Proceeding of the IEEE 28th
Canadian Conference on Electrical and Computer Engineering
Halifax, Canada, May 3-6, 2015

Segmentation of voiced newborns' cry sounds using Wavelet Packet based features

Lina Abou-Abbas, Hesam Fersai Alaei, and Chakib Tadj

***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 %.**

I. Introduction

Plusieurs études ont établi l'existence d'une information importante dans le signal du cri d'un nouveau-né [1-4]. 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) [5-7].

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) [1].

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 [3]. 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 [2, 9, 10]. 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 [11,12], 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 [13] 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 [14-16] calculé à partir de l'analyse en paquets d'ondelettes ainsi qu'une étape de classification automatique par l'approche *HMM*.

Manuscript received November 25, 2014.
All Authors are with the Electrical Engineering department, École de Technologie Supérieure, Montreal, Canada.
Lina Abou-Abbas (E-mail: Lina.Abou-Abbas.1@etsmtl.net).
Chakib Tadj (E-mail: Chakib.Tadj@etsmtl.ca).
Hesam Fersaie Alaie (E-mail: Hesam.farsaie-alaie.1@etsmtl.net).

978-1-4799-5829-0/15/$31.00 ©2015 IEEE

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.

II. EXTRACTION DES CARACTERISTIQUES 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.
Le schéma qui suit montre une architecture d'une décomposition en paquets d'ondelettes :

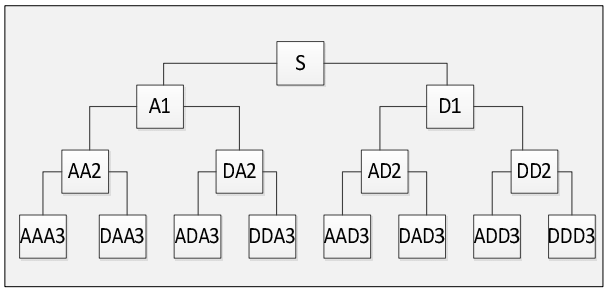

Fig. 1. Arbre de niveau 3 obtenu par décomposition en paquets d'ondelettes.

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 [17]. Ainsi, en se basant sur les coefficients d'ondelettes, nous calculons deux paramètres d'énergie :

1) 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.

2) 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 [18]. 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 [14-16]. 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 :

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

III. 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 I, contient 44.6% d'expiration, 21.9% d'inspiration et 33.5% d'autres activités acoustiques.

TABLEAU I
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

IV. SYSTEME 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 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).

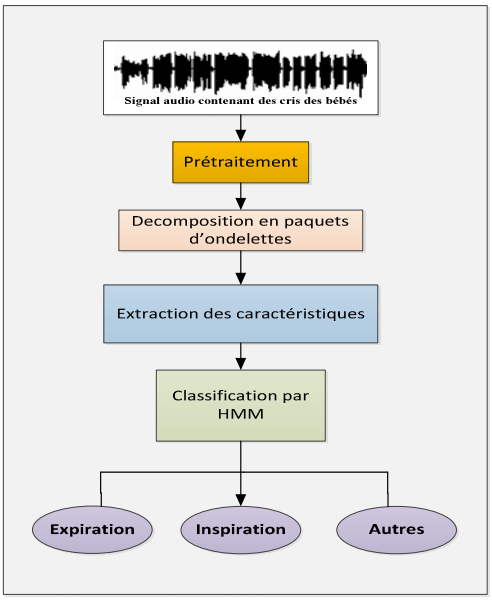

Fig. 2. Architecture du système de segmentation proposé

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 2^n 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 2^n 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 [19].

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.

V. RESULTATS EXPERIMENTAUX ET DISCUSSION

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 II, III, et IV exposent les résultats de la précision de la segmentation proposée en utilisant les niveaux de décompositions 5, 6 et 7 respectivement.

TABLEAU II
RESULTATS DE LA SEGMENTATION AUTOMATIQUE POUR LE NIVEAU DE DECOMPOSITION 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

TABLEAU III
RESULTATS DE LA SEGMENTATION AUTOMATIQUE POUR LE NIVEAU DE DECOMPOSITION 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

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.

TABLEAU IV
RESULTATS DE LA SEGMENTATION AUTOMATIQUE POUR LE NIVEAU DE DECOMPOSITION 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

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%.

VI. 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.

REFERENCES

[1] H. L. Golub, A Physioacoustic Model of the Infant Cry and Its Use for Medical Diagnosis and Prognosis: MIT Press, 1980.

[2] A. Proctor, "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, U.S. $28.00). (Distibutors: Kultura, Hungarian Foreign Trading Company, P.O.B. 149, H-1389, Budapest.)," Infant Mental Health Journal, vol. 5, pp. 245-247, 1984.

[3] R. F. Orlikoff, R. J. Baken, and D. H. Kraus, "Acoustic and physiologic characteristics of inspiratory phonation," J Acoust Soc Am, vol. 102, pp. 1838-45, Sep 1997.

[4] G. Várallyay, "Future Prospects of the Application of the Infant Cry in the Medicine," periodica polytechnica, vol. 50, pp. 47-62, april 28,2005 2005.

[5] Rui, x, M. A. z, L. C. Altamirano, C. A. Reyes, and O. Herrera, "Automatic identification of qualitatives characteristics in infant cry," in Spoken Language Technology Workshop (SLT), 2010 IEEE, 2010, pp. 442-447.

[6] H. Farsaie Alaie and C. Tadj, "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, 2012.

[7] M. Hariharan, J. Saraswathy, R. Sindhu, W. Khairunizam, and S. Yaacob, "Infant cry classification to identify asphyxia using time-frequency analysis and radial basis neural networks," Expert Systems with Applications, vol. 39, pp. 9515-9523, 8// 2012.

[8] A. Fort and C. Manfredi, "Acoustic analysis of newborn infant cry signals," Med Eng Phys, vol. 20, pp. 432-42, Sep 1998.

[9] K. Michelsson and O. Michelsson, "Phonation in the newborn, infant cry," International Journal of Pediatric Otorhinolaryngology, vol. 49, Supplement 1, pp. S297-S301, 10/5/ 1999.

[10] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni, "Developmental aspects of infant's cry melody and formants," Med Eng Phys, vol. 24, pp. 501-14, Sep-Oct 2002.

[11] A. I. G Várallyay Jr., Z. Benyó:, "The automatic segmentation of the infant cry," Előadás kivonatok. Méréstechnikai, Automatizálási és Informatikai Tudományos Egyesület, 2008.

[12] A. I. G Várallyay Jr., Z. Benyó:, "Automatic infant cry detection," in Models and analysis of vocal emissions for biomedical applications: 6th international workshop, Firenze, 2009.

[13] R. Cohen and Y. Lavner, "Infant cry analysis and detection," in Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of, 2012, pp. 1-5.

[14] J. F. Kaiser. "On a simple algorithm to calculate the 'Energy' of a signal." In Proc. IEEE ICASSP, pages 381-384, Albuquerque, NM, USA, 1990.

[15] P. Maragos, J.F. Kaiser and T.F. Quatieri: Energy separation in signal modulations with application to speech analysis. IEEE Trans. on Signal Processing, 41(10):3024–3051, October 1993

[16] H.M. Teager and S.M. Teager: Evidence for nonlinear speech production mechanisms in the vocal tract. In Proc. NATO Advanced Study Institute on Speech Production and Speech Modeling, pages 214–261, July 1989.

[17] S. G. Mallat, Une exploration des signaux en ondelettes / : Stéphane Mallat. Palaiseau, France: Éditions de l'École polytechnique, 2000.

[18] P. Maragos, J. F. KAISER, and T. F. Quatieri. On amplitude and frequency demodulation using energy operators. IEEE Trans. Sig.Process. 41:1532-1550, 1993.

[19] Rabiner, Lawrence R., et B. H. Juang. 1993. « Fundamentals of speech recognition ». Prentice-Hall signal processing series.