IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED INFORMATION SENSING AND LEARNING TECHNOLOGIES FOR DATA-CENTRIC SMART HEALTH APPLICATIONS

Received February 13, 2019, accepted March 31, 2019, date of publication April 15, 2019, date of current version April 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2911427

Automatic Detection of Cry Sounds in Neonatal Intensive Care Units by Using Deep Learning and Acoustic Scene Simulation

MARCO SEVERINI, DANIELE FERRETTI, EMANUELE PRINCIPI, AND STEFANO SQUARTINI, (Senior Member, IEEE)
Department of Information Engineering, Università Politecnica delle Marche, 60131 Ancona, Italy

Corresponding author: Emanuele Principi (e.principi@univpm.it)

This work was supported in part by the SINC - System Improvement for Neonatal Care project funded by the Regione Marche, Italy, within the POR MARCHE FESR 2014-2020 - ASSE 1 Program.

ABSTRACT Cry detection is an important facility in both residential and public environments, which can answer to different needs of both private and professional users. In this paper, we investigate the problem of cry detection in professional environments, such as Neonatal Intensive Care Units (NICUs). The aim of our work is to propose a cry detection method based on deep neural networks (DNNs) and also to evaluate whether a properly designed synthetic dataset can replace on-field acquired data for training the DNN-based cry detector. In this way, a massive data collection campaign in NICUs can be avoided, and the cry detector can be easily retargeted to different NICUs. The paper presents different solutions based on single-channel and multi-channel DNNs. The experimental evaluation is conducted on the synthetic dataset created by simulating the acoustic scene of a real NICU, and on a real dataset containing audio acquired on the same NICU. The evaluation revealed that using real data in the training phase allows achieving the overall highest performance, with an Area Under Precision-Recall Curve (PRC-AUC) equal to 87.28 %, when signals are processed with a beamformer and a post-filter and a single-channel DNN is used. The same method, however, reduces the performance to 70.61 % when training is performed on the synthetic dataset. On the contrary, under the same conditions, the new single-channel architecture introduced in this paper achieves the highest performance with a PRC-AUC equal to 80.48 %, thus proving that the acoustic scene simulation strategy can be used to train a cry detection method with positive results.

INDEX TERMS Infant cry detection, deep neural networks, neonatal intensive care unit, data augmentation, acoustic scene simulation, computational audio processing.

I. INTRODUCTION

Newborns' cry signals contain valuable information related to the state of the infant and their acoustic analysis represents a cost-effective and non-intrusive monitoring approach in different environments, from simple households to infant wards or Neonatal Intensive Care Units (NICUs) [1]. Cry detection consists in the identification of a cry signal within an audio stream, and it can serve as a pre-processing stage for deeper analysis, or as a support for the medical staff for evaluating the overall health status of the infant [2]–[5]. Further analysis of the audio signal can detect specific situations, such as the presence of a pathology [6]–[8], or the cause of a cry (e.g., hunger, pain) [9]–[11].

Cry detection has been already addressed in the literature, different techniques have been proposed to solve the problem, and diverse corpora have been used to evaluate them. In some earlier works, the authors [5], [12] proposed an algorithm for detecting voiced sounds by using the short-term energy measure of a signal and an automatic threshold selection algorithm. In [12], the authors assumed that undesired sounds contain lower energy and are shorter compared to voice sounds and evaluated the performance on a synthetic dataset composed of infants' cries. Reggiannini *et al.* [13] adopted a Cepstral-based acoustic analysis to identify cry utterances and detect the fundamental frequency of the cry.

The associate editor coordinating the review of this manuscript and approving it for publication was Qingxue Zhang.

In the recent years, approaches based on machine learning appeared in the literature and demonstrated promising results. Cohen and Lavner [14] suggested an algorithm that uses mel-frequency cepstral coefficients (MFCCs) and is based on k-nearest neighbors to classify both cry and non-cry frames. The algorithm was designed to alert parents when infants are being left alone (either in apartments or vehicles), thus it was evaluated on a synthetic corpus that includes street noises. Hidden Markov models (HMMs) have been used in [4] to detect and classify different elements in the recording including inspiratory and expiratory phases of the cry, as well as beeping sounds, speech, silence and background noise. Experiments have been conducted on a corpus collected in the neonatology departments of multiple hospitals by means of a hand-held recorder placed about 30 cm from the subject. In [2], the same authors extended their work by adopting different signal decomposition techniques and a Gaussian Mixture Model (GMM) classifier in addition to HMMs. Naithani *et al.* [3] also adopted HMMs to classify the expiratory and inspiratory phases of a cry, as well as a third class including all other noises. In this case, audio recordings have been captured in different acoustic environments within the target hospitals. Raboshchuk *et al.* [15] explicitly addressed the robustness of vocalization detection algorithms against noise. The authors proposed a pre-processing pipeline composed of Non-Negative Matrix Factorization (NMF) and spectral subtraction algorithms to reduce undesired disturbances. The paper evaluated a GMM and a Support Vector Machine (SVM) classifier, and the experiments demonstrated the superiority of the SVM-based solution.

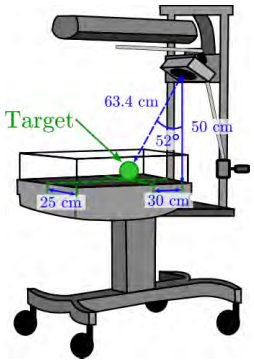
Methods based on deep neural networks (DNNs) have also been proposed for cry detection. In [16], Lavner and colleagues revised their previous work [14] by adopting a neural network composed of three convolutional layers and one fully-connected layer and evaluated the performance by using audio signals recorded in domestic environments. Torres *et al.* [17] presented a solution targeted at low-power devices and proposed a novel set of features. DNNs and support vector data description (SVDD) classifiers were evaluated by using a dataset composed of various recordings collected from public websites. A DNN-based algorithm has been proposed in a preliminary work by the authors [18]. The algorithm uses an eight-channel circular microphone array, and it was evaluated on a synthetic dataset and on a real dataset composed of 10 cry recordings acquired in a NICU.

A deeper analysis on additional data acquired in the NICU revealed significant problems related to the interaction of the medical staff with the microphone array, and important aspects related to the characteristics of the acoustic environment. More in detail, the analysis revealed that the interaction of the medical staff with the infants can misalign the microphone array, thus not allowing the multi-channel solution proposed in [18] to operate properly. The analysis suggested the need to further study the detection approach, and this paper extends the previous work by the authors [18] presenting additional solutions. More in detail, this paper presents a detailed study on single-channel and multi-channel neural networks, and on the beamforming plus post-filtering approach presented in [18]. The feature extraction stage has been modified in order to consider the spectral characteristics of the signals acquired in the NICU. Moreover, we propose the acoustic scene simulation strategy for generating synthetic data for training the neural networks. This solution allows for significant performance without requiring a massive amount of data acquired on-field for training. This represents an important result, since maternity wards and NICUs are very sensitive environments [19], and their access is often restricted to external personnel for multiple reasons, such for avoiding hazards due to the accidental introduction of pathogens. Therefore, the collection of data on-field may be a reason of concern from this perspective. On the other hand, the required safety precautions complicate the procedures and protocols to follow in order to collect data.

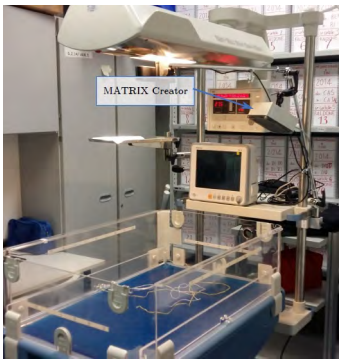
The experimental evaluation has been conducted on a synthetic dataset created by simulating the acoustic scene of a NICU and a real dataset. Compared to [18], the synthetic dataset has been modified in order to be more compliant with the acoustic characteristics of the NICU environment. The real dataset has been significantly extended with respect to the one used in [18], thus allowing a more insightful experimental evaluation of the cry detection algorithms. Moreover, the proposed approaches are compared to a state-of-the-art algorithm for vocalization detection in real NICUs [15]. Up to the authors' knowledge, this is the only work present in the literature that specifically addresses the problem of vocalization detection in real-life scenarios, and that proposes specific solutions to deal with the presence of undesired disturbances. The obtained results show that the proposed DNN-based approaches outperform the comparative method, and that the neural networks architectures introduced in this paper coupled with the acoustic scene simulation strategy are able to achieve significant performance without needing real data for training.

Compared to previous works on DNN-based cry detection [16], [17], this paper addresses the NICU environment, rather than domestic environments, and it proposes specific solutions for reducing the negative effects of undesired disturbances. Differently from [16], [17], here we study single-channel and multi-channel architectures, and we determine the topologies of the networks experimentally rather than fixing them a-priori. Moreover, the acoustic scene simulation strategy is adopted to create a synthetic dataset used in the training phase, and the experimental evaluation is conducted by using data acquired in a real NICU.

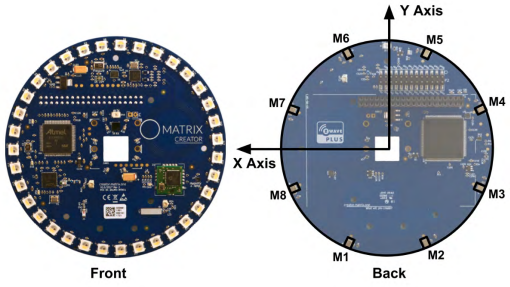
The outline of the paper is the following. Section II presents the case study and the hardware equipment used to collect data. Section III describes the proposed algorithmic framework for cry detection. The comparative method is briefly introduced in Section IV, whereas Section V presents the experiments performed to evaluate the proposed approach, and the obtained results. Finally, Section VI concludes the paper and presents future developments.



(a) Scheme of the recording setup.



(b) Picture of the recording setup.



(c) The MATRIX Creator circular microphone array.

FIGURE 1. Cry detection crib prototype and microphone array details: (a) schematic representation of the position and orientation of the microphone array with respect to the crib, (b) crib prototype adopted in the current research; (c) circular array board microphones placement and orientation, the board radius is 5.25 cm.

II. CASE STUDY

Cry detection technologies can support the NICUs medical staff by providing additional monitoring abilities with a non-intrusive technology. However, NICUs typically present noisy acoustic environments [20], [21] that implicate many challenges for an infant cry detection system. A microphone array allows the use of multi-channel techniques (such as beamforming) to enhance the audio signal by removing coherent noise sources. The prototype at the basis of the present work is shown in Fig. 1, and it consists in a microphone-array directly integrated in the crib and oriented towards the head of the infant. Fig. 1a shows a scheme where the microphone array is evidenced, Fig. 1b shows a picture of the actual prototype.

The crib is a Draeger Babytherm 8004/8010 that has been equipped with a Raspberry Pi 3 Model B v1.2 board, and the MATRIX Creator development board which includes a circular microphone array featuring 8 digital MEMS microphones (model MP34DB02 by ST Microelectronics). The microphone array is distributed uniformly on a circumference with radius of 5.25 cm[1] and is located above the crib by means of a supporting arm. The clamp that binds the array to the arm allows to tilt the array towards the head of the infant. The arm, on the other hand, allows for a partial rotation, to move the array whenever it hinders the activities of the medical staff.

[1] www.matrix.one

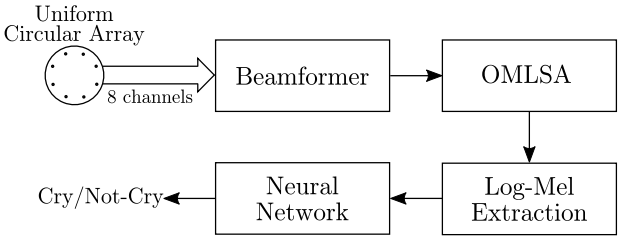


FIGURE 2. Block-scheme of the approach proposed in [18] (SE-DNN).

The monitoring device has been used to record the raw audio samples acquired by the MEMS microphones, without any processing over the audio stream, thus without any noise management, reduction or mitigation, from either the hardware or the software included in the prototype.

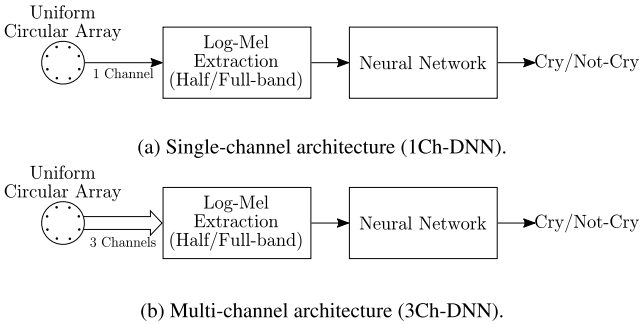
The complexity of the scenario requires a robust and effective cry detection method able to overcome interfering noises such as other infants' cries, the voices of the medical staff, the noise originating from the multiple devices in the NICU.

III. THE PROPOSED APPROACH

The cry detection approach proposed in this paper consists in extracting Log-Mel feature vectors from the audio signal and classifying them as "cry" or "not-cry" by using a DNN. For each feature vector, the network outputs a value comprised in the range [0, 1], where 0 and 1 represent respectively absence and presence of cry. Considering the system described in the previous section, the cry detection approach presented in [18] uses all the audio channels of the circular microphone array depicted in Fig. 1c, and prior to extracting Log-Mel coefficients it reduces undesired disturbances with a filter-and-sum beamformer and the Optimally Modified Log-Spectral Amplitude estimator (OMLSA) post-filter (Fig. 2).

A later investigation of the collected audio signals, however, revealed a few concerns related to the microphone array position and orientation. As an example, during the medical staff activity, the arm supporting the microphone is often moved to the side and it is not restored to its original position for a prolonged period. This problem is representative of the unpredictable nature of a NICU environment. In fact, in our previous work [18], although a part of the collected data has been used to evaluate the cry detection method, the data did not present this problem, thus resulting in a good performance. Further investigation on the collected data showed a performance drop, revealing the problem described above.

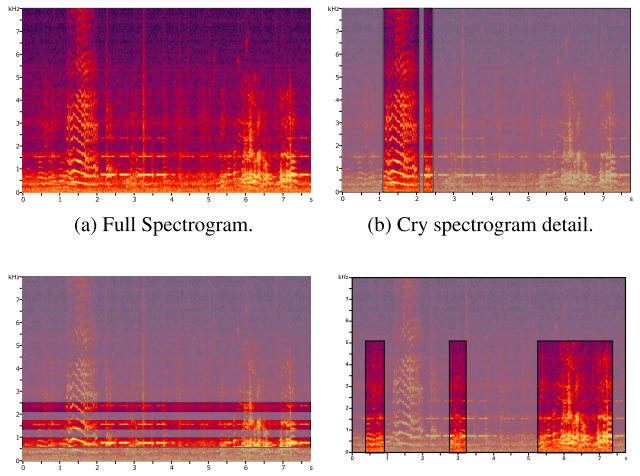
In this regard, some alternatives to the original approach presented in [18] have been investigated, and the related schemes are depicted in Fig. 3. One of the approaches operates on a single-channel, without additional pre-processing (Fig. 3a). The second approach uses 3 channels as input of the DNN, among the 8 provided by the array. In this way, the network directly incorporates the processing stages of multiple audio channels (Fig. 3a).



(a) Single-channel architecture (1Ch-DNN).

(b) Multi-channel architecture (3Ch-DNN).

FIGURE 3. Block-scheme of the single-channel and multi-channel approaches.



(a) Full Spectrogram.

(b) Cry spectrogram detail.

(c) "Beep" noise.

(d) Interfering voice.

FIGURE 4. Excerpt from the audio sample spectrograms and audio source spectra details (the ignored part is masked out): (a) full spectrogram with both cry and noises, (b) identification of the cry target detail, (c) example of "beep" noise, (d) identification of the interfering voice detail.

All the approaches share a similar feature extraction stage, where Log-Mel coefficients are calculated. However, as shown in Fig. 4, the observation of the spectrum of the audio stream in Fig. 4a reveals that the cry signals (Fig. 4b) occupy all the frequency components up to 8 kHz [11], [22], whereas most of the noise types, such as the "beep" noises (Fig. 4c) produced by medical equipment and the interfering voices from the medical staff (Fig. 4d), affect mostly the signal frequency components below 4 kHz. In this regard, an additional filtering stage has been included in the Log-Mel extraction process, where the lower frequency bands are discarded (see Section III-A for the details).

A. FEATURE EXTRACTION

All the methods presented above share the same feature extraction stage. For each input channel of the neural network, the signal is divided in frames 20 ms long and overlapped by 10 ms. The Fast-Fourier Transform of the frame is then filtered with a filter-bank composed of 40 triangular filters equally spaced in the mel-space. Log-Mel coefficients are obtained by calculating the energy in each band, and then by applying the logarithm operator. The final feature vector is composed of 40 elements. This approach, denoted as Full-band in the following, is evaluated both in the architecture of Fig. 2 and in the new architectures of Fig. 3. The Half-band approach, on the other hand, consists in reducing the bandwidth of the filter-bank to the range 4 kHz–8 kHz and number of filters to 20. The length of the feature vector is thus reduced accordingly.

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 $(2N + 1) \times F$ matrix, where N is the size of the temporal context, i.e., the number of frames preceding and following the frame being classified, whereas $F \in \{20, 40\}$ is the length of the feature vector. In this paper, N has been set to 49 frames, that leads to an input corresponding to about 1 s.

B. SINGLE-CHANNEL DNN APPROACH

The single-channel neural network architecture (1Ch-DNN) used for cry detection is shown in Fig. 5. The exact topology of the network is defined by exploring the hyperparameters space on a validation set (see Section V). 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 [23], rectifier linear unit (ReLU) activation function [24], dropout [25], 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. The network training is performed by minimizing the binary cross-entropy loss with the Adam algorithm [26].

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, the number of units in the fully-connected layers, as well as the learning rate, and the batch size used to train and validate the network.

C. MULTI-CHANNEL DNN APPROACH

The multi-channel neural network architecture uses 3 input channels (3Ch-DNN) and is shown in Fig. 6. As in the single-channel approach, the exact topology of the network is determined in the experimental phase by using a validation set (Section V). The general structure, however, is defined as follows: the first part of the network consists of three identical blocks, each with one or more convolutional layers, followed by batch normalization, rectifier linear unit (ReLU) activation function, dropout and max-pooling operator. These three blocks share the same exact topology and operate in parallel. Each channel input corresponds to a specific microphone of the array, that is, the first, the fourth and the seventh.

The outputs of the three blocks are then placed side by side. At one time each block produces one frame, the three frames

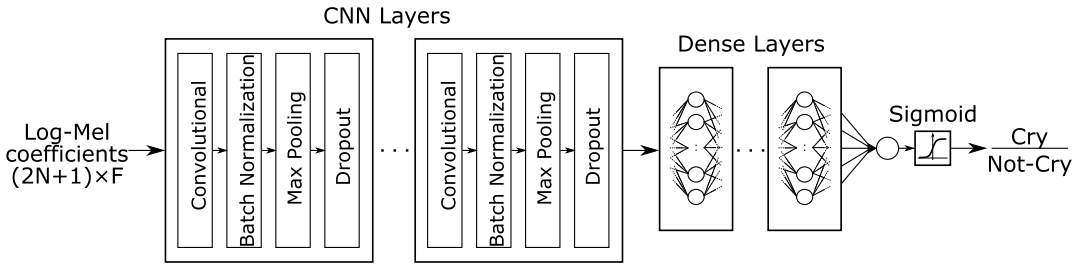


FIGURE 5. Single-channel DNN architecture used for cry detection.

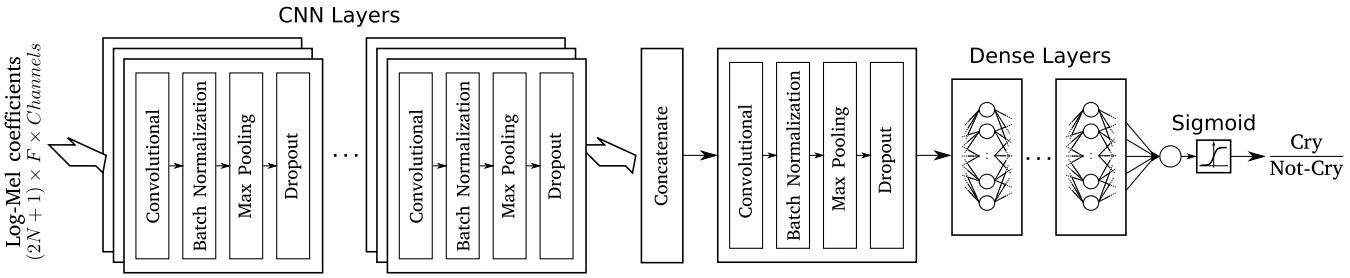


FIGURE 6. Multi-channel DNN architecture used for cry detection.

are then merged in a single frame with the same row number of the original frames and three times the number of columns of the original frames. The resulting frame is then processed by an additional convolutional layer also followed by batch normalization [26], ReLU activation function, dropout and max-pooling operator. The output of this convolutional layers is then processed by one or more fully connected layers, each followed by batch normalization, ReLU activation function, and dropout. As in the single-channel network, the output layer is composed of a single neuron with a sigmoid activation function, and training is performed by minimizing the binary cross-entropy loss with the Adam algorithm [26].

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, the number of units in the fully-connected layers, as well as the learning rate, and the batch size used to train and validate the network.

D. SIGNAL ENHANCEMENT APPROACH

The Signal Enhancement approach (SE-DNN) has been investigated in our previous work [18]. In the present work it is further investigated by means of an extended real dataset. In this approach, all the eight audio channels are used in the algorithm and processed according to the block-scheme shown in Fig. 2. The first stages of the algorithm consist in a filter-and-sum adaptive beamformer followed by a OMLSA post-filter to enhance the signal quality and reduce the noise. The output of the beamformer and of the post-filter is a single channel that follows the same processing steps described in Section III-A and Section III-B. The general structure of the neural network, thus, matches the one that has been already presented in Section III-B, but the exact topology has been determined separately in the experimental phase.

1) BEAMFORMER

A beamformer reduces the effects of coherent noise sources, such as the sounds from the medical equipments present in a NICU. The algorithm used in [18] is the linearly constrained minimum-variance (LCMV) beamformer [27], and it will be now briefly presented. Denoting with $s(t)$ the desired source as a function of time t, 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), (1) can be expressed in vector form as:

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

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 $W_m^*(k,l), m = 1,\ldots,M$, 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 [27] 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 residual diffuse noise is reduced by means of the OMLSA algorithm [27], that applies 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

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

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

and $v(k, l) = \gamma(k, l)\xi(k, l)/(1 + \xi(k, l))$. The noise variance $\sigma_n^2(k, l)$ is estimated using the improved minima controlled recursive averaging (IMCRA) [27]. 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 (6), $p(k, l)$ is the *speech presence probability* (SPP) and $G_{\min}$ is a lower threshold [27]. 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 [27].

IV. COMPARATIVE METHOD

Up to the authors' knowledge, the only work present in the literature that explicitly addresses the task of vocalization detection in real NICUs is the work by Raboshchuk *et al.* in [15]. This work describes a complete pipeline that handles the acoustic data recorded in NICUs and it has been compared to the proposed approaches. Other recent methods for infant cry detection have not been considered in the evaluation, since they lack important implementation details.

The first stage of the comparative method enhances the input signals by applying the Non-negative Matrix Factorization (NMF) and the spectral Subtraction (SS) algorithms. Vocalization detection is then performed by a Gaussian Mixture Model (GMM) based detector.

The NMF algorithm is used to reduce non-stationary noises. Denoting with $V_{F\times N}$ the matrix representing the spectrograms of the input signals, where F are the frequency bins and N the number of frames, it is possible to approximate it with two non-negative matrices:

$$V_{F\times N} \approx W_{F\times R} \cdot H_{R\times N}. \tag{10}$$

The columns of W should be intended as bases, whereas the rows of H as their corresponding activations in each frame, thus $R \leq F$. NMF attempts to find the matrices W and H through the solution of the minimization problem:

$$\text{argmin}_{W,H} D(V||WH) + \lambda |H|_1, \quad W, H \geq 0, \tag{11}$$

where D is Kullback-Leibler divergence and $\lambda \leq 0$ is used to promote a sparsity constraint on the activations. For each source, the matrix of bases is estimated on a training dataset and then is used in the source separation step of the whole dataset. In the cry detection problem, the ensemble matrix of bases $W_t = [W_{Cry}; W_{No-Cry}]$ is kept fixed in (11) in order to estimate the matrix $H = [H_{Cry}; H_{No-Cry}]$ for the test dataset. The spectrum of each source can be obtained as

$$\widehat{V}_i = \frac{W_i H_i}{\sum_i W_i H_i} \otimes V, \quad i \in \left[\text{Cry, No-Cry}\right] \tag{12}$$

where multiplication $\otimes$ and division operations are element-wise. The output enhanced signal is obtained joining the spectrogram $\widehat{V}_{Cry}$ and the phase of the original input audio.

The SS algorithm is applied to reduce the stationary noise contributions. The clean signal spectrum $\widehat{X}(n, k)$ can be estimated from the noisy input spectrum $Y(n, k)$ by subtracting an estimate of the noise spectrum $D(n, k)$:

$$|\widehat{X}(n, k)|^\gamma = \begin{cases} |\widehat{Y}(n, k)|^\gamma - \alpha|\widehat{D}(n, k)|^\gamma, \\ \quad \text{if } |\widehat{Y}(n, k)|^\gamma > (\alpha + \beta)|\widehat{D}(n, k)|^\gamma \\ \beta|\widehat{D}(n, k)|^\gamma, \qquad \text{otherwise} \end{cases} \tag{13}$$

where n is the frame index, k the frequency bin, $\gamma = 2$ corresponds to perform a power spectrum subtraction, α is the subtraction factor and $0 < \beta \ll 1$ is the spectral floor parameter. The noise estimate is obtained by using the Minima-Controlled Recursive-Averaging (MCRA) algorithm [28]:

$$|\widehat{D}(n, k)|^\gamma = \alpha_d(n, k)|\widehat{D}(n-1, k)|^\gamma + (1 - \alpha_d(n, k))|\widehat{Y}(n, k)|^\gamma, \tag{14}$$

with $\alpha_d(n, k) = \alpha + (1 - \alpha)p(n, k)$, where $p(n, k)$ is the speech-presence probability calculated exploiting the ratio between the noisy signal spectrum and its local minimum. The ratio is first smoothed by a factor α_s and then compared to a certain threshold value, where a higher ratio indicates presence of speech. Subsequently, a recursive temporal averaging is carried out, to reduce fluctuations between speech and non-speech segments.

A feature vector composed by 16 Frequency-Filtered Logarithmic FilterBank Energy (FF-LFBE) coefficients, along

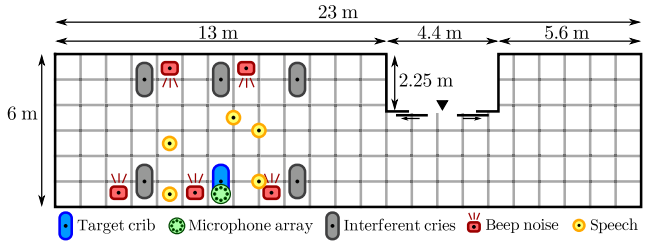


FIGURE 7. Plan of the NICU used to create the Synthetic Dataset.

with their 16 first temporal derivatives, is extracted from the enhanced audio signals divided into frames using 30 ms long Hamming windows, overlapped by 10 ms.

Vocalization detection is performed by a single Gaussian probability density function with a diagonal covariance matrix used to model each class, i.e., Cry and Non-Cry.

V. EXPERIMENTS

The methods described in Section III have been implemented in the Python programming language by using the Keras framework with Tensorflow as the back-end and *librosa* [29] for feature extraction. On the other hand, the source code of the comparative method has been provided us by the authors, and it is based on Matlab and HTK toolkit [30].

By combining the three methods described in Section III and the two feature extraction procedures, we investigated five cry detection strategies, namely:

- Full-band - 1Ch DNN: single-channel DNN with full feature vector as input.
- Full-band - 3Ch DNN: multi-channel DNN with full feature vector as input.
- Half-band - 1Ch DNN: single-channel DNN with half feature vector as input.
- Half-band - 3Ch DNN: multi-channel DNN with half feature vector as input.
- SE-DNN: single-channel DNN with signal enhancement and full feature vector as input.

In the following, we will describe the adopted datasets, the experimental setup and discuss the experimental results.

A. DATASETS

The methods presented above have been evaluated by using a synthetic and a real dataset. The first one is a revision of the synthetic dataset presented in [18], whereas the real dataset has been collected in the NICU of the Salesi Hospital (Ancona, Italy) by means of the prototype described in Section II.

1) SYNTHETIC DATASET

The synthetic dataset simulates the acoustic environment of the Salesi Hospital NICU (*acoustic scene simulation*). This has been performed by using Pyroomacoustics [31], which allows to create the impulse responses between an audio source and a microphone from their positions and the room characteristics.

As shown in Fig. 7, a simulated microphone array and a simulated crib (show in blue) have been placed in the model along with different types of noise sources. The microphone array has been based on the geometry of its real counterpart, with 8 channels placed on a circular pattern with a radius of 5.25 cm. Its position and orientation towards the audio source within the crib (target source) also matches the prototype geometry (i.e., towards the head of the monitored baby). In the model, the noise sources are placed within a radius of about 5.5 m from the microphone array. Three coherent noise source types and two incoherent noise source types have been considered. Among the coherent noise source types are:

- human speech: it emulates the presence of the medical staff;
- infant cry: it emulates other infants within other cribs nearby the target;
- "beep" sound: it emulates the typical noises of a medical equipment.

Regarding the incoherent noise source types, the sounds of a fan and of an oxygen concentrator have been used. The sampling frequency is 16 kHz for all audio data. A total of 64 infant cry recordings belonging to 29 different subjects have been combined with 12 background realization, 23 beep sounds and 26 human speech recordings in order to create 64 audio sequences of 30 s that simulate realistic acoustic scenarios. Half of simulated scenarios presents a SNR of 0 dB, whereas the other half presents a SNR of 5 dB. The total amount of cry signal is 15 minutes and 1 second, whereas the cry/silence ratio in each recording is about 50%.

The speech signals are extracted from a widely used mono clean speech dataset with American English sentences (WSJ0) [32]. All the other audio signals are collected from different web sources.[2,3]

With respect to the synthetic dataset proposed in our previous work [18], SNR values above 5 dB have not been used.

2) REAL DATASET

The real dataset is composed of about 2 hours and 57 minutes of audio data sampled at 16 kHz. The total duration of cry signals is 45 minutes and 55 seconds. As shown in Table 1, a total of 2 female and 3 male infants have been monitored. All the infants were born premature with gestational age between 28 weeks and 34 weeks and 2 days, whereas their age span from 2 days up to 208 days. All of them were suffering or had suffered from some illness, including respiratory, which are very common in preterm children. The dataset is composed of 535 audio fragments whose durations are comprised between 2 and 150 seconds.

B. EXPERIMENTAL SETUP

To evaluate the proposed approaches and to define the topology of each DNN, we used a random search approach [33], defining a pool of 300 configurations. The hyperparameters

[2]http://www.freesound.org
[3]http://www.youtube.com

TABLE 1. Real dataset composition by subjects.

Subject	Sex	Gestational Age (weeks$^{+\text{days}}$)	Age (days)	Nr. of Audio Fragments	Recording Time (s)	Cry Time (s)
Baby 1	M	28	24	113	2280	930
Baby 2	M	34	2	259	4892	897
Baby 3	M	28^{+1}	208	108	2014	497
Baby 4	F	31^{+1}	34	24	606	62
Baby 5	F	34^{+2}	5	31	833	369

TABLE 2. Hyperparameters explored in the random search and network architectures for the proposed configurations. "*U*": Uniform distribution; log *U* uniform distribution in the log-domain.

Parameter (Distribution)	Range	Full-band 1Ch-DNN	Full-band 3Ch-DNN	Half-band 1Ch-DNN	Half-band 3Ch-DNN	SE-DNN
Batch size (U)	{512, 1024, 2048}	512	2048	512	1024	1024
Learning Rate ($\log U$)	$[4.88 \cdot 10^{-4}, 5.52 \cdot 10^{-3}]$	$1.02 \cdot 10^{-3}$	$8.54 \cdot 10^{-4}$	$2.18 \cdot 10^{-3}$	$5.55 \cdot 10^{-4}$	$2.89 \cdot 10^{-3}$
CNN layers						
Nr. of CNN layers (U)	$[1, 3]$	1	1	3	2	3
Kernel shape (U)	$[1, 10] \times [1, 10]$	2×1	1×1	1×1, 1×1, 1×1	2×2, 1×1	5×3, 2×2, 2×1
Kernel number ($\log U$)	$[16, 64]$	18	36	63, 18, 19	4×1, 2×1	29, 54, 61
Strides ($\log U$)	$[1, 6] \times [1, 6]$	3×1	2×1	2×4, 5×1, 4×1	27, 32	2×2, 2×5, 3×1
Pooling Shape (U)	$\{1, 2\} \times \{1, 2\}$	1×2	1×2	2×1, 2×2, 2×1	1×1, 1×1	1×1, 2×1, 1×2
Pooling Strides (U)	$\{1, 2\} \times \{1, 2\}$	1×1	1×2	1×2, 1×2, 1×2	1×1, 2×2	1×1, 1×1, 1×2
Dropout Rate (U)	$\{0, 0.1\}$	0.1	0	0.1, 0.2, 0.3	0.0, 0.1	0.1, 0.2, 0.3
Last CNN Layer (Multi-Channel DNN Only)						
Nr. of CNN layers (U)	1	-	1	-	1	-
Kernel shape (U)	$[1, 4] \times [1, 4]$	-	1×1	-	2×2	-
Kernel number ($\log U$)	$[16, 64]$	-	35	-	53	-
Strides (U)	$[1, 7] \times [1, 7]$	-	2×2	-	2×1	-
Fully-connected layers						
Nr. of fully-connected layers (U)	$[1, 3]$	1	3	3	1	2
Units $\log U$	$[100, 1024]$	181	251, 153, 127	154, 113, 107	796	148, 140
Dropout Rate (U)	$\{0, 0.5\}$	0	0.5, 0.5, 0.5	0.5, 0.5, 0.5	0	0.5, 0.5
Number of trainable parameters	-	4.193.535	4.455.596	49.570	4.117.515	305.996

distributions and ranges reported respectively in the first and second column of Table 2. The synthetic dataset has been divided in 4 folds with the same number of audio sequences, corresponding to the 25% of the dataset each. On the other hand, the real dataset has been divided in three parts. One third has been used as test set, whereas the remaining part has been further divided in training set (75%) and validation sets (25%), corresponding to 50% and 16.7% of the whole real dataset respectively. Each subject is present only in the training set, the validation set, or the test set.

The experimentation consists of three main phases:

- Hyperparameters search conducted on the synthetic dataset: 3 folds have been used as a training set and 1 fold has been used as a validation set. The process has been carried out for each method and each configuration. The best performing topology of each method is reported in Table 2 and the results are discussed in Section V-C.1.
- Training conducted on the real dataset: in this experiment, the network architectures determined in the previous phase and reported in Table 2 have been trained and evaluated on the real dataset. The validation set has been used for early stopping, and the evaluation has been conducted on test set of the real dataset. The obtained results are discussed in Section V-C.2.
- Training conducted on the synthetic dataset, followed by a test on the real dataset: the entire synthetic dataset has been used to train the network architectures reported in Table 2. The evaluation has been conducted on the overall real dataset and on the test set of the real dataset. This experiment evaluates the effectiveness of the acoustic scene simulation strategy. The obtained results are discussed in Section V-C.3.

Regarding the comparative method, we used the hyperparameters values reported in [15], and training is performed on the same training set of the proposed methods.

The performance has been evaluated by using the Area Under the Precision-Recall Curve (PR-AUC) [34]. The PR-AUC is calculated as follows:

$$\text{PR-AUC} = \sum_{n} (R_n - R_{n-1}) \cdot P_n, \tag{15}$$

where R_n and P_n are respectively the Recall and Precision for the threshold n. Precision and Recall are calculated from the true positives TP, the true negatives TN, the false positives FP, and the false negatives FN as follows:

$$R = \frac{TP}{TP + FN}, \quad P = \frac{TP}{TP + FP}, \tag{16}$$

where the subscript n has been omitted for simplicity.

C. RESULTS

The results of the experiments described above are reported in Tables 3, 4 and 5. More in detail, Table 3 summarizes

TABLE 3. PR-AUC (%) on synthetic validation dataset (training on synthetic dataset).

Algorithm	Validation
Full-band 1Ch-DNN	85.31
Full-band 3Ch-DNN	90.54
Half-band 1Ch-DNN	82.97
Half-band 3Ch-DNN	89.12
SE-DNN	**90.55**
Raboshchuk *et al.* [15]	76.37

TABLE 4. PR-AUC (%) on the test set of the real dataset (training on real dataset).

Algorithm	Validation	Test
Full-band 1Ch-DNN	**97.50**	86.18
Full-band 3Ch-DNN	97.00	81.72
Half-band 1Ch-DNN	91.63	84.53
Half-band 3Ch-DNN	96.29	81.28
SE-DNN	97.09	**87.28**
Raboshchuk *et al.* [15]	92.89	74.96

the results obtained during the hyperparameters search process by using the synthetic dataset. Table 4 summarizes the results that the DNNs trained over the real dataset achieve during the validation and the test conducted on the real dataset. In Table 5, training has been performed on the whole synthetic dataset, whereas the testing has been carried out, respectively, on the overall real dataset and on the test set of the real dataset.

From a general standpoint, we observe that the use of a synthetic dataset for training can produce good cry-detection results, up to 83.25% over the overall real dataset and up to 80.48% over the test set of the real dataset (Table 5). This result proves that training the DNNs over a synthetic dataset can represent a viable alternative to real life counterparts. This performance is even more notable if we consider that the synthetic dataset does not model the problem encountered with the real dataset, i.e., the misalignment of the microphone array with respect to its reference position.

1) HYPERPARAMETERS SEARCH

The results obtained in the hyperparameters search phase are reported in Table 3, and the best performing methods are the SE-DNN approach, and the multi-channel approach applied to the full-band feature set, which achieve almost identical results. The half-band multi-channel approach achieves a PR-AUC equal to 89.12%, whereas the worst performing method is the half-band single-channel approach with a PR-AUC equal to 82.97%. The comparative method [15] achieves a PR-AUC equal 76.37%, resulting the less performing approach.

2) REAL DATASET TRAINING

The results obtained by training the algorithms on the real dataset are reported in Table 4. On the validation set, the Full-band 1Ch-DNN is the best performing algorithm with a PR-AUC equal to 97.50%. The Full-band 3Ch-DNN and the SE-DNN attain almost identical results with a PR-AUC equal to 97.00% and 97.09% respectively. The Half-band 3Ch-DNN achieves slightly lower results with a 96.29% detection rate, whereas the half-band single-channel network is the worst performer with a 91.63% score. The approach proposed by Raboshchuk *et al.* [15] scores 92.89%, thus superior to the Half-band 1Ch-DNN.

The second column of Table 4 shows the results obtained on the test set. Comparing these results to the ones obtained on the validation set, the performance drop is evident. Multi-channel networks exhibit the highest reduction, with a PRC-AUC below 82%, whereas the Full-band 1Ch-DNN is more robust, with a PRC-AUC equal to 86.18%. The SE-DNN is the best performer with an 87.28% detection rate and a drop of roughly 10 percentage points (pp). The Half-band 1Ch-DNN, although it does not achieve the highest PR-AUC, shows the lowest performance drop and a PRC-AUC 2.75 pp lower than the most performing network. The approach proposed by Raboshchuk *et al.* [15] is the least robust of the evaluated methods, with PR-AUC reduction of about 18 pp.

TABLE 5. PR-AUC (%) on real dataset (training on synthetic dataset - test on the overall real dataset and on the test set of the real dataset).

Algorithm	Overall	Test
Full-band 1Ch-DNN	80.80	76.77
Full-band 3Ch-DNN	77.95	72.71
Half-band 1Ch-DNN	**83.25**	**80.48**
Half-band 3Ch-DNN	69.15	59.05
SE-DNN	74.06	70.61
Raboshchuk *et al.* [15]	54.12	46.18

3) SYNTHETIC DATASET TRAINING

In this experiment, training is performed on the entire synthetic dataset and testing on the real dataset. Table 5 reports both the results obtained on the overall real dataset and on the test subset in order to compare them with results obtained in Section V-C.2.

On the overall real dataset, the best performing algorithm is the Half-band 1Ch-DNN, with a PR-AUC equal to 83.25%. Notably, the PR-AUC is 0.28 pp greater than the one obtained on the synthetic dataset during the hyperparameters search phase (Table 3). The other investigated methods, on the other hand, show a performance drop. The Full-band 1Ch-DNN achieves a PR-AUC equal to 80.80%, with a reduction of 4.51 pp. The SE-DNN approach reduces the PR-AUC to 74.06% and the multi-channel networks to 77.95% in the full-band case, and to 69.15% in the half-band case.

A motivation for this behavior is that the SE-DNN method and the multi-channel networks are affected by the deviation of the microphone array from its target position, whereas single-channel networks are not. Rather, the main difference between single-channel networks is that the Full-band 1Ch-DNN shows a performance reduction, whereas the Half-band 1Ch-DNN shows an improvement. This suggests that in this case the additional filtering step performed in the feature extraction stage improves the detection abilities.

TABLE 6. PR-AUC (%) on the test set of the real dataset when training is performed on the real dataset and on the synthetic dataset.

Algorithm	Training set		Difference
	Real	Synthetic	
Full-band 1Ch-DNN	86.18	76.77	-9.41
Full-band 3Ch-DNN	81.72	72.71	-9.01
Half-band 1Ch-DNN	84.53	**80.48**	-4.05
Half-band 3Ch-DNN	81.28	59.05	-22.23
SE-DNN	**87.28**	70.61	-16.67
Raboshchuk *et al.* [15]	74.96	46.18	-28.78

The third column of Table 5 reports the performance of the different strategies limited to the test set of the real dataset. Although a further performance reduction affects all the evaluated strategies, the overall situation remains unchanged: the Half-band 1Ch-DNN approach is subject to a performance reduction, but it is still the best performing method and it exhibits the lowest performance drop among the different approaches.

The comparative method [15] exhibits a significant performance reduction: the PR-AUC on the overall real dataset is equal to 54.13%, and it further reduces when the experiment is carried out over the test set.

4) DISCUSSION

The third column of Table 4 and Table 5 reports the results obtained on the test set of the real dataset. The same results are reported in Table 6 to ease the comparison. Observing Table 6, it is evident that training the algorithms over the synthetic dataset causes a general performance reduction. This is a common behavior that affects machine learning algorithms when training and testing is conducted on mismatched conditions. In the case study of this paper, the performance reduction is particularly evident for certain approaches due the misalignment between the microphone array and the target position that affects the real dataset.

Indeed, the performance reduction highly depends on the algorithm, with the Half-band 3Ch-DNN being the most affected method (22.23 pp), and the Half-band 1Ch-DNN being the least affected one (4.05 pp). The SE-DNN approach achieves the highest PR-AUC when training is performed on the real dataset, but it is also affected by a significant performance reduction (16.67 pp). Observing Table 2, it is also evident that the SE-DNN and the Half-band 1Ch-DNN are the networks with lowest number of trainable parameters. More in detail, the Half-band 1Ch-DNN has one hundredth the number of parameters of the Full-band 1Ch-DNN. Reducing the size of the feature vector has the effect of reducing also the size of the network, thus to improve its generalization capabilities. Similarly, the SE-DNN has less than one tenth the number of parameters of the multi-channel networks, and better generalization capabilities, confirming the benefits of introducing a pre-processing stage.

The notable performance exhibited by SE-DNN method may give the impression that acquiring cry signals on-field is worth the effort. NICUs and maternity wards, however, are very sensitive environments, and it can be difficult if not impossible to properly record a significant amount of data for training a learning algorithm. Moreover, on-field acquisition has intrinsic constraints on the number of subjects that can be monitored, which can affect the diversity of cry samples. The same conclusion holds also if the hardware requirements are considered. From this perspective, the Half-band 1Ch-DNN may appear sub-optimal, but it represents be the most cost effective and non-intrusive solution when coupled with the acoustic data simulation strategy. It should be also noted that, by taking into account the crib structure, the microphone array deviation problems may be simulated, thus better simulating the conditions of the real dataset.

Similarly to the proposed DNN-based algorithms, the comparative method exhibits a significant performance reduction when training is performed on the synthetic dataset and testing on the real dataset. Differently from the DNN-based solutions, however, the PR-AUC reduces to 46.18%, suggesting that a synthetic dataset cannot be used as an alternative to a real dataset.

VI. CONCLUSION

In this paper, DNN-based methods for infant cry detection have been proposed and the effectiveness of the acoustic scene simulation strategy has been investigated. This strategy has been proposed with the aim of avoiding the dependence of the training algorithms on real data, and for reducing the intrusiveness level required for collecting data in NICUs. Moreover, algorithms can be more easily retargeted to other NICUs without the need for additional acquisitions.

The analysis of real-life dataset highlighted a few problems, among which the deviation of the microphone array at the basis of the monitoring device from its reference position. From this, we proposed additional solutions based on single-channel and multi-channel networks in order to better evaluate the robustness and performance with respect to the approach presented in [18].

The experiments have been conducted on a synthetic dataset created by simulating the acoustic scene, and the real dataset containing data acquired in the NICU. The proposed methods have been compared with a state-of-the-art algorithm for vocalization detection NICUs [15]. The evaluation revealed that the SE-DNN method is the best performing approach when training and evaluation are performed on the same dataset, but that it shows a significant performance reduction if trained on the synthetic dataset and evaluated on the real dataset. At the same time, under the same conditions the Half-band 1Ch-DNN approach is the best performing solution with a PR-AUC equal to 80.48%. This allows us to confirm about the effectiveness of the acoustic scene simulation strategy.

The achieved results proved that a synthetic dataset can be a useful replacement with respect to a real-life dataset. Indeed, it allows to reduce the interaction with a sensitive environment such as a NICU, to the bare minimum. Moreover, it can be adjusted to include changes to the environment as needed,

without requiring an additional acquisition session. In this regard, one of our future aims is to revise the simulation dataset, to include some occurrences where the microphone array does not target the intended subject, to bridge the still existing gap with the DNN training by using real data. Further studies will be addressed to cry classification, aimed at the recognition of the cause of the cry.

ACKNOWLEDGMENT

The authors would like to thank the authors of paper [15] for sharing the source code of their algorithm. They also want to acknowledge the Italian University and Research Consortium CINECA for the availability of high-performance computing resources and support.

REFERENCES

[1] L. LaGasse, A. R. Neal, and B. M. Lester, "Assessment of infant cry: Acoustic cry analysis and parental perception," *Mental Retardation Develop. Disabilities Res. Rev.*, vol. 11, no. 1, pp. 83–93, Feb. 2005.
[2] L. Abou-Abbas, C. Tadj, C. Gargour, and L. Montazeri, "Expiratory and inspiratory cries detection using different signals' decomposition techniques," *J. Voice*, vol. 31, no. 2, pp. 259.e13–259.e28, Mar. 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, Dec. 2018.
[4] L. Abou-Abbas, H. F. 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, May 2015.
[5] M. A. R. Díaz, C. A. R. García, L. C. A. Robles, J. E. X. Altamirano, and A. V. Mendoza, "Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis," *Biomed. Signal Process. Control*, vol. 7, no. 1, pp. 43–49, Jan. 2012.
[6] A. Chittora and H. A. Patil, "Classification of normal and pathological infant cries using bispectrum features," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Nice, France, Aug./Sep. 2015, pp. 639–643.
[7] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García, "Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies," in *Proc. 7th Mexican Int. Conf. Artif. Intell.*, Atizapan de Zaragoza, Mexico, Oct. 2008, pp. 330–335.
[8] Z. Benyó, Z. Farkas, A. Illényi, G. Katona, and G. Várallyay, Jr., "Information transfer of sound signals. A case study: the infant cry. Is it a noise or an information?" in *Proc. Int. Congr. Expo. Noise Control Eng.*, Aug. 2004, no. 5, pp. 2774–2781.
[9] V. K. Mittal, "Discriminating features of infant cry acoustic signal for automated detection of cause of crying," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Tianjin, China, Oct. 2016, pp. 1–5.
[10] S. Ntalampiras, "Audio pattern recognition of baby crying sound events," *J. Audio Eng. Soc.*, vol. 63, no. 5, pp. 358–369, Jun. 2015.
[11] A. Chittora and H. A. Patil, "Newborn infant's cry analysis," *Int. J. Speech Technol.*, vol. 19, no. 4, pp. 919–928, Dec. 2016.
[12] S. Orlandi, P. H. Dejonckere, J. Schoentgen, J. Lebacq, N. Rruqja, and C. Manfredi, "Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 799–810, Nov. 2013.
[13] B. Reggiannini, S. J. Sheinkopf, H. F. Silverman, X. Li, and B. M. Lester, "A flexible analysis tool for the quantitative acoustic assessment of infant cry," *J. Speech Lang. Hearing Res.*, vol. 56, no. 5, pp. 1416–1428, Oct. 2013.
[14] R. Cohen and Y. Lavner, "Infant cry analysis and detection," in *Proc. IEEE 27th Conv. Elect. Electron. Eng. Isr.*, Eilat, Israel, Nov. 2012, pp. 1–5.
[15] G. Raboshchuk, C. Nadeu, S. V. Pinto, O. R. Fornells, B. M. Mahamud, and A. R. de Veciana, "Pre-processing techniques for improved detection of vocalization sounds in a neonatal intensive care unit," *Biomed. Signal Process. Control*, vol. 39, pp. 390–395, Jan. 2018.
[16] Y. Lavner, R. Cohen, D. Ruinskiy, and H. Ijzerman, "Baby cry detection in domestic environment using deep learning," in *Proc. IEEE Int. Conf. Sci. Elect. Eng. (ICSEE)*, Eilat, Israel, Nov. 2016, pp. 1–5.
[17] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: A comparison of hand crafted features and deep learning approach," in *Proc. Int. Conf. Eng. Appl. Neural Netw. (EANN)*, Athens, Greece, Aug. 2017, pp. 168–179.
[18] D. Ferretti, M. Severini, E. Principi, A. Cenci, and S. Squartini, "Infant cry detection in adverse acoustic environments by using deep neural networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 997–1001.
[19] R. A. Polin, S. Denson, and M. T. Brady, "Strategies for prevention of health care-associated infections in the NICU," *Pediatrics*, vol. 129, no. 4, pp. e1085–e1093, 2012.
[20] H. Shoemark, E. Harcourt, S. J. Arnup, and R. W. Hunt, "Characterising the ambient sound environment for infants in intensive care wards," *J. Paediatrics Child Health*, vol. 52, no. 4, pp. 436–440, Apr. 2016.
[21] S. M. Hassanein, N. M. E. Raggal, and A. A. Shalaby, "Neonatal nursery noise: Practice-based learning and improvement," *J. Maternal-Fetal Neonatal Med.*, vol. 26, no. 4, pp. 392–395, 2013.
[22] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni, "Developmental aspects of infant's cry melody and formants," *Med. Eng. Phys.*, vol. 24, nos. 7–8, pp. 501–514, Sep./Oct. 2002.
[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lile, France, Jul. 2015, pp. 448–456.
[24] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, Jun. 2010, pp. 807–814.
[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.
[27] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Berlin, Germany: Springer, 2008, ch. 47, pp. 945–978.
[28] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[29] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proc. 14th Python Sci. Conf.*, Austin, TX, USA, Jul. 2015, pp. 18–25.
[30] S. J. Young *et al.*, *The HTK Book*. Cambridge, U.K.: Cambridge Univ. Press, 2002.
[31] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 351–355.
[32] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British english corpus for large vocabulary continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Detroit, MI, USA, May 1994, pp. 81–84.
[33] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
[34] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds. Berlin, Germany: Springer, 2013, pp. 451–466.

MARCO SEVERINI received the degree in electronics engineering from the Università Politecnica delle Marche, Italy, in 2012, where he is currently a Research Fellow with the Department of Information Engineering. His current research interests include the design and development of task and resource scheduling algorithms, energy and power management optimization, mixed integer nonlinear programming, wireless sensor networks, embedded system programming, and smart grids.

DANIELE FERRETTI was born in Ancona, Italy, in 1985. He received the M.Sc. and Ph.D. degrees from the Università Politecnica delle Marche, Italy, in 2015 and 2019. He is currently a freelance Software Engineer, working at the development of efficient machine learning algorithms for embedded platforms.

EMANUELE PRINCIPI was born in Senigallia, Italy, in 1978. He received the Italian Laurea (Hons.) degree in electronic engineering from the University of Ancona (now Università Politecnica delle Marche), Italy, in 2004, and the Ph.D. degree from the Università Politecnica delle Marche, in 2009. He has been a Postdoctoral Researcher, since 2010. His current research interests are in the area of digital signal processing and computational intelligence, with the special focus on energy management and speech/audio processing. He has actively participated to various (funded) regional, national, and European projects on multimedia Digital Signal Processing. He is the author and coauthor of many international scientific peer-reviewed articles, and has been serving as a Reviewer for several international journals and conference proceedings. He is a member of the Editorial Board of the *Neural Computing and Applications* (Springer) and the *Artificial Intelligence Review* (Springer), since 2017, and a member of the Program Committee of several international conferences. He has also served as the Guest Editor for the Special Issue on *Theory and Application of Computational Intelligence in Electric Vehicles and their Integration within Smart Energy Networks*Ï (Energies, MDPI). He has organized the Special Session on Deep Neural Audio processing within the IEEE International Joint Conference on Neural Networks in 2017, 2018, and 2019, respectively. He is also a member of the Texas Instrument Expert Advisory Panel and of the International Neural Networks Society.

STEFANO SQUARTINI was born in Ancona, Italy, in 1976. He received the Italian Laurea (Hons.) in electronic engineering from the University of Ancona (now Polytechnic University of Marche, UnivPM), Italy, in 2002, and the Ph.D. degree from the University of Marche, in 2005. He also worked as Postdoctoral Researcher with UnivPM, from 2006 to 2007, and subsequently, he joined the Department of Information Engineering (DII) as Assistant Professor in circuit theory. He has been an Associate Professor with UnivPM, since 2014. His current research interests are in the area of computational intelligence and digital signal processing, with the special focus on speech/audio/music processing and energy management. He is the author and coauthor of more than 200 international scientific peer-reviewed articles. He joined the Organizing and the Technical Program Committees of more than 70 International Conferences and Workshops in the recent past. He is a Senior Member of the IEEE and a member of the IEEE CIS. He is the Organizing Chair of the IEEE CIS Task Force on Computational Audio Processing. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, and the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, and also a member of the Cognitive Computation, Big Data Analytics, and Artificial Intelligence Reviews Editorial Boards.