Infant Cry Detection in Adverse Acoustic Environments by Using Deep Neural Networks

Daniele Ferretti, Marco Severini, Emanuele Principi, Annalisa Cenci, and Stefano Squartini
Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy
Email: d.ferretti@pm.univpm.it, {m.severini,e.principi,a.cenci,s.squartini,}@univpm.it

***Abstract*—The amount of time an infant cries in a day helps the medical staff in the evaluation of his/her health conditions. Extracting this information requires a cry detection algorithm able to operate in environments with challenging acoustic conditions, since multiple noise sources, such as interferent cries, medical equipments, and persons may be present. This paper proposes an algorithm for detecting infant cries in such environments. The proposed solution is a multiple stage detection algorithm: the first stage is composed of an eight-channel filter-and-sum beamformer, followed by an Optimally Modified Log-Spectral Amplitude estimator (OMLSA) post-filter for reducing the effect of interferences. The second stage is the Deep Neural Network (DNN) based cry detector, having audio Log-Mel features as inputs. A synthetic dataset mimicking a real neonatal hospital scenario has been created for training the network and evaluating the performance. Additionally, a dataset containing cries acquired in a real neonatology department has been used for assessing the performance in a real scenario. The algorithm has been compared to a popular approach for voice activity detection based on Long-Term Spectral Divergence, and the results show that the proposed solution achieves superior detection performance both on synthetic data and on real data.**

I. Introduction

Acoustic analysis of infant cries has been devoted particular attention in the last years, since it offers a non-invasive and cost-effective method for monitoring the health conditions of a newborn. Indeed, cry signals contain valuable information related to the state of an infant [1]. Cry detection (or segmentation), consists in identifying the portions of the audio signals where a cry is present [2]–[4]. Further analysis can evaluate whether an infant is affected by a pathology or not, i.e., a binary classification problem whose two classes are "healthy" and "non-healthy", the latter class representing a specific pathology or not [5]. An even more advanced cry analysis, i.e., a multiclass classification problem, can detect either the infant's pathology [6], [7], or the cause of a cry (e.g., hunger, pain) [8], [9].

This paper focuses on cry detection, a fundamental part of any cry analysis system. Several works in the literature addressed this task [10], [11], and more recently machine learning methods have been proposed [2]–[4]. Among them, Cohen and Lavner [12] proposed an algorithm based on k-nearest neighbors to classify each frame as cry or non-cry for alerting parents when infants are being left alone in closed apartments or vehicles. Several acoustic features have been used, such as the fundamental frequency, mel-frequency cepstral coefficients (MFCCs) [13], among others. The evaluation corpus is synthetic, and it considers street noises. In [4], Abou-Abbas *et al.* presented a method based on MFCCs and hidden Markov models (HMMs). In a later work [2], they introduced a signal decomposition stage and extracted the related features. In both papers, the corpus included cries acquired in real environments and different acoustic conditions. In [3], Naithani *et al.* also used MFCCs and HMMs, but they augmented them with fundamental frequency and aperiodicity. In their case, also, the experimentation dataset contained cries recorded in a real environment, in presence of noise. Conversely, in [14] the extraction of Log-Mel features and the detection of cry sounds is achieved by using a deep neural network (DNN) composed of three convolutional layers and one fully-connected layer. The evaluation dataset contained cries recorded in a domestic environment. Torres *et al.* [15] modified the neural network topology of [14] by introducing dropout and batch normalization. The experiments were conducted on a synthetic dataset containing cries collected from free on-line resources.

In the examined literature, the algorithm robustness against noise or reverberation is not explicitly addressed, thus detection issues may arise when used in maternity wards or Neonatal Intensive Care Units (NICUs). In this paper, we propose a cry detection algorithm based on DNN able to operate in realistic acoustic environments for identifying the portions of the audio signals where a cry is present. The robustness of the detection algorithm against noise is increased by acquiring cry signals with an eight-channel circular array, and by pre-processing them with a linear-constraint minimum-variance (LCMV) beamformer [16] followed by the optimally modified log-spectral amplitude (OMLSA) post-filter [16]. The DNN-based cry detector operates on Log-Mel features and is composed of 3 convolutional layers followed by 1 fully connected layer. The experiments have been conducted on a "Simulated" dataset and on a "Real" dataset. The first has been created by generating the impulse responses of a real NICU, and by synthetically adding several kinds of noises at different Signal to Noise Ratios (SNRs) to clean cry recordings. The "Real" dataset contains recordings acquired in a NICU by using a circular microphone array positioned above the target crib. By training the DNN on the Simulated dataset only, we demonstrate that the proposed approach is effective in real scenarios and it does not require a large amount of data to be acquired in sensible environments such as NICUs. This strategy can be considered as an extreme data augmentation technique, since a model of the target environment

ISBN 978-90-827970-1-5

is created and the related synthetic data are generated and used for training. This approach has been compared to the voice-activity detector described in [17]. The results show that in both datasets the proposed approach outperforms the comparative algorithm. Compared to the recent works on infant cry detection, the contributions of this paper are: the introduction of a signal enhancement stage for pre-processing the acoustic signals, the use of a DNN-based classifier, data augmentation to ease the data requirements for cry detection in real conditions, and the evaluation on a Simulated dataset in controlled conditions and on a Real dataset in a realistic scenario.

The outline of the paper is the following. Section II describes in details the beamformer, the post filter, and the neural network cry detector. The comparative method is briefly introduced in Section III, whereas Section IV presents the experiments performed to evaluate the proposed approach, and the obtained results. Finally, Section V concludes the paper.

II. The Proposed Approach

A block-scheme of the proposed approach is shown in Fig. 1. Acoustic signals are acquired with an eight-channel circular microphone array and processed by a filter-and-sum beamformer for reducing coherent noise source, followed by the OMLSA post-filter that reduces residual diffuse noise. The feature extraction stage calculates the Log-Mel spectrogram, whereas the DNN takes in multiple frames and classifies the central one as a cry frame or not, exploiting the information from temporally adjacent frames. The details of the signal enhancement stage and of the DNN-based classifier follows.

A. Signal Enhancement

1) Beamformer: Cry signals are acquired by using an eight-channel circular microphone array. This allows to apply a beamforming algorithm for reducing the effect of coherent noise sources. The beamformer used in this work is the linearly constrained minimum-variance (LCMV) algorithm [16].

Denoting with $s(t)$ the desired source, with $a_m(t)$ the room impulse response between the m-th microphone and $s(t)$, and with $n_m(t)$ the noise term related to microphone m, the signal acquired by the m-th microphone is given by:

$$z_m(t) = a_m(t) * s(t) + n_m(t). \tag{1}$$

Analyzing the signals with the short-time Fourier transform (STFT), Eq. (1) can be expressed in vector form as:

$$\mathbf{Z}(k,l) = \mathbf{A}(k)S(k,l) + \mathbf{N}(k,l), \tag{2}$$

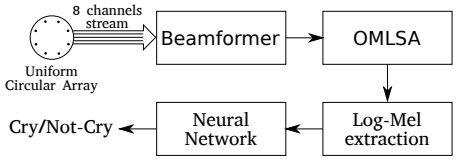


Fig. 1. Block-scheme of the proposed approach.

where l is the frame index and k is the frequency bin index. Beamforming consists in filtering the signal acquired by each microphone with the filter $W_m^*(k,l)$, $m = 1,\dots,M$, and summing the outputs. The vector formulation of the beamforming operation is:

$$Y(k,l) = \mathbf{W}^H(k,l)\mathbf{Z}(k,l). \tag{3}$$

Filters coefficients $\mathbf{W}^H(k,l)$ are obtained by minimizing the output power $E\{Y(k,l)Y^*(k,l)\}$, and constraining the signal component of $Y(k,l)$ to be equal to $S(k,l)$. It can be demonstrated [16] that the steepest descent formulation of the adaptive solution is given by the following expression:

$$\mathbf{W}(k,l+1) = P(k)[\mathbf{W}(k,l) - \mu\mathbf{Z}(k,l)Y^*(k,l)] + \mathbf{F}(k), \tag{4}$$

where $P(k) = \mathbf{I} - \mathbf{A}(k)\mathbf{A}^H(k)/||\mathbf{A}(k)||^2$ and $\mathbf{F}(k) = \mathbf{A}(k)/||\mathbf{A}(k)||^2$.

2) Post-filter: The output of the beamformer is further processed by a post-filter algorithm in order to reduce the residual diffuse noise. The post-filter used in this work is the OMLSA algorithm [16], that operates by applying an adaptive gain function $G(k,l)$ to the output of the beamformer:

$$|\hat{Y}(k,l)|^2 = G(k,l)|Y(k,l)|^2, \tag{5}$$

where

$$G(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{\nu(k,l)}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{6}$$

$$\xi(k,l) = \frac{\sigma_x^2(k,l)}{\sigma_n^2(k,l)}, \quad \gamma(k,l) = \frac{|Y(k,l)|^2}{\sigma_n^2(k,l)}, \tag{7}$$

and $\nu(k,l) = \gamma(k,l)\xi(k,l)/(1+\xi(k,l))$. The noise variance $\sigma_n^2(k,l)$ is estimated using the improved minima controlled recursive averaging (IMCRA) [16]. In OMLSA, the optimal spectral gain function is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty. The modified gain function takes the following form:

$$G(k,l) = [G_{H_1}(k,l)]^{p(k,l)} G_{\min}^{1-p(k,l)}, \tag{8}$$

where $G_{H_1}(k,l)$ is the same as Eq. (6), $p(k,l)$ is the *speech presence probabilty* (SPP) and $G_{\min}$ is a lower threshold [16]. The speech presence probability is computed as

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))e^{-v(k,l)}\right\}^{-1}, \tag{9}$$

where $q(k,l)$ is the *a priori* speech absence probability estimated using a soft-decision approach [16].

B. Feature Extraction

Log-Mel coefficients are widely used acoustic features in audio analysis with Convolutional Neural Networks, since they allow a compact representation of the audio signals while retaining discriminative information [18], [19]. Log-Mels are obtained by dividing the signal in frames 20 ms long and overlapped by 10 ms. After calculating their Fast-Fourier Transform, each frame is filtered with a filter-bank composed

ISBN 978-90-827970-1-5 © EURASIP 2018

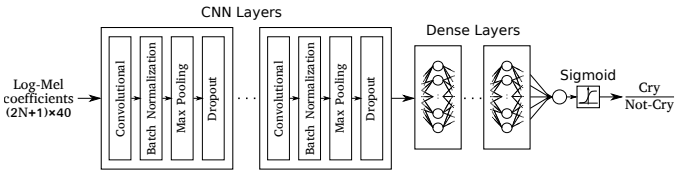


Fig. 2. The general neural network architecture used for cry detection.

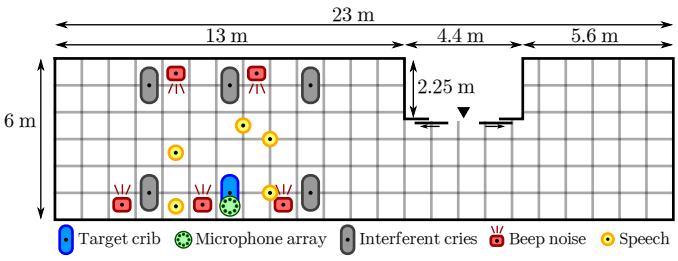


Fig. 3. Plan of the NICU used to create the Simulated Dataset.

of 40 triangular filters equally spaced in the mel-space. Log-Mel coefficients are finally obtained by calculating the energy in each band and applying the logarithm operator. The final feature vector is thus composed of 40 coefficients.

As aforementioned, the classifier does not operate on individual feature vectors, but it exploits the temporal information contained in adjacent frames. The input of the neural network is thus a $(2N+1) \times 40$ matrix, where N is the size of the temporal context, i.e., the number of frames preceding and following the frame being classified. In this paper, N has been set to 49 frames, corresponding to about 0.5 s.

C. Neural Network

The neural network architecture used for cry detection is shown in Fig. 2. The exact topology of the network is determined in the experimental phase by using a validation set (Section IV), however its general structure is defined as follows: the first part of the network consists in one or more convolutional layers, each followed by batch normalization [20], rectifier linear unit (ReLU) activation function, dropout and max-pooling operator. The output of convolutional layers is processed by one or more fully connected layers, each followed by batch normalization, ReLU activation function, and dropout. The output layer is composed of a single neuron with a sigmoid activation function. that outputs the probability of the central frame of being a cry. Training of the network is performed by minimizing the binary cross-entropy loss with the Adam algorithm [20].

The hyperparameters related to the network topology that are determined in the experimental phase are the number of convolutional and fully connected-layers, the size and the number of the kernels of convolutional layers, the size of the max-pooling operator, the dropout rate, and the number of units in the fully-connected layers.

III. Comparative Method

The proposed approach has been compared to a popular algorithm commonly used for voice activity detection [17]. The algorithm will be denoted as "Ramírez" in the following sections. Recalling the notation of Section II-A, the algorithm operates by calculating the long-term spectral estimation (LTSE) from the input signal as:

$$\mathrm{LTSE}(k,l) = \max\{Y(k,l+j)\}_{j=-6}^{j=+6}, \tag{10}$$

and the long-term spectral divergence (LTSD) as:

$$\mathrm{LTSD}(l) = 10\log 10\left(\frac{1}{\mathrm{NFFT}}\sum_{k=0}^{\mathrm{NFFT}-1}\frac{\mathrm{LTSE}^2(k,l)}{N^2(l)}\right), \tag{11}$$

where $N(l)$ is the noise spectrum magnitude and NFFT is the number of FFT points. A frame is classified as cry if the value of the LTSD exceeds a predefined threshold. The algorithm includes a hangover scheme that delays cry/non-cry transitions during 8 frames. Additionally, the noise spectrum magnitude $N(l)$ is updated if a non-cry is detected, thus improving the robustness of the algorithm in noisy conditions.

IV. Experiments

The proposed approach as well as the comparative method have been evaluated on a Simulated dataset and on a Real dataset containing data acquired in a NICU.

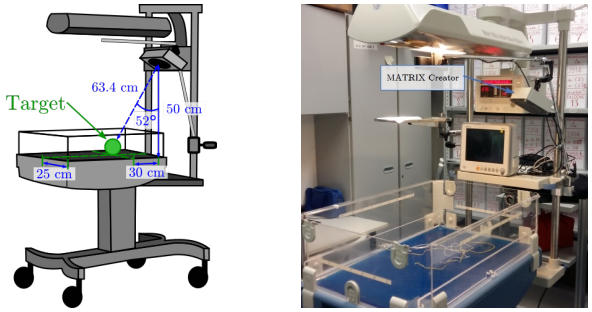
A. Simulated Dataset

The Simulated dataset has been created by considering the scenario in Fig. 3, showing the plan of a NICU. The eight-channel circular microphone array with radius 5.25 cm has been positioned close to the crib shown in blue, and it has been oriented towards the head of the infant as shown in Fig. 4a. The impulse responses between the cry source and the microphones have been created by using Pyroomacoustics[1]. A "clean" set has been created by convolving 64 infant cry recordings at 16 kHz[2] with the synthetic impulse responses. Each recording contains the sound of a single subject, and it has been zero-padded in order to be 30 s long. The cry/silence ratio in each recording is about 50%, while the total duration of cry signals is 16 minutes and 57 seconds, with 15 minutes and 1 second of expiratory phases and 1 minutes and 56 seconds of inspiratory phases. The total number of subjects is 29.

Additionally to the "clean" set, 240 "noisy" conditions have been created by reproducing a realistic acoustic scenario. Noisy conditions have been created by synthetically adding four noises: human speech, infant cry, "beep" sounds, and background noise. Human speech considers the presence of other persons in the room, infant cry the presence of other infants in the cribs nearby the target, and "beep" sounds considers the noise produced by medical equipment. These noises represent coherent sources positioned as shown in Fig. 3, and they have been convolved with the related synthetic impulse responses. As incoherent background noises, the sounds of a fan and of an oxygen concentrator have been used. Noises and clean data have been synthetically combined in order to

[1] https://github.com/LCAV/pyroomacoustics
[2] Sources: www.freesound.org and www.youtube.com

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) (b)

Fig. 4. Acquisition setup used for creating the Real dataset.

produce sets with SNRs equal to 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. For each coherent noise source and SNR, 12 acoustic scenes have created, each differing for the position of the source. Regarding the background noise, 12 acoustic scenes have been created by using a different noise realization.

B. Real Dataset

The Real dataset has been acquired in the NICU of the Salesi Children's Hospital (Ancona, Italy). The plan of the room and the crib disposition are the same of the Simulated dataset shown in Fig. 3. Signals have been acquired at 16 kHz by using the MATRIX Creator development board, equipped with 8 digital MEMS microphones (model MP34DB02 by STMicroelectronics) arranged in a uniform circular array with radius 5.25 cm. The MATRIX Creator has been mounted in a support that allows to direct it towards the desired position. The model of the crib is Babytherm 8004/8010 by Draeger. The dataset is composed of 10 sequences of 30 s of a single subject, for a total duration of the cry signal equal to 122.95 s (234 cry units). The average SNR is 0.09 dB$\pm$0.06 dB.

C. Experimental Procedure

The first phase of the evaluation procedure is finding the values of the hyperparameters that provide the best performance. This phase has been conducted on the clean data of the Simulated dataset, with a 4-fold cross-validation and a random search strategy [20]. As reported in Table I, among the explored parameters we included the number of layers, the number of kernels and their shape, as well as the dropout rates. The parameters of the convolutional layers have been constrained so that the number of kernels, and their shape decreases with depth, while the dropout rate increases.

After this phase, the DNN-based cry detector and the comparative method have been evaluated on the noisy and enhanced signals of the Simulated dataset. The former have been obtained by using one channel of the microphone array, while the latter by processing the noisy sets with the signal enhancement pipeline described in Section II-A.

Additionally, the proposed approach has been evaluated on two training conditions, "clean" and "multicondition". The "clean" training set contains uncorrupted files described in Section IV-A, while the "multicondition" training set contains a combination of files with all the SNRs and all the noises

TABLE I
HYPERPARAMETERS EXPLORED IN THE RANDOM SEARCH. "U": UNIFORM DISTRIBUTION; $\log U$ UNIFORM DISTRIBUTION IN THE LOG-DOMAIN.

Parameter (Distribution)	Range	Final
Batch size (U)	{128, 256, 512}	256
Nr. of CNN layers (U)	$[1, 3]$	3
Nr. of fully-connected layers (U)	$[1, 3]$	1
CNN layers		
Kernel shape (U)	$[1, 10] \times [1, 10]$	10×10, 6×2, 3×2,
Kernel number ($\log U$)	$[8, 64]$	32, 32, 32
Activation function (U)	{ReLu, Elu}	ReLu, ReLu, ReLu
Strides (U)	$\{1, 2, 3\} \times \{1, 2, 3\}$	1×1, 1×1, 1×1
Pooling (U)	{yes, no}	yes, yes, no
Pooling Strides (U)	$\{1, 2, 3\} \times \{1, 2, 3\}$	2×2, 2×2
Pooling Shape (U)	$\{1, 2, 3\} \times \{1, 2, 3\}$	2×1, 2×2
Dropout Rate (U)	$\{0, 0.1, 0.2, 0.3\}$	0.1, 0.2, 0.3
Fully-connected layers		
Units $\log U$	$[128, 1024]$	1024
Activation function:	ReLu, Elu	ReLu
Dropout Rate (U)	$\{0, 0.5\}$	0.5

TABLE II
COMPOSITION OF THE MULTICONDITION TRAINING SET.

SNR	Beep	Speech	Interferent Cry	Background	Total
0 dB	2	4	4	4	14
5 dB	2	4	4	2	12
10 dB	6	4	2	2	14
15 dB	4	2	2	4	12
20 dB	2	2	4	4	12

processed by the signal enhancement pipeline (see Table II for details). The neural network trained on clean data will be indicated with "DNN-Clean", while the one trained on multicondition data with "DNN-Multi". The performance on the Real dataset has been obtained by using the same models trained on the Simulated dataset.

The performance has been evaluated in terms of Average Precision score (AP), a metric that summarizes the Precision and Recall curve. AP score is calculated as $\text{AP} = \sum_n (R_n - R_{n-1}) P_n$ where R_n and P_n are respectively the Recall and Precision for threshold n. Precision and Recall are calculated as $R_n = TP_n/(TP_n + FN_n)$ and $P_n = TP_n/(TP_n + FP_n)$, where TP_n is the number of cry frames correctly detected, FN_n is the number of false negatives, and FP_n is the number of false positives.

D. Results on the Simulated Dataset

The network topology that provided the best performance is composed of three convolutional layers and one fully-connected layer (details are reported in Table I). On clean data, this network achieves an AP score equal to 94.83%.

The results obtained on the Simulated dataset are shown in Table III. Without the signal enhancement stage, the DNN-based cry detector outperforms "Ramírez [17]" by 12.98% despite it has been trained on clean data and the latter includes a noise robust adaptive stage. The effectiveness of the signal enhancement stage is evident by observing the result reported in the lower part of Table III. The effect on the proposed approach can be observed by comparing the DNN-Clean results, where introducing the signal enhancement stage improves the performance by 2.40%. Similarly for the

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE III
RESULTS OBTAINED ON THE SIMULATED DATASET.

Algorithm / SNR	0 dB	5 dB	10 dB	15 dB	20 dB	Average
Without Signal Enhancement						
DNN-Clean	83.35	85.11	86.70	88.45	90.11	**86.74**
Ramírez [17]	71.39	72.03	73.00	74.95	77.46	73.76
With Signal Enhancement						
DNN-Clean	86.86	88.19	89.25	90.29	91.09	89.14
DNN-Multi	91.63	92.58	93.18	93.52	93.71	**92.92**
Ramírez [17]	74.08	75.43	76.41	77.82	78.86	76.52

TABLE IV
RESULTS OBTAINED ON THE REAL DATASET.

	DNN-Clean	DNN-Multi	Ramírez [17]
AP Score (%)	67.41	**86.58**	75.67

comparative method, the AUC improves by 2.76%. The overall best result is obtained by using the multicondition training set (DNN-multi), where the AP score exceeds 90%.

E. Results on the Real Dataset

The results on the Real dataset are shown in Table IV. In this case, all the results have been obtained by using the signal enhancement stage. Training the neural network on clean data provides an AP score equal to 67.51%, while using the multicondition training set improves the performance by 19.17%. Moreover, the DNN-Multi approach outperforms the comparative method by 10.91%. It is worth reminding that the DNN-Multi has been trained on multicondition data of the Simulated dataset, thus demonstrating the good generalization capabilities of the proposed method.

V. CONCLUSION

In this work, a neural network based approach for infant cry detection has been proposed. It makes use of a Convolutional Neural Network having as input Log-Mel features extracted from the audio signals. Log-Mels are extracted from audio pre-processed version of signals acquired in real acoustic environments by using a circular Digital-MEMS microphone array, mounted on the embedded system composed of the commercial MATRIX Creator and Raspberry PI boards. The pre-processing stage is characterized by a LVCM beamformer and a speech enhancement algorithm (OMLSA). The neural network has been trained by solely using a synthetic dataset mimicking the acoustic condition of the hospital room in which the cribs equipped with the audio system mentioned above are located. This makes the proposed approach versatile and easily applicable in different environmental contexts. The Simulated dataset is made of 128 hours of audio, in which the presence of different noise sources at diverse SNRs is simulated. Obtained experimental results show that the proposed algorithm is able to achieve an AP Score equal to 86.58% in the Real dataset, superior with respect to the comparative method [17] by 19.17%, thus allowing to positively conclude about the effectiveness of the proposed approach.

Future works are targeted to enhance the DNN generalization capability by using different feature sets and other regularization techniques. Moreover, suitable adaptation strategies will be implemented to refine the DNN parameters by exploiting specific in-domain audio streams. Finally, non-acoustic cues coming from diverse infant activity monitoring systems (e.g., video) will be fused with the available audio signals to improve the overall classification performance.

REFERENCES

[1] L. LaGasse, A. Neal, and B. Lester, "Assessment of infant cry: Acoustic cry analysis and parental perception," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 11, no. 1, pp. 83–93, 2005.

[2] L. Abou-Abbas, C. Tadj, C. Gargour, and L. Montazeri, "Expiratory and inspiratory cries detection using different signals' decomposition techniques," *Journal of Voice*, vol. 31, no. 2, pp. 259.e13–259.e28, 2017.

[3] G. Naithani, J. Kivinummi, T. Virtanen, O. Tammela, M. J. Peltola, and J. M. Leppänen, "Automatic segmentation of infant cry signals using hidden Markov models," *EURASIP J. Audio, Speech, Music Process.*, vol. 2018, no. 1, p. 1, 2018.

[4] L. Abou-Abbas, H. Fersaie Alaie, and C. Tadj, "Automatic detection of the expiratory and inspiratory phases in newborn cry signals," *Biomed. Signal Process. Control*, vol. 19, pp. 35–43, 2015.

[5] A. Chittora and H. Patil, "Classification of normal and pathological infant cries using bispectrum features," in *Proc. of EUSIPCO*, Nice, France, Aug. 31 - Sept. 4 2015, pp. 639–643.

[6] H. Farsaie Alaie, L. Abou-Abbas, and C. Tadj, "Cry-based infant pathology classification using GMMs," *Speech Communication*, vol. 77, pp. 28–52, 2015.

[7] N. Wahid, P. Saad, and M. Hariharan, "Automatic infant cry pattern classification for a multiclass problem," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 8, no. 9, pp. 45–52, 2016.

[8] V. Mittal, "Discriminating features of infant cry acoustic signal for automated detection of cause of crying," in *Proc. of ISCSLP*, Tianjin, China, Oct. 17-20 2016, pp. 1–5.

[9] S. Ntalampiras, "Audio pattern recognition of baby crying sound events," *Journal of the AES*, vol. 63, no. 5, pp. 358–369, 2015.

[10] S. Orlandi, P. Dejonckere, J. Schoentgen, J. Lebacq, N. Rruqja, and C. Manfredi, "Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 799–810, 2013.

[11] B. Reggiannini, S. Sheinkopf, H. Silverman, X. Li, and B. Lester, "A flexible analysis tool for the quantitative acoustic assessment of infant cry," *J. of Speech Lang. Hear. Res.*, vol. 56, no. 5, pp. 1416–1428, 2013.

[12] R. Cohen and Y. Lavner, "Infant cry analysis and detection," in *Proc. of IEEEI*, Eilat, Israel, Nov. 14-17 2012, pp. 1–5.

[13] E. Principi, D. Droghini, S. Squartini, P. Olivetti, and F. Piazza, "Acoustic cues from the floor: a new approach for fall classification," *Expert Systems with Applications*, vol. 60, pp. 51–61, 2016.

[14] Y. Lavner, R. Cohen, D. Ruinskiy, and H. Ijzerman, "Baby cry detection in domestic environment using deep learning," in *Proc. of ICSEE*, Eilat, Israel, Nov. 16-18 2016, pp. 1–5.

[15] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: A comparison of hand crafted features and deep learning approach," in *Engineering Applications of Neural Networks*, G. Boracchi, L. Iliadis, C. Jayne, and A. Likas, Eds. Cham: Springer, 2017, pp. 168–179.

[16] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York, NY, USA: Springer, 2007, ch. 47.

[17] J. Ramírez, J. Segura, C. Benítez, A. De la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Comm.*, vol. 42, no. 3-4, pp. 271–287, 2004.

[18] P. Vecchiotti, F. Vesperini, E. Principi, S. Squartini, and F. Piazza, "Convolutional Neural Networks with 3-D Kernels for Voice Activity Detection in a Multiroom Environment," in *Multidisciplinary Approaches to Neural Computing*, A. Esposito, M. Faudez-Zanuy, F. C. Morabito, and E. Pasero, Eds. Cham: Springer, 2018, vol. 69, pp. 161–170.

[19] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *Proc. of ICASSP*, Shanghai, China, Mar. 20-25 2016, pp. 2742–2746.

[20] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

ISBN 978-90-827970-1-5 © EURASIP 2018