; Proctor, 1984; Várallyay, 2006). En se basant sur cette hypothèse, de nombreuses recherches se sont consacrées à l'analyse de ce signal dans le but de classifier d'une part, le type du cri (cri de naissance, douleur, faim, inconfort, etc.) et d'autre part, l'état pathologique du nouveau-né (sain, malade) (Farsaie Alaie et Tadj, 2012; Hariharan et al., 2012; Rui et al., 2010).

En 1985, Corwin et Golub, ont classé les parties audibles du cri en quatre catégories importantes : a) Phonation expiratoire (avec F0 entre 250-750 Hz) b) Hyperphonation expiratoire (avec F0 entre 1000-2000 Hz) c) Dysphonation expiratoire (segment expiratoire apériodique) d) Phonation inspiratoire (qui est associée à une partie du cri audible générée durant une phase d'inspiration) (Golub, 1979).

Des études récentes réalisées sur l'analyse du cri ont révélé qu'il existe une différence entre les caractéristiques des vocalisations expiratoire et inspiratoire et il a été prouvé que la partie inspiratoire serait utile pour la classification des bébés malades (Orlikoff, Baken et Kraus, 1997). De ce fait, il est important de distinguer dans les signaux audio de cris, les parties expiratoires et inspiratoires des autres activités acoustiques inutiles.

Dans les différentes approches de traitement du signal utilisées dans l'analyse du cri du bébé, la segmentation des signaux enregistrés s'avère un des problèmes les plus complexes. Elle constitue une étape essentielle pour la classification du signal.

Des signaux de cris des nouveau-nés enregistrés peuvent comporter des cris qui sont des suites d'expirations et d'inspirations, des paroles de qualités variables (infirmières, parents, etc.), des périodes de silence et de bruits, etc. Ainsi, les activités acoustiques inutiles nuisent aux processus de l'analyse et du traitement. Jusqu'à présent, et dans la plupart des travaux de recherche, la segmentation a été réalisée manuellement (Michelsson et Michelsson, 1999; Proctor, 1984; Wermke et al., 2002). Ceci représente une tâche fastidieuse, pouvant requérir des heures de travail par signal.

De rares études ont été menées spécifiquement sur la segmentation automatique des signaux de cris. Dans des travaux récents (Bajaj et Pachori, 2012; Várallyay, Illényi et Benyó, 2009; Várallyay Jr, Illényi et Benyó, 2008), des approches basées sur une méthode de détection de la fréquence fondamentale « Harmonic Product Spectrum » ont été introduites. Les auteurs ont montré qu'en utilisant ses méthodes, il est en fait possible de classifier la structure spectrale d'un signal donné en détectant ainsi les parties de cri voisées parmi autres activités acoustiques.

Une autre recherche en 2012 portante spécifiquement sur la segmentation des signaux de cris a été menée en 2012. Le but était de marquer chaque segment qu'il soit un cri/non cri/non-activité. Cette recherche était basée sur l'étude du contenu spectral ainsi que l'harmonicité du signal. À la différence de notre base de données variée, la base de données utilisée dans ces travaux n'était formée que des suites d'expirations et d'inspirations qui peuvent être alternées par des périodes de silence ou de faible bruit de l'environnement.

Dans ce travail, nous nous intéressons à une méthode de segmentation automatique des signaux audio des cris de nouveau-nés. Cette méthode est basée sur une étape de paramétrisation sur l'opérateur d'énergie TKEO introduit par Teager-Kaiser (Kaiser, 1990; Maragos, Kaiser et Quatieri, 1993a; Teager et Teager, 1989) calculé à partir de l'analyse en paquets d'ondelettes ainsi qu'une étape de classification automatique par l'approche HMM.

L'article est structuré comme suit : dans la deuxième section, la phase d'extraction des caractéristiques fondée sur les paquets d'ondelettes est présentée. La troisième section est réservée à la description du corpus utilisé. La quatrième section décrit le système de segmentation automatique proposé. Les résultats obtenus ainsi qu'une discussion sont exposés dans la cinquième section. Enfin nous terminons par une conclusion.

A I-2 Extraction des caractéristiques basées sur les paquets d'ondelettes

L'analyse par paquets d'ondelettes a connu beaucoup de succès dans le domaine de l'analyse du signal audio et en particulier dans l'analyse de la parole. Une caractéristique essentielle des ondelettes réside dans leur capacité à contrôler à la fois les variables temps et fréquence d'un signal. Dans le cas d'une analyse par paquets d'ondelettes, le signal original est décomposé en deux vecteurs : *Approximation* et *Détails* dans un premier niveau, puis ces vecteurs à leur tour sont de nouveau décomposés en deux sous vecteurs détails et approximations et ainsi de suite. La Figure-A I-1 montre une architecture d'une décomposition en paquets d'ondelettes :

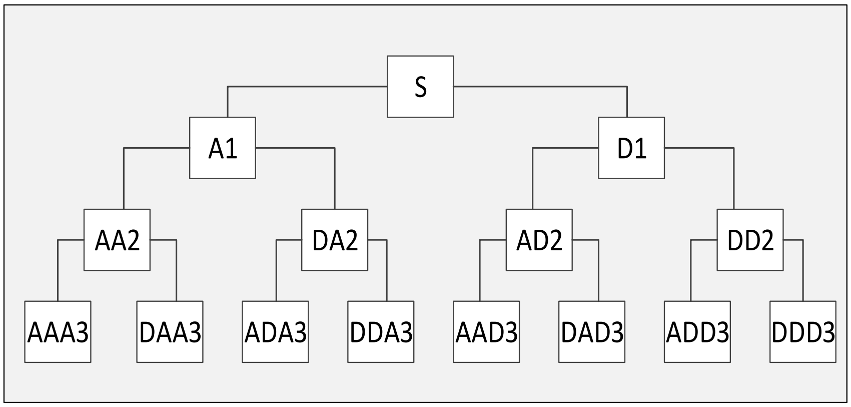


Figure-A I-1 Arbre de niveau 3 obtenu par décomposition en paquets d'ondelettes

Les paquets d'ondelettes sont représentés par les équations suivantes :

$$W_{2n}(x) = \sqrt{2}\sum_{k=0}^{2N-1} h(k)W_n(2x-k)$$
$$W_{2n+1}(x) = \sqrt{2}\sum_{k=0}^{2N-1} g(k)W_n(2x-k)$$

Avec h et g sont de filtres miroirs en quadrature respectivement passe-bas et passe-haut. Le détail du calcul des coefficients des ondelettes sont disponibles dans (Mallat, 2000). Ainsi, en se basant sur les coefficients d'ondelettes, nous calculons deux paramètres d'énergie :

Energie du paquet i

C'est l'énergie instantanée qui est calculée à partir de la formule :

$$E_i = \sum_k W_i(k)^2$$

Avec W $_i$(k) les coefficients d'ondelettes.

Operateur d'énergie Teager-Kaiser (TKEO)

C'est un opérateur différentiel non linéaire qui permet d'estimer l'énergie d'un signal basée sur ses caractéristiques physiques réelles (Maragos, Kaiser et Quatieri, 1993b). Il a été utilisé avec succès dans plusieurs domaines de traitement du signal et plus spécialement dans le domaine de l'analyse de la parole (Kaiser, 1990; Maragos, Kaiser et Quatieri, 1993a; Teager et Teager, 1989). Cet opérateur calcule une énergie instantanée en prenant compte des échantillons voisins. Un aspect important qui montre la simplicité du calcul de l'opérateur TKEO est que trois échantillons à chaque instant du temps sont nécessaires.

L'opérateur TKEO appliqué à un signal x(t) est défini par :

$$TKEO[x(t)] = (\frac{dx(t)}{dt})^2 - x(t)\frac{d^2x(t)}{dt^2}$$

Dans le cas d'un signal x(n) discret, TKEO est défini comme suit :

$$TKEO[x(n)] = x^2(n) - x(n+1)x(n-1)$$

A I-3 Corpus utilisé

Notre base de données utilisée pour l'entrainement et pour le test de notre système automatique de segmentation comporte des enregistrements appartenant à 64 nouveau-nés de six semaines ou moins, qui peuvent être prématurés ou à terme, sains ou malades. Ces enregistrements ont été effectués dans divers hôpitaux. La base de données créée comporte des signaux audio formés de longues périodes de cri qui sont une suite d'expiration et d'inspiration ainsi que d'autres activités acoustiques comme la parole, le bruit, le bip d'une machine médicale et le silence. Pour évaluer le système proposé, nous avons utilisé 200 signaux : 100 pour le corpus d'entrainement et 100 pour le corpus de test. La durée d'un signal peut aller de deux à trois minutes. Nous avons ainsi obtenu un total de 450.5 minutes.

Il est important de noter que nous nous sommes basés par la suite sur les annotations « Expiration » et « Inspiration » afin de désigner les parties des signaux de cris audibles sans tenir compte des inspirations et des expirations non perceptibles ou plus précisément respiratoires. Les parties, qui sont les sujets de notre étude, sont celles qui sont générées par les nouveau-nés pendant les phases d'expiration et d'inspiration de l'air durant un cri et non pas durant une phase de respiration normale.

Au total, notre corpus, détaillé dans le tableau-A I-1, contient 44.6% d'expiration, 21.9% d'inspiration et 33.5% d'autres activités acoustiques.

Table-A I-1 Détails sur les données annotées du corpus utilisé

Symbole	**Activités acoustiques**	**Temps en min**
EXP	Expiration	200.78
INS	Inspiration	98.5
Autres	Parole, Bip, bruit, silence…	151.22

A I-4 Système de segmentation proposé

Nos signaux sont enregistrés avec un taux d'échantillonnage de 44.1 KHz. Les différentes étapes de notre système de segmentation proposé sont illustrées par la Figure-A I-2. Le prétraitement, incluant le filtre de préaccentuation de coefficient 0.97 ainsi que le fenêtrage constituent la première étape de la procédure de segmentation. Nous avons analysé le signal en employant la fenêtre de Hamming de durée 30 ms et d'un recouvrement de 21ms. La décomposition du signal en paquets d'ondelettes est ensuite effectuée sur le signal. Nous avons choisi d'utiliser la famille d'ondelettes la plus connue et la plus utilisée dans le domaine de l'analyse de parole qui est celle de Daubechies (db-4) ainsi que de tester différents niveaux de décomposition de paquets d'ondelettes n (de 5 à 7).

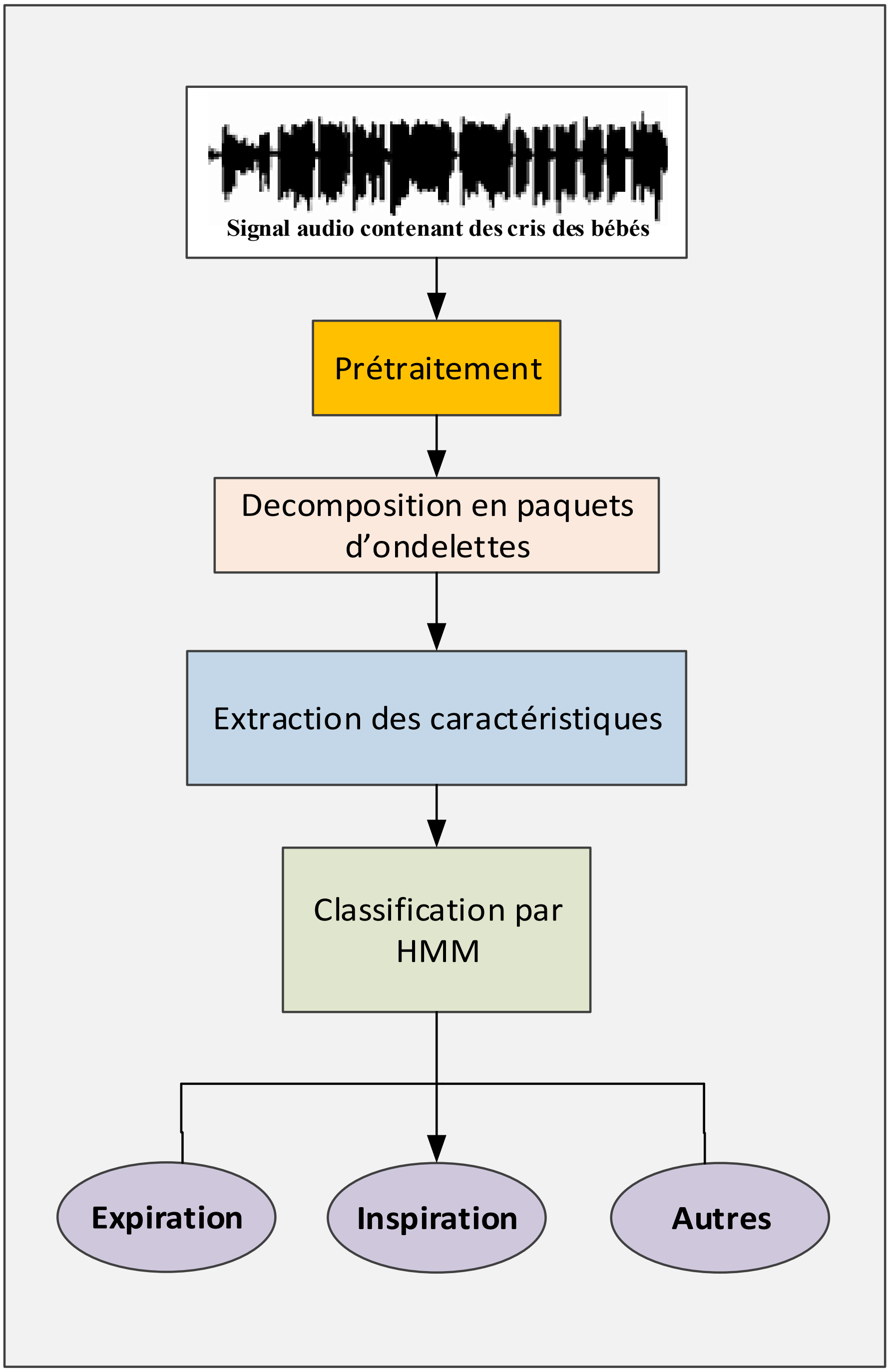

Signal audio contenant des cris des bébés
Prétraitement
Decomposition en paquets d'ondelettes
Extraction des caractéristiques
Classification par HMM
Expiration
Inspiration
Autres

Figure-A I-2 Architecture du système de segmentation proposé

Comme toute méthode de segmentation d'un signal audio, l'étape d'extraction des paramètres s'avère la plus importante. Après avoir appliqué la décomposition en paquets d'ondelettes, nous nous retrouvons avec 2n bandes de fréquences. Nous nous proposons de calculer dans chaque bande l'opérateur TKEO et l'énergie moyenne. Nous obtenons ainsi deux vecteurs caractéristiques de longueurs 2n pour chaque segment d'entrée X.

$$Y1 = (TKEO_1, TKEO_2, ..., TKEO_{2^n})$$
$$Y2 = (e_1, e_2, ..., e_{2^n})$$

Après l'étape d'extraction des paramètres, nous procédons à la classification des signaux audio. Nous avons choisi de déployer l'approche HMM qui nécessite deux étapes essentielles: entrainement et reconnaissance. Pour plus de détails sur les phases de l'approche HMM, nous referons le lecteur à une publication intéressante de Rabiner (Rabiner et Juang, 1993b).

Nous avons construit trois classes HMM formant les différentes activités acoustiques qui composent nos enregistrements :

1. Classe Expiration : Elle regroupe toute activité vocale générée par le bébé durant une phase d'expiration de l'air. Cette activité peut avoir lieu durant un épisode de cri ou autre comme le pseudo-cri;

2. Classe Inspiration : Il existe deux types d'inspiration : inspiration sonore et inspiration sourde. Dans notre cas, il s'agit de tout type de longue inspiration sonore que nous avons pu détecter fréquemment chez des bébés malades ayant des problèmes de respiration;

3. Classe « Autres» : C'est la classe qui regroupe les sons des machines médicales, la parole, le silence et le bruit.

A I-5 Résultats expérimentaux et discussions

L'objectif principal de notre étude est la détection des parties expiration et inspiration audibles dans des signaux audio composés de nombreuses activités acoustiques autre que les cris comme la parole, les « bip », les bruits divers, et les périodes de silence.

Afin d'optimiser la performance de notre système de segmentation, plusieurs variables sont évalués :

1. Niveaux de décomposition de paquets d'ondelettes : 5, 6 et 7 ;
2. Nombres d'états HMM : 6, 7 et 8 ;
3. Nombre de gaussiennes : 20, 30, 32, 35 et 40.

Les performances du système proposé sont évaluées à l'aide du corpus annoté et segmenté manuellement.

Les mesures de performance sont calculées en fonction du nombre d'erreurs de substitution S, de suppression D et d'insertion I. Nous avons ainsi calculé le taux de précision A:

$$A = \frac{N - D - I - S}{N} \times 100\%$$

Avec N : nombre total d'annotations dans un signal étiqueté manuellement.
D : nombre d'erreurs de suppression
S : nombre d'erreurs de substitution
I : nombre d'erreurs d'insertion
Les tableaux-A I-2, A I-3 et A I-4 exposent les résultats de la précision de la segmentation proposée en utilisant les niveaux de décompositions 5, 6 et 7 respectivement.

Table-A I-2 Résultats de la segmentation automatique pour le niveau de décomposition 5 d'ondelettes

Nombre d'états HMM	**Nb de Gaussiennes**	**Energy %**	**TKEO %**
6	20	63.27	65.58
6	30	65.39	68.45
6	32	68.25	69.24
6	35	70.35	71.26
6	40	73.52	75.33
7	20	70.17	72.34
7	30	73.28	73.85
7	32	73.62	74.69
7	35	74.01	74.87
7	40	74.35	76.09
8	20	72.21	74.53
8	30	72.48	76.26
8	32	74.23	77.08
8	35	74.65	77.43
8	40	75.3	78.56

Table-A I-3 Résultats de la segmentation automatique pour le niveau de décomposition 6 d'ondelettes

Nombre d'états HMM	**Nb de Gaussiennes**	**Energy %**	**TKEO %**
6	20	67.03	70.47
6	30	70.49	73.33
6	32	73.72	74.53
6	35	70.38	75.27
6	40	75.22	76.31
7	20	72.1	74.82
7	30	75.14	77.91
7	32	76.83	77.97
7	35	77.19	78.1
7	40	77.7	79.03
8	20	73.3	74.5
8	30	75.86	77.52
8	32	77.8	79.31
8	35	78.59	80.47
8	40	79.15	81.23

Table-A I-4 Résultats de la segmentation automatique pour le niveau de décomposition 7 d'ondelettes

Nombre d'états HMM	Nb de Gaussiennes	Energy %	TKEO %
6	20	70.81	72.95
6	30	74.12	75.41
6	32	74.87	75.83
6	35	75.45	76.05
6	40	76.22	77.76
7	20	75.84	76.12
7	30	76.13	78.26
7	32	76.52	79.05
7	35	76.98	79.88
7	40	77.29	80.37
8	20	76.51	78.7
8	30	80.04	81.95
8	32	80.56	82.62
8	35	80.85	83.21
8	40	81.41	84.08

Ces trois tableaux montrent l'influence du nombre d'états HMM et du nombre de gaussiennes sur le taux de précision.

En effet, il apparait clairement que le taux de performance augmente avec l'augmentation du nombre d'états et du nombre de gaussiennes.

Nous constatons également d'après toutes les expériences menées que l'opérateur TKEO a donné des résultats plus efficaces que de la mesure de l'énergie dans chaque bande de fréquence.

Le meilleur taux global est obtenu avec l'opérateur TKEO lorsque le niveau de décomposition est 7. Il a atteint un pourcentage de 84.08%.

A-I-6 Conclusion

Nous avons proposé une méthode de segmentation automatique des signaux de cris des nouveau-nés en se basant essentiellement sur l'analyse par paquets d'ondelettes en raison de sa richesse dans la résolution fréquentielle. Par rapport à l'énergie calculée dans chaque bande de fréquence, nous avons constaté que l'opérateur TKEO donne des résultats plus efficaces dans la segmentation. Nous avons testé des décompositions en paquets d'ondelette de niveaux 5, 6 et 7.

Nous avons montré qu'il est possible de détecter les phases d'expirations et d'inspirations audibles du cri avec un taux de précision allant jusqu'à 84.08%. Cette phase de segmentation est indispensable dans le cadre de l'analyse des cris des nouveau-nés sains et malades. Les résultats de l'approche utilisée sont satisfaisants. Nous avons constaté que le nombre d'états et de mixtures de gaussiennes améliorent les performances du système.

Cependant nous constatons l'existence des erreurs de segmentation qui sont dues à la présence de quelques signaux très bruités. Afin d'améliorer les résultats obtenus, nous envisageons de tester d'autres types d'ondelettes ainsi que d'ajouter une phase de post-traitement qui sert à bien localiser les segments étiquetés. Nous ajoutons aussi que la taille de fenêtre d'analyse doit être un sujet de test.

Remerciements

Nous remercions la fondation Bill et Melinda Gates qui a soutenu financièrement notre projet. Nous tenons à remercier également tous les membres du groupe de la néonatologie de l'hôpital Sainte-Justine à Montréal et de l'hôpital Sahel au Liban pour leur collaboration concernant la création de notre base de données de cris des nouveau-nés.

ANNEXE II

ON THE USE OF EMD FOR AUTOMATIC NEWBORN CRY SEGMENTATION

L.Abou-Abbas[a], L. Montazeri[b], C. Gargour[c] and C. Tadj[d]

[a,c,d] Department of Electrical Engineering, École de Technologie Supérieure,
1100 Rue Notre Dame-West, Montreal, Quebec, Canada H3C1K3
[b] Department of Electrical Engineering, Polytechnique Montreal, Quebec, Canada H3T 1J4

This article is published in
The International Conference on Advances in Biomedical Engineering (ICABME)
In 16-18 Sept. 2016, pp. 262-265

Abstract

Cry segmentation is an essential preprocessing step in any infant crying diagnosis system. Besides crying sounds consisting of expiration phases followed by short periods of inspiration episodes, each recording of newborn cries also includes silence sections as well as other sounds such as speech of caregivers, noise and sound of medical equipments. This paper is devoted to a newly developed Empirical Mode Decomposition (EMD) application to cry segmentation. The goal of the segmentation is to detect cry episodes automatically from unimportant acoustic activities existed inside the recorded signals. EMD decomposes a multicomponent non stationary signal into a set of monocomponent signals called Intrinsic Mode Functions (IMFs). The cry signals are segmented using Hidden Markov Models (HMMs) applied to the features extracted by employing EMD combined with Mel-Frequency Cepstral coefficients to the recorded cry signals. The performance of the proposed approach is evaluated on a database of 200 cry signals recorded in a real clinical environment. The experimental results demonstrate the effectiveness and suitability of the proposed method for the automatic cry segmentation.

Keywords : Automatic cry segmentation; Empirical mode decomposition; Features extraction; Mel-frequency Cepstral coefficients; Classification; Hidden Markov Models.

A II-1 Introduction

Although it might be hard sometimes to listen to infants crying, it is their only means of communication with their parents or caregivers, allowing them to express their needs for support. While cry might be considered as a simple behavior, it is actually, rather complicated. Studies show that newborns' cries take on different patterns depending on the reason of cries. Different features of newborn cries such as raising-falling pitch pattern, rapid pitch shifts, ascending-descending melody, high intensity, or intensity lower than normal, length of the cry and other time relations, can give some hints concerning different pathological conditions (Davis, 1983; Golub, 1979; Orozco-García et Reyes-García, 2003b). Therefore, it becomes essential to put effort into newborn cry signal analysis.

Automatic cry segmentation serves as a primary step towards infant cry signal analysis applications such healthy and pathological classification. A segmentation system should be able to identify cry segments and distinguish them from other audio types. In most studies, cry segmentation has been made manually by experts. This technique, while highly effective, is time-consuming and cannot serve the needs of a real-time newborn diagnostic tool. Reviewing the literature, it has been found some efforts on automatic segmentation of cry sounds using different voice activity detection methods like zero crossing rate and short term energy (Kuo, 2010; Rui et al., 2010; Zabidi et al., 2009a), Harmonic Product Spectrum (Várallyay, Illényi et Benyó, 2009) and methods based on HMM classification (Abou-Abbas, Fersaei Alaei et Tadj, 2015; Abou-Abbas, Fersaie Alaie et Tadj, 2015). Unlike most of these previous works on this topic which uses a database composed only of cry sounds, this paper is meant for segmenting a database recorded in a real clinical environment. In this work, a novel approach which relies on EMD-based features extraction technique has been proposed for cry segmentation. EMD has been widely used and applied by researchers to different areas and has demonstrated to be effective in the domain of biomedical signal and speech signal processing (Bajaj et Pachori, 2012; Martis et al., 2012; Wu et Huang, 2009).

In this framework, the input signal is hierarchically decomposed using empirical mode decomposition (EMD) and the given signal is divided into a set of Intrinsic Mode Functions (IMFs) and residual. Mel-frequency Cepstral coefficients are then extracted from the obtained IMF sequence through FFT algorithm and inverse cosine transform. The performance of the features extracted from the IMFs components is evaluated using Hidden Markov Models classifier. It has been demonstrated in the proposed approach that cepstral characteristics derived from IMFs components are useful to draw a distinction between voiced cries segments and other acoustical activities. Voiced cry segment is associated to the audible expiratory or inspiratory phase during the cry.

The rest of the paper is organized as follows. Corpus used in this paper is presented in section II. EMD technique is explained in section III while section IV provides description of the proposed approach. Results are presented in section V and the paper is finally concluded in section VI.

A II-2 Corpus

200 cry signals used in this study were recorded in the neonatology departments of several hospitals in Canada and Lebanon. In order to participate in the project, newborns had to have just born up to 6 month old, regardless of gender and prematurity. The corpus collection is still in its middle stage, and will continue for a while. The cry signals were recorded with a sampling rate of 44.1 KHz and a sample resolution of 16 bits. The recorder was placed at a distance of 10 to 30 cm from the baby's mouth. The details of the generated database are described in our previous work (Abou-Abbas, Fersaie Alaie et Tadj, 2015). The recordings collected are produced by 64 babies including both healthy and sick cases, for a total of 200 signals, in different environments and conditions from silent to very noisy: the average duration of one signal is 90 seconds. To evaluate the proposed segmentation system, cry signals were segmented and labeled manually through combined of visual and auditory techniques by authors prior to the approach's development. Overall, our corpus had 66.5 % of voiced cry

sounds and 33.5 % of other acoustic activities like speech, machines' sounds, noise, silence, etc.

A II-3 Empirical mode decomposition

Huang et al. (1998) developed a novel decomposition technique which is known as empirical mode decomposition (EMD) and is used for analyzing nonlinear and non-stationary signals. EMD decomposes the given signal into a set of intrinsic mode functions (IMFs) and a residual function during a so-called sifting process.

The main benefit of EMD is that basic functions are derived directly from the signal. An IMF represents simple oscillatory mode of the signal and is defined so as to draw position of the signal in time. An IMF must satisfy two basic criteria (Huang et al., 1998):

- the number of zero crossing rate and the number of extremes in the whole data set are equal or differ by at most one;

- the mean value of the envelope defined by the local maxima and the envelope defined by local minima is zero at any point.

To extract an IMF from an observed signal, the following sifting approach is adopted (Huang et al., 1998):

1. Identify the extrema of an observed signal x(t) (local maxima and local minima separately);

2. Interpolate the local maxima to form the upper envelope u(t);

3. Interpolate the local minima to form the lower envelope l(t) Calculate local mean value of the upper and lower envelopes $m(t)=\frac{u(t)+l(t)}{2}$;

4. Retrieve the local mean value from the original signal $h(t)=x(t)-m(t)$;

5. If h (t) satisfies the criteria to be an IMF, stop sifting else repeat the procedure with
h (t) regarded as original signal.

This process could be repeated k times until h (t) is an IMF, therefore: c_1 (t) = h_{1k} (t) is the first IMF of the original signal x (t).

The remainder or so-called residue r_1 (t) =x (t)-c_1 (t) is regarded as the new signal to repeat the sifting process for the extraction of the second IMF and so on.

By summing IMFs and residue, the original signal can be reconstructed:

$$x(t)=\sum_{i=1}^{n}c_i(t)+r_n(t)$$

In Figure-A II-1 below the leading five IMFs are shown since all the cry signal components exist in these IMFs.

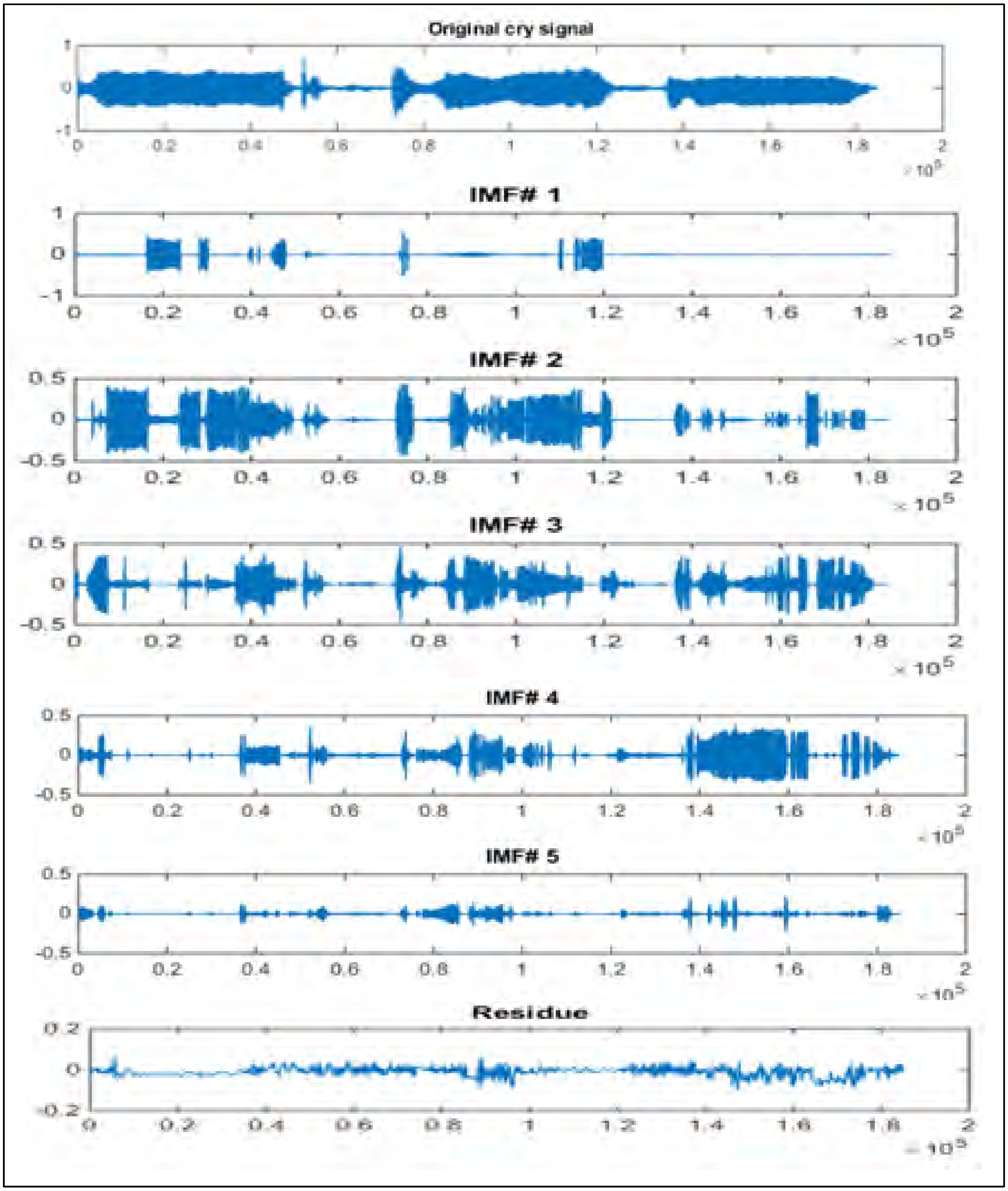


Figure-A II-1 EMD of a cry signal: from top to bottom, the original signal, five IMFs and the residue

A II-4 System Implementation

The proposed segmentation system is divided into three essential stages: Empirical mode decomposition, features extraction, and classification into two distinct classes: Voiced cry and other sounds. The block diagram of the proposed system is illustrated in Figure-A II-2.

To segment subjected cry signals, original signals have been broken down into several small frames of size 50ms with overlap of 35ms. Frames have been decomposed into five IMFs using EMD technique and following parameters: resolution of 50dB and residual energy of 40 dB. Once these IMFs obtained, sets of different IMFs are computed. The 12 Mel-Frequency Cepstral components and their respective energy are derived from the first five IMFs and from the different combination of IMFs. Note that the first IMF corresponds to the fastest oscillation while the last IMF corresponds to the slower one.

The stochastic analysis of the characteristics' space has been carried out using an HMM classifier of 5 states and 32 Gaussians in order to detect voiced cry segments among other acoustic activities. A training step has been achieved using the expectation maximization algorithm for the estimation of the maximum likelihood probabilities. The obtained results will be compared to the standard MFCC characteristics extracted directly from the original signal (Abou-Abbas, Fersaie Alaie et Tadj, 2015).

For the features extraction phase, signals obtained from different set of decompositions from one, two and three IMFs resulting in total 9 sets have been considered and presented in Table-A II-1. From each set, 39 MFCC have been extracted and then applied to HMM for the classification stage. Two groups of recordings were used. A training group used to develop and optimize the model and a testing group used for test. Both groups included recordings contaminated with various types of noise such as respiration sounds, speech and machine equipments. Experiments were designed in order to identify the best IMF combinations.

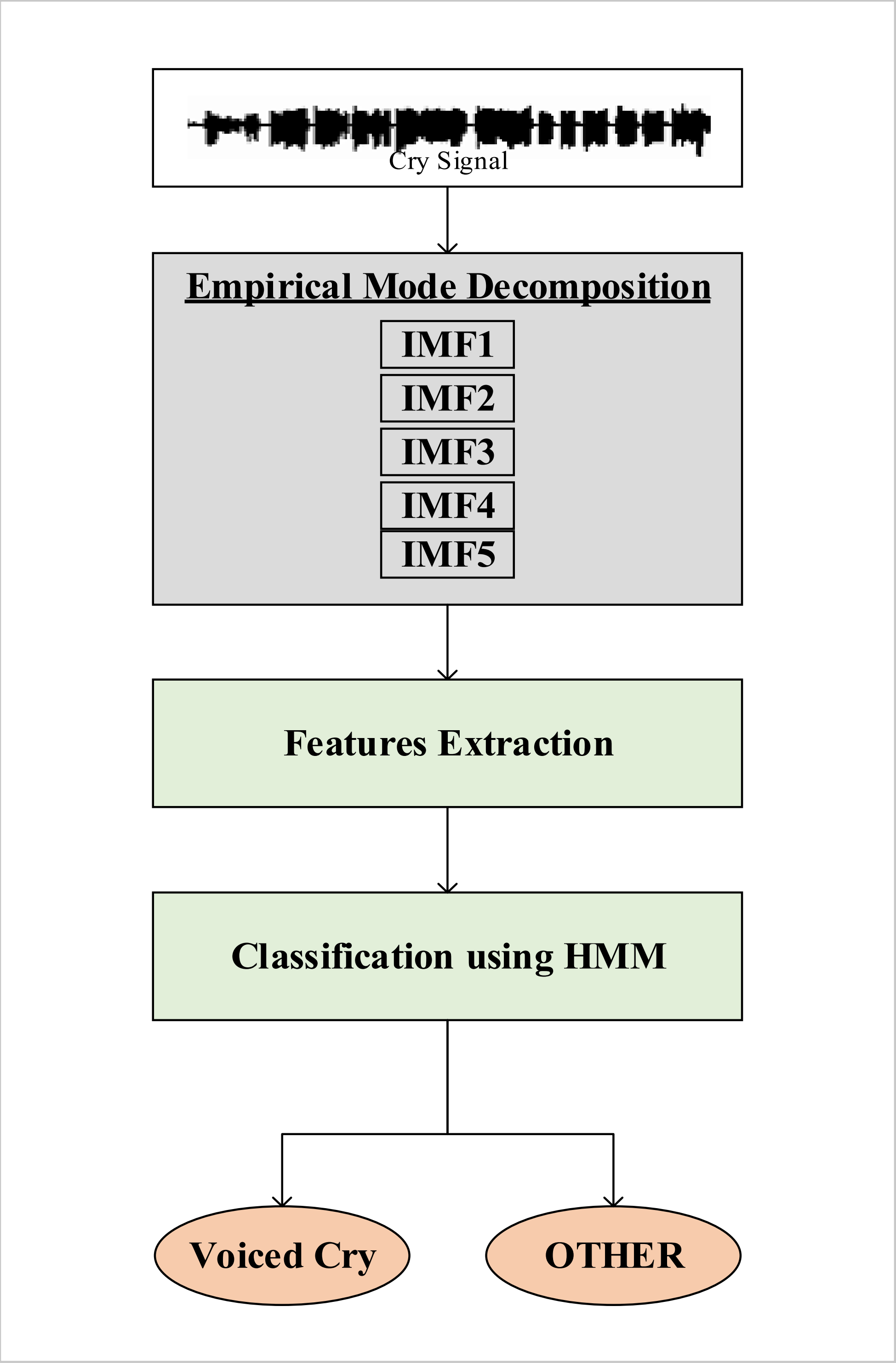


Figure-A II-2 Block diagram of the segmentation system

Table-A II-1 List of different EMD decomposition used

Set 1	IMF1
Set 2	IMF2
Set 3	IMF3
Set 4	IMF4
Set 5	IMF5
Set 6	IMF34=IMF3+IMF4
Set 7	IMF45=IMF4+IMF5
Set 8	IMF234=IMF2+IMF3+IMF4
Set 9	IMF345=IMF3+IMF4+IMF5

A II-5 Results

As we have considered a classification approach that implies training and testing stages, we have decomposed our database equally into two parts: 50% for the training stage and 50% for the testing stage.

To evaluate the performance of our system, a comparison between the automatic transcription file and the reference file (manually segmented) is done.

Three types of errors are calculated: insertion (I), deletion (D) and substitution (S). Readers are referred to our previous work (Abou-Abbas, Fersaie Alaie et Tadj, 2015) for more details about these errors. The accuracy rate of the system could be estimated as follows:

$$AR = \frac{N - D - I - S}{N} \times 100\%$$

Where N represents the total number of labels in the reference file.

The results presented in Table-A II-2 represent that our approach applied to IMF3+IMF4+IMF5 have an improvement of accuracy rate of 8.76% compared with the standard one which is: signal without EMD considered in a previous work (Abou-Abbas, Fersaie Alaie et Tadj, 2015).

We considered each IMF separately. We noted that IMF3, IMF4, and IMF5 contained more discriminant information than the other IMFs and allowed some improvement of accuracy rate in comparison with the standard approach without EMD.

The summation of the IMF3 or IMF5 to the IMF4 allows also some improvement in comparing to the other IMFs considered separately. From other side, the summation of three IMFs: 2, 3, 4 and 3, 4, 5 was considered and we noted that the combination of IMF345 achieves the best compromises for the accuracy rate.

Table-A II-2 Accuracy Rate of the overall system for the different experiments

Experiments	**Accuracy Rate**
Without EMD	77.93
IMF1	68.36
IMF2	73.85
IMF3	77.26
IMF4	79.84
IMF5	78.62
IMF3+IMF4	84.05
IMF2+IMF3+IMF4	79.6
IMF4+IMF5	83.62
IMF3+IMF4+IMF5	86.69

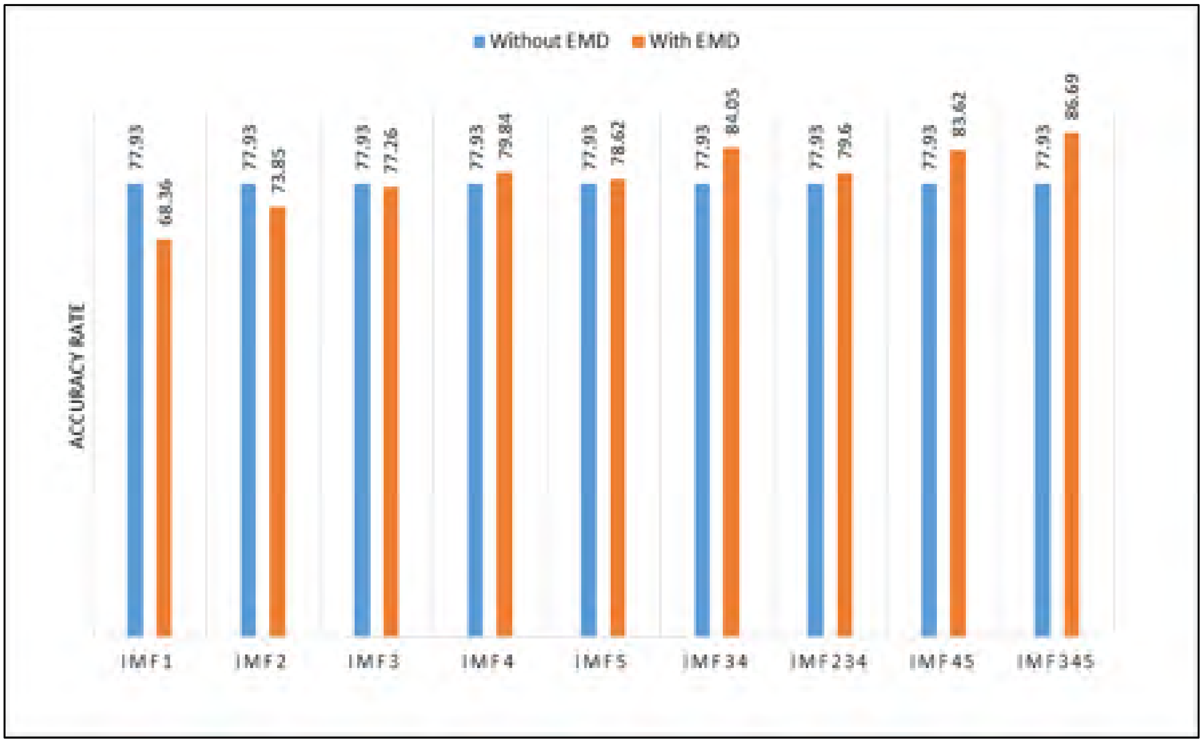


Figure-A II-3 Performance evaluation of segmentation of crying signals- Comparison between experiments based on EMD and without EMD

A II-6 Conclusion

This paper introduces a novel Empirical Mode Decomposition based decision HMM approach for crying signals segmentation. EMD deals with nonlinear and nonstationary signals by decomposing them into intrinsic Mode Functions. Features extracted from the IMFs of crying signals have been found useful in detecting the voiced cry parts.

We concluded that the combination of EMD with MFCC analysis gives interesting features for the classification of voiced crying parts and other acoustics parts. The classification results indicated that studied approach had provided 86.69% accuracy. Future direction of research may include application of EMD combined with wavelet to improve the accuracy of the segmentation system.

Acknowledgment

This work is supported by Bill and Melinda gates foundation. The authors thank the personnel at the Department of neonatology at Sainte Justine hospital and Sahel Hospital for their cooperation in the data collection process.

BIBLIOGRAPHIE

Abdalla, Mahmoud I, et Hanaa S Ali. 2010. « Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden markov models ». *arXiv preprint arXiv:1003.5627*.

Abou-Abbas, L., L. Montazeri, C. Gargour et C. Tadj. 2015. « On the use of EMD for automatic newborn cry segmentation ». In *Advances in Biomedical Engineering (ICABME), 2015 International Conference on*. (16-18 Sept. 2015), p. 262-265.

Abou-Abbas, Lina, Tadj Chakib, Gargour Christian et Leila Montazeri. 2016. « Expiratory and Inspiratory Cries Detection Using Different Signals' Decomposition Techniques ». *Voice Journal*.

Abou-Abbas, Lina, Hesam Fersaei Alaei et Chakib Tadj. 2015. « Segmentation of voiced newborns' cry sounds using Wavelet Packet based features ». In *Electrical and Computer Engineering (CCECE), 2015 IEEE 28th Canadian Conference* (Halifax, Canada, May 3-6).

Abou-Abbas, Lina, Hesam Fersaie Alaie et Chakib Tadj. 2015. « Automatic detection of the expiratory and inspiratory phases in newborn cry signals ». *Biomedical Signal Processing and Control,* vol. 19, p. 35-43.

Ajmera, Jitendra, Iain A. McCowan et H. Bourlard. 2002. « Robust HMM-based speech/music segmentation ». In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. (13-17 May 2002) Vol. 1, p. I-297-I-300.

Ajmera, Jitendra, Iain A. McCowan et Hervé Bourlard. 2003. « Speech/Music Discrimination using Entropy and Dynamism Features in a HMM Classification Framework ». *Speech Communication,* vol. 40, p. 351-363.

Alaie, Hesam Farsaie, Lina Abou-Abbas et Chakib Tadj. 2016. « Cry-based infant pathology classification using GMMs ». *Speech Communication,* vol. 77, p. 28-52.

Alani, A., et M. Deriche. 1999. « A novel approach to speech segmentation using the wavelet transform ». In *Signal Processing and Its Applications, 1999. ISSPA '99. Proceedings of the Fifth International Symposium on*. (1999) Vol. 1, p. 127-130 vol.1.

Amaro-Camargo, Erika, et CarlosA Reyes-García. 2007. « Applying Statistical Vectors of Acoustic Characteristics for the Automatic Classification of Infant Cry ». In *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, sous la dir. de Huang, De-Shuang, Laurent Heutte et Marco Loog. Vol. 4681, p. 1078-1085. Coll. « Lecture Notes in Computer Science »: Springer Berlin Heidelberg. < http://dx.doi.org/10.1007/978-3-540-74171-8_109 >.

Aucouturier, J. J., Y. Nonaka, K. Katahira et K. Okanoya. 2011. « Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden Markov models ». *J Acoust Soc Am,* vol. 130, nº 5, p. 2969-77.

Bachu, RG, S Kopparthi, B Adapa et Buket D Barkana. 2010. « Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy ». In *Advanced Techniques in Computing Sciences and Software Engineering*. p. 279-282. Springer.

Bai, Jing, Jie Wang et Xueying Zhang. 2013. « A parameters optimization method of v-support vector machine and its application in speech recognition ». *Journal of Computers,* vol. 8, nº 1, p. 113-120.

Bajaj, Varun, et Ram Bilas Pachori. 2012. « EEG signal classification using empirical mode decomposition and support vector machine ». In *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011*. p. 623-635. Springer.

Barr, Ronald G. 2006. « Crying behavior and its importance for psychosocial development in children ». *Encyclopedia on early childhood development. Montreal (QC): Centre of Excellence for Early Childhood Development*, p. 1-10.

Becerra, MA, DA Orrego, Carolina Mejia et Edilson Delgado-Trejos. 2012. « Stochastic analysis and classification of 4-area cardiac auscultation signals using Empirical Mode Decomposition and acoustic features ». In *Computing in Cardiology (CinC), 2012*. p. 529-532. IEEE.

Benyassine, Adit, Eyal Shlomot, Huan-Yu Su, Dominique Massaloux, Claude Lamblin et Jean-Pierre Petit. 1997. « ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications ». *Communications Magazine, IEEE,* vol. 35, nº 9, p. 64-73.

Boite, René, et M. Kunt. 1987. « Traitement de la parole ». *Complement au traite d'electricite*.

CanoOrtiz, Sergio D, Daniel I Escobedo Beceiro et Taco Ekkel. « A RADIAL BASIS FUNCTION NETWORK ORIENTED FOR INFANT CRY CLASSIFICATION ».

Carey, M. J., E. S. Parris et H. Lloyd-Thomas. 1999. « A comparison of features for speech, music discrimination ». In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. (15-19 Mar 1999) Vol. 1, p. 149-152 vol.1.

Chang, Chuan-Yu, Chuan-Wang Chang, S Kathiravan, Chen Lin et Szu-Ta Chen. 2016. « DAG-SVM based infant cry classification system using sequential forward floating feature selection ». *Multidimensional Systems and Signal Processing*, p. 1-16.

Charleston-Villalobos, S, AT Aljama-Corrales et R Gonzalez-Camarena. 2006. « Analysis of simulated heart sounds by intrinsic mode functions ». In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. p. 2848-2851. IEEE.

Chu, Yun Yun, Wei Hua Xiong, Wei Wei Shi et Yu Liu. 2013. « The Extraction of Differential MFCC based on EMD ». In *Applied Mechanics and Materials*. Vol. 313, p. 1167-1170. Trans Tech Publ.

Corwin, M. J., B. M. Lester et H. L. Golub. 1996. « The infant cry: what can it tell us? ». *Curr Probl Pediatr,* vol. 26, n° 9, p. 325-334.

Cyril Goizet, Didier Lacombe. 2004. « Le dépistage néonatal ». < http://college-genetique.igh.cnrs.fr/Enseignement/genformclin/depistneonat.html >.

Davis, A., S. Nordholm et R. Togneri. 2006. « Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold ». *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, n° 2, p. 412-424.

Davis, John A. 1983. « Pathological Cry, Stridor, and Cough in Infants ». *Archives of Disease in Childhood,* vol. 58, n° 4, p. 319-320.

Davis, Steven B, et Paul Mermelstein. 1980. « Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences ». *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 28, n° 4, p. 357-366.

Didiot, Emmanuel, Irina Illina, Dominique Fohr et Odile Mella. 2010. « A wavelet-based parameterization for speech/music discrimination ». *Computer Speech & Language,* vol. 24, n° 2, p. 341-357.

Didiot, Emmanuel, Irina Illina, Odile Mella, Dominique Fohr et Jean-Paul Haton. 2006. « Une nouvelle approche fondée sur les ondelettes pour la discrimination parole/musique ». *XXVIes Journées d'Etude sur la Parole-JEP 2006*.

Didiot, Emmanuelle. 2007. « Segmentation parole/musique pour la transcription automatique de parole continue ». Henri Poincaré-Nancy 1, 158 p.

Dongwen, Ying, Yan Yonghong, Dang Jianwu et F. K. Soong. 2011. « Voice Activity Detection Based on an Unsupervised Learning Framework ». *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, n° 8, p. 2624-2633.

Farhid, M, et MA Tinati. 2008. « Robust voice conversion systems using MFDWC ». In *Telecommunications, 2008. IST 2008. International Symposium on*. p. 778-781. IEEE.

Farooq, O, et S Datta. 2001. « Mel filter-like admissible wavelet packet structure for speech recognition ». *IEEE Signal Processing Letters,* vol. 8, n^{o} 7, p. 196-198.

Farooq, Omar, et Sekharjit Datta. 2003. « Phoneme recognition using wavelet based features ». *Information Sciences,* vol. 150, n^{o} 1, p. 5-15.

Farooq, Omar, Sekharjit Datta et MC Shrotriya. 2010. « Wavelet sub-band based temporal features for robust Hindi phoneme recognition ». *International Journal of Wavelets, Multiresolution and Information Processing,* vol. 8, n^{o} 06, p. 847-859.

Farsaie Alaie, Hesam, Lina Abou-Abbas et Chakib Tadj. 2016. « Cry-based infant pathology classification using GMMs ». *Speech Communication,* vol. 77, p. 28-52.

Farsaie Alaie, Hesam, et Chakib Tadj. 2012. « Cry-Based Classification of Healthy and Sick Infants Using Adapted Boosting Mixture Learning Method for Gaussian Mixture Models ». *Modelling and Simulation in Engineering,* vol. 2012, p. 10.

Fort, A., et C. Manfredi. 1998. « Acoustic analysis of newborn infant cry signals ». *Med Eng Phys,* vol. 20, n^{o} 6, p. 432-42.

G Várallyay Jr., A. Illényi, Z. Benyó:. 2008. « The automatic segmentation of the infant cry ». *Előadás kivonatok. Méréstechnikai, Automatizálási és Informatikai Tudományos Egyesület.*

Gales, Mark, et Steve Young. 2008a. « The application of hidden Markov models in speech recognition ». *Foundations and Trends in Signal Processing,* vol. 1, n^{o} 3, p. 195-304.

Gales, Mark, et Steve Young. 2008b. « The Application of Hidden Markov Models in Speech Recognition ». *Foundations and Trends® in Signal Processing,* vol. 1, n^{o} 3, p. 195-304.

Galka, J., et M. Ziolko. 2008. « Wavelets in speech segmentation ». In *Electrotechnical Conference, 2008. MELECON 2008. The 14th IEEE Mediterranean*. (5-7 May 2008), p. 876-879.

Gauvain, Jean-Luc, Lori Lamel et Gilles Adda. 2002. « The LIMSI Broadcast News transcription system ». *Speech Communication,* vol. 37, n^{o} 1–2, p. 89-108.

Gemello, Roberto, Dario Albesano, Loreta Moisa et Renato De Mori. 2001. « Integration of fixed and multiple resolution analysis in a speech recognition system ». In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. Vol. 1, p. 121-124. IEEE.

Golub, Howard L. 1979. « A physioacoustic model of the infant cry and its use for medical diagnosis and prognosis ». *The Journal of the Acoustical Society of America,* vol. 65, n° S1, p. S25-S26.

Gowdy, John N, et Zekeriya Tufekci. 2000. « Mel-scaled discrete wavelet coefficients for speech recognition ». In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. Vol. 3, p. 1351-1354. IEEE.

Grau, S. M., M. P. Robb et A. T. Cacace. 1995. « Acoustic correlates of inspiratory phonation during infant cry ». *J Speech Hear Res,* vol. 38, n° 2, p. 373-81.

Hariharan, M., J. Saraswathy, R. Sindhu, Wan Khairunizam et Sazali Yaacob. 2012. « Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks ». *Expert Systems with Applications,* vol. 39, n° 10, p. 9515-9523.

Hariharan, Muthusamy, R Sindhu et Sazali Yaacob. 2012. « Normal and hypoacoustic infant cry signal classification using time–frequency analysis and general regression neural network ». *Computer methods and programs in biomedicine,* vol. 108, n° 2, p. 559-569.

He, Ling, Margaret Lech, Namunu C Maddage et Nicholas B Allen. 2011. « Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech ». *Biomedical Signal Processing and Control,* vol. 6, n° 2, p. 139-146.

Huang, Liang-sheng, et Chung-ho Yang. 2000. « A novel approach to robust speech endpoint detection in car environments ». In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. Vol. 3, p. 1751-1754. IEEE.

Huang, Norden E, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung et Henry H Liu. 1998. « The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis ». In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 454, p. 903-995. The Royal Society.

Huang, Xuedong, Alex Acero et Hsiao-Wuen Hon. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 960 p.

Hubin-Gayte. 2012. *Tout savoir sur son bébé- idées reçues sur les enfants de 0 à 3 ans.*, 160 p.

Huiqun, Deng, et D. O'Shaughnessy. 2007. « Voiced-Unvoiced-Silence Speech Sound Classification Based on Unsupervised Learning ». In *Multimedia and Expo, 2007 IEEE International Conference on*. (2-5 July 2007), p. 176-179.

ITU, A. 1996. « silence compression scheme for G. 729 optimized for terminals conforming to recommendation V. 70 ». *ITU-T Recommendation G,* vol. 729.

J. Pinquier, C. Sénac et R. André-Obrecht. 2002a. « Indexation de la bande sonore :recherche des composantes Parole et Musique ». In *RFIA*. (Angers, France).

J. Pinquier, J.L. Rouas , R. André-Obrecht. 2002b. « Fusion de paramètres pour une classification automatique Parole / Musique robuste ».

John R. Deller, Jr., John G. Proakis et John H. Hansen. 1993. *Discrete Time Processing of Speech Signals*. Prentice Hall PTR, 800 p.

Juang, Biing Hwang, et Laurence R Rabiner. 1991. « Hidden Markov models for speech recognition ». *Technometrics,* vol. 33, n^{o} 3, p. 251-272.

Kaiser, James F. 1990. « On a simple algorithm to calculate theenergy'of a signal ». In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. p. 381-384. IEEE.

Khaled, Zaabi. 2004. « Implémentation d'une méthode de reconnaissance de la parole sur le processeur de traitement numérique du signal TMS320C6711 ». Montreal, Ecole de Technologie supérieure, 149 p.

Kim, Myung Jong, Younggwan Kim, Seungki Hong et Hoirin Kim. 2013. « ROBUST detection of infant crying in adverse environments using weighted segmental two-dimensional linear frequency cepstral coefficients ». In *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. p. 1-4. IEEE.

Kuo, Kevin. 2010. « Feature extraction and recognition of infant cries ». In *Electro/Information Technology (EIT), 2010 IEEE International Conference on*. p. 1-5. IEEE.

Kyoung-Ho, Woo, Yang Tae-Young, Park Kun-Jung et Lee Chungyong. 2000. « Robust voice activity detection algorithm for estimating noise spectrum ». *Electronics Letters,* vol. 36, n^{o} 2, p. 180-181.

L.R. Rabiner, M.R. Sambur. 1975. « An Algorithm for Determining the Endpoints for Isolated Utterances ». *The bell system technical journal,* vol. 54, n^{o} 2, p. 297-315.

LaGasse, L. L., A. R. Neal et B. M. Lester. 2005. « Assessment of infant cry: acoustic cry analysis and parental perception ». *Ment Retard Dev Disabil Res Rev,* vol. 11, n^{o} 1, p. 83-93.

Lederman, Dror. 2010. « Estimation of Infants' Cry Fundamental Frequency using a Modified SIFT algorithm ». *CoRR,* vol. abs/1009.2796.

Lewis, Finley R. 2007. *Focus on Nonverbal Communication Research*. Nova Publishers.

Lind, J, V Vuorenkoski, G Rosberg, TJ Partanen et O Wasz-Höckert. 1970. « Spectographic analysis of vocal response to pain stimuli in infants with Down's syndrome ». *Developmental Medicine & Child Neurology,* vol. 12, nº 4, p. 478-486.

Liu, Dan, Lie Lu et HongJiang Zhang. 2003. « Automatic mood detection from acoustic music data ». In *ISMIR*. p. 81-87.

Liu, Lihan, Haibin Wang, Yan Wang, Ting Tao et Xiaochen Wu. 2010. « Notice of Retraction Feature analysis of heart sound based on the improved Hilbert-Huang Transform ». In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*. Vol. 6, p. 378-381. IEEE.

Lu, Lie, Hao Jiang et HongJiang Zhang. 2001. « A robust audio classification and segmentation method ». In *Proceedings of the ninth ACM international conference on Multimedia*. (Ottawa, Canada), p. 203-211. 500173: ACM.

Lu, Lie, Dan Liu et Hong-Jiang Zhang. 2006. « Automatic mood detection and tracking of music audio signals ». *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, nº 1, p. 5-18.

Mallat, Stéphane. 2000. *Une exploration des signaux en ondelettes*. Editions Ecole Polytechnique.

Manfredi, C, L Bocchi, S Orlandi, L Spaccaterra et GP Donzelli. 2009. « High-resolution cry analysis in preterm newborn infants ». *Medical engineering & physics,* vol. 31, nº 5, p. 528-532.

Maragos, Petros, James F Kaiser et Thomas F Quatieri. 1993a. « Energy separation in signal modulations with application to speech analysis ». *Signal Processing, IEEE Transactions on,* vol. 41, nº 10, p. 3024-3051.

Maragos, Petros, James F Kaiser et Thomas F Quatieri. 1993b. « On amplitude and frequency demodulation using energy operators ». *Signal Processing, IEEE Transactions on,* vol. 41, nº 4, p. 1532-1550.

Martis, R. J., U. R. Acharya, J. H. Tan, A. Petznick, R. Yanti, C. K. Chua, E. Y. Ng et L. Tong. 2012. « Application of empirical mode decomposition (emd) for automated detection of epilepsy using EEG signals ». *Int J Neural Syst,* vol. 22, nº 6, p. 1250027.

Marzinzik, Mark, et Birger Kollmeier. 2002. « Speech pause detection for noise spectrum estimation by tracking power envelope dynamics ». *Speech and Audio Processing, IEEE Transactions on,* vol. 10, nº 2, p. 109-118.

Messaoud, Ali, et Chakib Tadj. 2010. « A cry-based babies identification system ». In *Image and Signal Processing*. p. 192-199. Springer.

Michelsson, Katarina, et Oliver Michelsson. 1999. « Phonation in the newborn, infant cry ». *International Journal of Pediatric Otorhinolaryngology,* vol. 49, Supplement 1, n° 0, p. S297-S301.

Misiti, Michel, Yves Misiti, Georges Oppenheim et Jean-Michel Poggi. 1996. « Wavelet toolbox ».

Moukadem, A, A Dieterlen, N Hueber et C Brandt. 2011. « Localization of heart sounds based on S-transform and radial basis function neural network ». In *15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011)*. p. 168-171. Springer.

Moukadem, Ali, Alain Dieterlen et Christian Brandt. 2012. « Study of Two Feature Extraction Methods to Distinguish between the First and the Second Heart Sounds ». In *BIOSIGNALS*. p. 346-350.

Orlandi, S., C. Manfredi, L. Bocchi et M. L. Scattoni. 2012. « Automatic newborn cry analysis: A Non-invasive tool to help autism early diagnosis ». In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. p. 2953-2956. In *Scopus*. < http://www.scopus.com/inward/record.url?eid=2-s2.0-84870849024&partnerID=40&md5=36e2fa2a32c09e64179544710079f383 >.

Orlandi, Silvia, PH Dejonckere, Jean Schoentgen, Jean Lebacq, N Rruqja et Claudia Manfredi. 2013. « Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring ». *Biomedical Signal Processing and Control,* vol. 8, n° 6, p. 799-810.

Orlikoff, R. F., R. J. Baken et D. H. Kraus. 1997. « Acoustic and physiologic characteristics of inspiratory phonation ». *J Acoust Soc Am,* vol. 102, n° 3, p. 1838-45.

Orozco-García, José, et Carlos A Reyes-García. 2003a. « A study on the Recognition of Patterns of Infant Cry for the Identification of Deafness in Just Born Babies with neural Networks ». In *Progress in Pattern Recognition, Speech and Image Analysis*. p. 342-349. Springer.

Orozco-García, José, et CarlosA Reyes-García. 2003b. « A Study on the Recognition of Patterns of Infant Cry for the Identification of Deafness in Just Born Babies with Neural Networks ». In *Progress in Pattern Recognition, Speech and Image Analysis*, sous la dir. de Sanfeliu, Alberto, et José Ruiz-Shulcloper. Vol. 2905, p. 342-349. Coll. « Lecture Notes in Computer Science »: Springer Berlin Heidelberg. < http://dx.doi.org/10.1007/978-3-540-24586-5_42 >.

Papoulis, Athanasios. 1991. « Probability, random variables, and stochastic processes ». *McGraw-Hill series in electrical engineering. Communications and signal processing*.

Petroni, Marco, Alfred S Malowany, C Celeste Johnston et Bonnie J Stevens. 1995. « Classification of infant cry vocalizations using artificial neural networks (ANNs) ». In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. Vol. 5, p. 3475-3478. IEEE.

Pinquier, Julien. 2004. « Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle ». Institut de Recherche en Informatique de Toulouse.

Pinquier, Julien, et Nicolas Chambert. 2001. « La première étape d'un système d'Indexation audio (Parole/Musique/Bruit) ».

Proctor, Adele. 1984. « Pathological cry, stridor and cough in infants: A clinical-acoustic study, J. Hirschberg & T. Szende, Akademiai Kiado, 1982.(156 pp., 109 Illustrations: Two 33 1/3 RPM Records, US $28.00).(Distibutors: Kultura, Hungarian Foreign Trading Company, POB 149, H-1389, Budapest.) ». *Infant Mental Health Journal,* vol. 5, n^{o} 4, p. 245-247.

Rabiner, L. 1989. « A tutorial on hidden Markov models and selected applications in speech recognition ». *Proceedings of the IEEE,* vol. 77, n^{o} 2, p. 257-286.

Rabiner, Lawrence R, et Marvin R Sambur. 1975. « An algorithm for determining the endpoints of isolated utterances ». *Bell System Technical Journal,* vol. 54, n^{o} 2, p. 297-315.

Rabiner, Lawrence R., et B. H. Juang. 1993a. « Fundamentals of speech recognition ». *Prentice-Hall signal processing series*.

Rabiner, Lawrence R., et B. H. Juang. 1993b. *Fundamentals of speech recognition* (1993). Englewood Cliffs, N.J.: PTR Prentice Hall, xxxv, 507 p. p.

Ramírez, Javier, José C. Segura, Carmen Benítez, Ángel de la Torre et Antonio Rubio. 2004. « Efficient voice activity detection algorithms using long-term speech information ». *Speech Communication,* vol. 42, n^{o} 3–4, p. 271-287.

Ramona, Mathieu. 2010. « Classification automatique de flux radiophoniques par Machines à Vecteurs de Support ». Telecom ParisTech, 200 p.

Razik, Joseph. 2003. « segmentation parole/musique ». Henri Poincaré-Nancy 46 p.

Razik, Joseph, Christine Sénac, Dominique Fohr, Odile Mella et Nathalie Parlangeau-Vallès. 2003. « Comparison of Two Speech/Music Segmentation Systems For Audio Indexing on the Web ». In *World Multiconference on Systemics, Cybernetics and Informatics - SCI'2003*. p. 6p.

Reyes-Galaviz, Orion Fausto, Sergio Daniel Cano-Ortiz et Carlos A Reyes-Garcia. 2008. « Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies ». In *Artificial Intelligence, 2008. MICAI'08. Seventh Mexican International Conference on*. p. 330-335. IEEE.

Reynolds, Douglas, et Richard C Rose. 1995. « Robust text-independent speaker identification using Gaussian mixture speaker models ». *Speech and Audio Processing, IEEE Transactions on,* vol. 3, n° 1, p. 72-83.

Rui, x, M. A. z, L. C. Altamirano, C. A. Reyes et O. Herrera. 2010. « Automatic identification of qualitatives characteristics in infant cry ». In *Spoken Language Technology Workshop (SLT), 2010 IEEE*. (12-15 Dec. 2010), p. 442-447.

Rúız, Maŕıa Antonia, Carlos Alberto Reyes et Luis Carlos Altamirano. 2012. « On the implementation of a method for automatic detection of infant cry units ». *Procedia Engineering,* vol. 35, p. 217-222.

Sahak, R, YK Lee, W Mansor, AIM Yassin et A Zabidi. 2010. « Optimized Support Vector Machine for classifying infant cries with asphyxia using Orthogonal Least Square ». In *Computer Applications and Industrial Electronics (ICCAIE), 2010 International Conference on*. p. 692-696. IEEE.

Saïdi, Mohamed, Olivier Pietquin et Régine André-Obrecht. 2010. « Application of the EMD decomposition to discriminate nasalized vs. vowels phones in French ». In *Proceedings of the International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA 2010)*. p. 128-132.

Saraswathy, J., M. Hariharan, V. Vijean, S. Yaacob et W. Khairunizam. 2012. « Performance comparison of Daubechies wavelet family in infant cry classification ». In *Signal Processing and its Applications (CSPA), 2012 IEEE 8th International Colloquium on*. (23-25 March 2012), p. 451-455.

Sarikaya, Ruhi, Bryan L Pellom et John HL Hansen. 1998. « Wavelet packet transform features with application to speaker identification ». In *IEEE Nordic Signal Processing Symposium*. p. 81-84. Citeseer.

Saunders, J. 1996. « Real-time discrimination of broadcast speech/music ». In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. (7-10 May 1996) Vol. 2, p. 993-996 vol. 2.

Scheirer, E., et M. Slaney. 1997. « Construction and evaluation of a robust multifeature speech/music discriminator ». In *Acoustics, Speech, and Signal Processing, 1997.*

ICASSP-97., 1997 IEEE International Conference on. (21-24 Apr 1997) Vol. 2, p. 1331-1334 vol.2.

Scherer, Klaus R. 1982. « The assessment of vocal expression in infants and children ». *Measuring emotions in infants and children*, p. 127-163.

Shen, Jia-Lin, Jeih-Weih Hung et Lin-Shan Lee. 1998. « Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments ». In *Fifth International Conference on Spoken Language Processing*.

Shete, DS, SB Patil et PSB Patil. 2014. « Zero crossing rate and Energy of the Speech Signal of Devanagari Script ». *IOSR-JVSP,* vol. 4, n^{o} 1, p. 1-5.

Shi, Wei Wei, Wei Hua Xiong et Wei Chen. 2014. « Speech Recognition Algorithm Based on Empirical Mode Decomposition and RBF Neural Network ». In *Advanced Materials Research*. Vol. 831, p. 465-469. Trans Tech Publ.

Siafarikas, M., T. Ganchev et N. Fakotakis. 2007. « Wavelet Packet Bases for Speaker Recognition ». In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*. (29-31 Oct. 2007) Vol. 2, p. 514-517.

Siafarikas, Mihalis, Todor Ganchev et Nikos Fakotakis. « Objective wavelet packet features for speaker verification ». In.

Siafarikas, Mihalis, Todor Ganchev et Nikos Fakotakis. 2004. « Wavelet packet based speaker verification ». In *ODYSSEY04-The Speaker and Language Recognition Workshop*.

Sjölander, Kåre, et Jonas Beskow. 2000. « Wavesurfer-an open source speech tool ». In *INTERSPEECH*. p. 464-467.

Soltis, Joseph. 2004. « The signal functions of early infant crying ». *Behavioral and Brain Sciences,* vol. 27, n^{o} 04, p. 443-458.

Teager, HM, et SM Teager. 1989. « Evidence for nonlinear speech production mechanisms in the vocal tract ». *Proc. NATO Advanced Study Institute on Speech Production and Speech Modeling*, p. 214-261.

Tu, Binbin, et Fengqin Yu. 2012. « Speech emotion recognition based on improved MFCC with EMD ». *Jisuanji Gongcheng yu Yingyong(Computer Engineering and Applications),* vol. 48, n^{o} 18, p. 119-122.

Várallyay, György. 2005. « Future Prospects of the Application of the Infant Cry in the Medicine ». *periodica polytechnica,* vol. 50, n^{o} 1-2, p. 47-62.

Várallyay, György. 2006. « Future Prospects of the Application of the Infant Cry in the Medicine ». *Electrical Engineering,* vol. 50, n^{o} 1-2, p. 47-62.

Várallyay, György, András Illényi et Zoltán Benyó. 2009. « Automatic infant cry detection ». In *MAVEBA*. p. 11-14.

Várallyay Jr, G, A Illényi et Z Benyó. 2008. « The automatic segmentation of the infant cry ». In *Proc. BUDAMED'08 Conference, Budapest*. p. 28-32.

Verduzco-Mendoza, A., E. Arch-Tirado, C. A. Reyes-Garcia, J. Leybon-Ibarra et J. Licona-Bonilla. 2012. « Spectrographic cry analysis in newborns with profound hearing loss and perinatal high-risk newborns ». *Cir Cir,* vol. 80, n° 1, p. 3-10.

Wang, Hongzhi, Yuchao Xu et Meijing Li. 2011. « Study on the MFCC similarity-based voice activity detection algorithm ». In *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*. p. 4391-4394. IEEE.

Wang, W. Q., W. Gao et D. W. Ying. 2003. « A fast and robust speech/music discrimination approach ». In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*. (15-18 Dec. 2003) Vol. 3, p. 1325-1329 vol.3.

Wasz-Höckert, Ole, Katarina Michelsson et John Lind. 1985. « Twenty-Five Years of Scandinavian Cry Research ». In *Infant Crying*, sous la dir. de Lester, BarryM, et C. F. Zachariah Boukydis. p. 83-104. Springer US. < http://dx.doi.org/10.1007/978-1-4613-2381-5_4 >.

Wermke, K., W. Mende, C. Manfredi et P. Bruscaglioni. 2002. « Developmental aspects of infant's cry melody and formants ». *Med Eng Phys,* vol. 24, n° 7-8, p. 501-14.

West, K., et S. J. Cox. 2004. « Features and Classifiers for the Automatic Classification of Musical Audio Signals ». In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*. (Barcelona, Spain).

Williams, Gethin, et Daniel P. W. Ellis. 1999. « Speech/music discrimination based on posterior probability features ». In *European Conference on Speech Communication and Technology*. (Budapest, Hungary, September 5-9, 1999). ESCA.

Woodland, P. C., T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk et S. J. Young. 1998. « Experiments in broadcast news transcription ». In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*. (12-15 May 1998) Vol. 2, p. 909-912 vol.2.

Wu, Zhaohua, et Norden E Huang. 2009. « Ensemble empirical mode decomposition: a noise-assisted data analysis method ». *Advances in adaptive data analysis,* vol. 1, n° 01, p. 1-41.

Yamamoto, Shota, Yasunari Yoshitomi, Masayoshi Tabuse, Kou Kushida et Taro Asada. 2010. « Detection of baby voice and its application using speech recognition system and fundamental frequency analysis ». In *Proc. of 10th WSEAS Int. Conf. on Applied Computer Science*. Vol. 341345.

Yamamoto, Shota, Yasunari Yoshitomi, Masayoshi Tabuse, Kou Kushida et Taro Asada. 2013. « Recognition of a Baby's Emotional Cry Towards Robotics Baby Caregiver ». *International Journal of Advanced Robotic Systems,* vol. 10.

Young, Steve, et Gunnar Evermann. 1996. « The HTK book ».

Zabidi, A., W. Mansor, Khuan Lee Yoot, R. Sahak et F. Y. A. Rahman. 2009a. « Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism ». In *Signal Processing & Its Applications, 2009. CSPA 2009. 5th International Colloquium on*. (6-8 March 2009), p. 204-208.

Zabidi, Azlee, Lee Yoot Khuan, Wahidah Mansor, Ihsan Mohd Yassin et Rohilah Sahak. 2010. « Classification of infant cries with asphyxia using multilayer perceptron neural network ». In *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*. Vol. 1, p. 204-208. IEEE.

Zabidi, Azlee, Wahidah Mansor, Lee Yoot Khuan, Rohilah Sahak et Farah Yasmin Abd Rahman. 2009b. « Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism ». In *Signal Processing & Its Applications, 2009. CSPA 2009. 5th International Colloquium on*. p. 204-208. IEEE.

Zhang Tong, Kuo C, J.C. 1998. « Hierarchical system for content-based audio classification and retrieval ». In *Conference on Multimedia storage and Archiving Systems III*. Vol. 3527, p. 398-409.

Ziółko, Bartosz, Suresh Manandhar, Richard C Wilson et Mariusz Ziółko. 2006. « Wavelet method of speech segmentation ». In *Proceedings of 14th European Signal Processing Conference EUSIPCO, Florence.*