Who Cried When: Infant Cry Diarization with Dilated Fully-Convolutional Neural Networks

Marco Severini, Emanuele Principi, Samuele Cornell, Leonardo Gabrielli, and Stefano Squartini
Department of Information Engineering
Università Politecnica delle Marche, Ancona, Italy
{m.severini,e.principi,l.gabrielli,s.squartini}@univpm.it, s.cornell@pm.univpm.it

***Abstract*—In this paper, we address the problem of the concurrent detection of multiple infant cries by using microphones located in the cribs of a Neonatal Intensive Care Unit (NICU). We term this task as *infant cry diarization* in resemblance with the "speaker diarization" task related to the speech signal: instead of determining "who spoke when", here the problem is determining "who cried when". The proposed algorithm consists of a fully-convolutional neural network (Conv-DetNet) that processes simultaneously all the audio signals acquired from the microphone in each crib and detects if the infants cried or not. The neural network takes as input Log-Mel coefficients and it is composed of stacked dilated convolutional blocks with increasing dilation factors. Each block is composed of pointwise and depthwise convolutional layers that replace standard convolutions with a mathematically equivalent but more efficient operation. The architecture has been compared to its single-channel equivalent and to single and multi-channel architectures presented in a previous work, composed of standard convolutional layers and fully-connected layers. The experiments have been conducted on a synthetic dataset that simulates the acoustic environment of the Salesi Hospital NICU located in Ancona (Italy). The results have been evaluated in terms of Area Under Precision-Recall Curve (PRC-AUC) and they showed that the proposed multi-channel Conv-DetNet achieves the highest performance with a PRC-AUC equal to 87.58%, outperforming all the comparative methods.**

***Index Terms*—Infant Cry Detection, Deep Neural Networks, Dilated Convolutions, Fully-convolutional networks**

I. Introduction

The acoustic analysis of infants' vocalizations provides valuable support to the medical staff for monitoring the health status of an infant and for detecting specific pathologies [1]. An advantage of this approach is its low level of intrusiveness since monitoring is performed by using contact-less sensors (i.e., microphones).

Infants' vocalizations can be analyzed at multiple abstraction levels, depending on the specific aspect the interest is on. A possible classification divides the different approaches in cry detection, pathology detection, pathology identification, and cause identification. Cry detection consists of determining the time boundaries of a vocalization [2], [3], and it can help the medical staff to determine the general health status of an infant by evaluating the total amount of crying activity in a period. Moreover, cry detection can be a pre-processing step that segments the audio signal, which is then analyzed by pathology detection, identification, or cause identification algorithms [4]. Pathology detection is a binary classification task where a cry is classified as normal or pathological [5], [6]. Pathology identification, on the other hand, analyses the cry signal for determining the type of pathology an infant is affected by (e.g., perinatal asphyxia) [7]–[9]. Differently, cause identification aims at discovering the underlying reason that elicited the cry (e.g., hunger, pain) [10]–[12].

This research was carried out within the "SINC - System Improvement for Neonatal Care Project", funded by Regione Marche, Italy within the POR MARCHE FESR 2014-2020 - ASSE 1 program.

The focus of this paper is on cry detection, specifically addressing the acoustic environment typically found in Neonatal Intensive Care Units (NICUs). In the literature, several works have been proposed for detecting infant cries, both in domestic environments and in hospital wards. Early works were based on pure signal processing methods [13]–[15]. In [13], the algorithm is based on the short-term energy measure of the audio signal, and a cry is detected if the value exceeds a certain threshold. A similar approach has been also adopted in [14]. A different method has been presented in [15], where the authors detected cry utterances by using Cepstral-based acoustic analysis.

More recent works are based on machine-learning methods that learn to identify cry signals directly from data. Mel-frequency cepstral coefficients (MFCCs) and k-nearest neighbors have been used in [16] to classify cry and non-cry units and to alert parents when infants are being left alone (either in apartments or vehicles). Abou-Abbas *et al.* [17] proposed Hidden Markov Models (HMMs) to detect and classify the inspiratory and expiratory phases of the cry. Gaussian Mixture Model (GMM) classifier and HMMs have been proposed in [3] along with short-time Fourier transform, empirical mode decomposition (EMD), and wavelet packet transform. Naithani *et al.* [2] discriminated the expiratory and inspiratory phases of a cry, and a third class including all other noises by using HMMs. Raboshchuk *et al.* [18] 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.

Convolutional Neural Networks have been used in [19]–[21]. In [19], the authors proposed a neural network for cry detection in domestic environments composed of three con-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

volutional layers and one fully-connected layer. The method provided significantly better results compared to a logistic regression classifier. The authors of [22] presented a solution targeted at low-power devices and proposed a novel set of features. Deep neural network (DNN) and support vector data description (SVDD) classifiers were evaluated by using a dataset composed of various recordings collected from public websites. Chang *et al.* [4] described a two-stage algorithm with the first stage devoted to the detection of cry segments. The method consists of pre-emphasizing the input signal and then in calculating the short-time Fourier transform. Cry detection is performed by using a network composed of 6 convolutional layers and one fully-connected layer. In our previous work [20], we presented a cry detection method that uses eight channels of a circular microphone array. The algorithm comprised a linear-constraint minimum-variance (LCMV) beamformer followed by an optimally-modified log-spectral amplitude (OMLSA) post-filter for reducing the noise contributions of the acoustic environment. Cry detection, then, was performed by using a neural network composed of three convolutional layers followed by one fully-connected layer. In our later work [21], the study was extended and the method presented in [20] was compared to different single-channel and multi-channel neural networks. Moreover, the feature extraction stage was modified in order to consider the spectral characteristics of the NICU acoustic environment, and the acoustic scene simulation strategy was adopted to train the network and reduce the need for real data acquired in NICUs. The results showed the superiority of pure DNN approaches and evidenced the effectiveness of the acoustic scene simulation strategy.

The case study explored in this paper is similar to the one considered in [21], where we addressed cry detection in NICUs by using a single crib equipped with a microphone. NICUs, however, represent a particularly challenging environment since multiple cribs are present and the possibility that several infants cry simultaneously is high. Here we address the problem by considering the case where multiple cribs in the NICU are equipped with a microphone for detecting the cries of the infants located in them. Differently from [21], thus, here the task consists in the simultaneous detection of the cries coming from every crib equipped with a microphone, i.e., in determining "who cried when". From here on, we will term this task as *infant cry diarization* in resemblance with the "speaker diarization" task [23] for the speech signal, where the objective is to determine "who spoke when".

For this task, we propose a fully-convolutional neural network composed of multiple stacked dilated convolutional stages that operate at increasing dilation factors. The network uses Log-Mel coefficients as input extracted from the microphone signals of all the cribs equipped with a microphone, thus exploiting the information in all the acquired audio signals. In the final layer, the network gives as output if a cry is present or not in each audio signal.

For training the network and evaluating the performance we created a new synthetic dataset that simulates the acoustic environment of the NICU located in the Salesi Hospital (Ancona, Italy). Differently from [21], the dataset used in this paper simulates the presence of two microphones located in two target cribs. The proposed approach has been compared to single and multi-channel architectures composed of standard convolutional layers and fully-connected layers, and to the single-channel architecture presented in [21]. Instead of operating on all the acquired audio signals, single-channel architectures operate on the signals of each crib individually. Apart from the architecture presented in [21], the hyperparameters of all the networks have been determined by performing a Bayesian search [24] and 4-fold cross-validation. The results show that the proposed approach outperforms all the comparative methods in terms of Area Under Precision-Recall Curve (PRC-AUC).

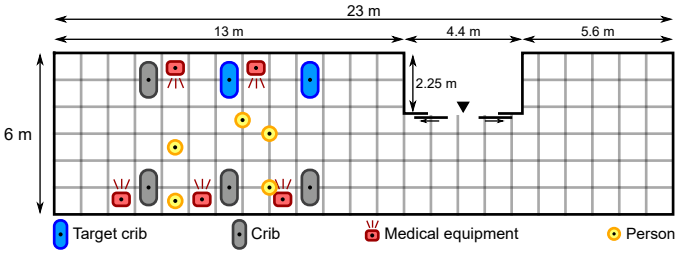


Fig. 1. Plan of the NICU of the Salesi Hospital (Ancona, Italy). Target cribs are highlighted in blue. Other cribs, medical equipments, and persons represent possible sources of interference.

The outline of the paper is the following. Section II describes the case study and the infant cry diarization task. Section III describes the proposed algorithm for cry detection. The comparative method is briefly introduced in Section IV-C, whereas Section IV presents the experiments performed to evaluate the proposed approach, and the obtained results. Finally, Section V concludes the paper and presents future developments.

II. Case Study

In this paper, we consider a NICU environment where multiple cribs are present, and a subset of them is equipped with a microphone. Fig. 1 shows the plan of the NICU of the Salesi Hospital located in Ancona, Italy, where two cribs are equipped with a microphone. The figure shows also possible sources of noise, such as medical equipment, persons, or other cries we are not interested in recognizing. Fig. 2 shows the scheme of a crib and the position of the microphone. In this scenario, the interest is on detecting when the infants located in the cribs equipped with a microphone are crying or not, i.e., "who cried when".

III. Proposed Method

The method proposed here for infant cry diarization is depicted in Fig. 3. The figure shows D audio signals coming from the cribs equipped with a microphone that are processed individually by a Log-Mel extraction stage. Log-Mels are then concatenated and used as input to the neural network (ConvDetNet), that, for each signal, outputs a 1 if it detected a cry

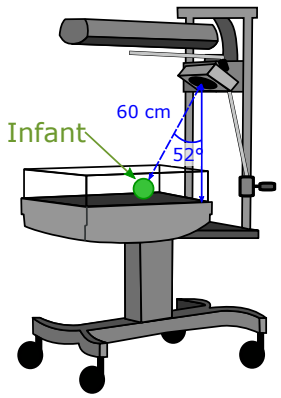


Fig. 2. Scheme of the recording setup.

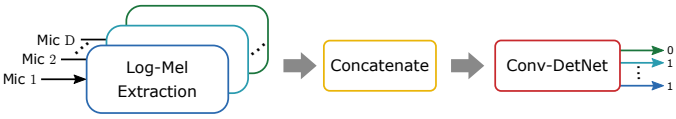


Fig. 3. Block diagram of the proposed infant cry diarization method. For each audio signal, the network outputs a 1 if a cry is detected, or a 0 otherwise.

and a 0 otherwise. The remainder of this section describes in detail the feature extraction stage and the architecture of Conv-DetNet.

A. Feature Extraction

The audio signals acquired from the microphones in the cribs are sampled at 16 kHz, and Log-Mel coefficients are then calculated by dividing them in frames 20 ms long and overlapped by 10 ms. After calculating the Fast-Fourier Transform of each frame, the signal is filtered by using a triangular filter-bank composed of 20 filters equally spaced in the mel-frequency space. The frequency range of the filter-bank is 4 kHz-8 kHz, thus discarding the spectral components below 4 kHz. This choice derives from our previous work, where we showed in NICUs the majority of the energy of the noise signals is concentrated below 4 kHz [21]. The final feature vector is obtained by calculating the energy in each sub-band, and then by applying the logarithm operator. For each frame, a vector of 20 coefficients is therefore obtained. Log-Mel coefficients are calculated for each audio channel independently. The Log-Mel feature vector related to the microphone of crib i extracted at the time-frame k will be denoted as $\mathbf{x}_k^{(i)}$ in the following. The feature vector has dimension $N \times 1$, with $N = 20$ in this case.

B. Multi-Channel Conv-DetNet

The neural network architecture used here for infant cry diarization is inspired to the Conv-TasNet proposed in [25], and will be referred to as Conv-DetNet from here on. Conv-TasNet is a fully-convolutional network that processes the raw audio waveforms and it is used for single-channel audio source separation. The network is composed of an encoder that extracts higher-level features from the raw audio signal, a separation stage that estimates a set of multiplicative masks, and a decoder that converts the higher-level representation back to the audio domain. Differently from [25], in the proposed method the encoder and the decoder are not present, and the final layers have been modified to output the decision on the presence of infant cries.

Details of the architecture of the network are shown in Fig. 4. Supposing that D cribs are equipped with a microphone, the input to the network is composed by concatenating the Log-Mels of all cribs to create the feature vector $\mathbf{x}_k = [\mathbf{x}_k^{(1)}, \ldots, \mathbf{x}_k^{(D)}] \in \mathbb{R}^{1 \times D \cdot N}$. Then, C vectors preceding and following $\mathbf{x}_k$ are concatenated to form the final feature matrix:

$$\mathbf{X}_k = \begin{bmatrix} \mathbf{x}_{k-C} \\ \mathbf{x}_{k-C+1} \\ \vdots \\ \mathbf{x}_k \\ \vdots \\ \mathbf{x}_{k+C-1} \\ \mathbf{x}_{k+C} \end{bmatrix} \in \mathbb{R}^{(2C+1) \times D \cdot N}. \quad (1)$$

Concatenating $2C$ vectors preceding and following the Log-Mel coefficients extracted from the k-the frame allows to exploit the temporal information of the surrounding features.

As shown in Fig. 4, the first processing stage of the network is layer normalization on $\mathbf{X}_k$ and it is performed as follows [26]:

$$\tilde{\mathbf{X}}_k = \frac{\mathbf{X}_k - E[\mathbf{X}_k]}{\sqrt{Var[\mathbf{X}_k] + \epsilon}} \odot \gamma + \beta, \quad (2)$$

$$E[\mathbf{X}_k] = \frac{1}{2N(2C+1)} \sum_{2N(2C+1)} \mathbf{X}_k, \quad (3)$$

$$Var[\mathbf{X}_k] = \frac{1}{2N(2C+1)} \sum_{2N(2C+1)} (\mathbf{X}_k - E[\mathbf{X}_k])^2, \quad (4)$$

where ϵ is a small constant that avoids division by zero, and (γ, β) are trainable parameters.

After normalization, $\tilde{\mathbf{X}}_k$ is processed by a convolutional layer with kernel size 1 and B channels (1×1-Conv), then by several groups of stacked 1-D dilated convolutional blocks (1-D Conv). The architecture of these blocks is shown in Fig. 5 [27], and it consists of a 1×1-Conv layer with H channels followed by Parametric Rectified Linear Unit (PReLU) activation function [28], layer normalization, depthwise convolution (D-Conv), PReLU, Dropout, layer normalization, and two 1×1-Conv blocks at the end. D-Conv and 1×1-Conv blocks are used to form the so-called *depthwise separable convolution* operation. This has been proposed in the literature since it is mathematically equivalent to standard convolution, but it requires a significantly less number of trainable parameters [29]. The D-Conv block performs a dilated convolution operation with kernel size P. Details on depthwise separable convolution are reported in [25], [29].

Referring to Fig. 5, the left output is a skip-connection with dimension $(2C + 1) \times Sc$ that is summed to the skip-

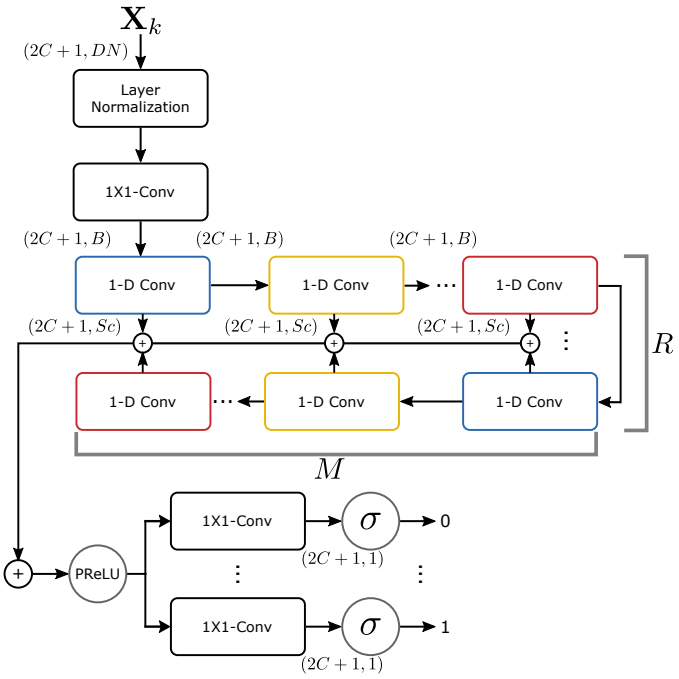


Fig. 4. Architecture of the neural network used for infant cry detection. 1-D Conv blocks in the same row have increasing dilation factors, while blocks with the same colors in different rows share the same factor. The symbol "σ" represents the sigmoid activation function. The notation (X, Y) indicates that the dimension of the matrix at the input or the output of a block is $X \times Y$.

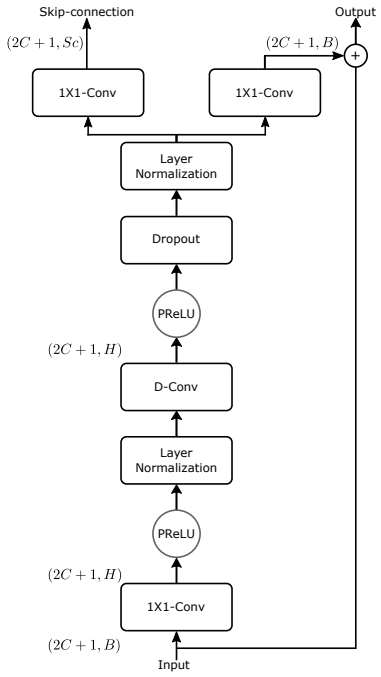


Fig. 5. Architecture of the 1-D Conv blocks.

connections of all the 1-D Conv blocks. The right output is a residual path with dimension $(2C+1) \times B$, and it is used as input for the following 1-D Conv blocks. 1-D Conv blocks are stacked with increasing dilation factors, allowing the network to exploit the information of the temporal context. As in [25], dilation factors are increased exponentially $1, 2, 4, \ldots, 2^{M-1}$, where M is the number 1-D Conv block, and each group of M 1-D Conv blocks is repeated R times. In the final stages, the sum of the skip-connections of all the 1-D Conv blocks is processed by a PReLU activation function, and by D 1×1-Conv layers followed by a sigmoid activation function the produce the final output. Note that the input is padded so that the output of each block has the same time dimension $(2C+1)$.

The network has been trained by using the binary cross-entropy loss for each output and the Adam optimizer [30].

C. Single-Channel Conv-DetNet

Along with the multi-channel architecture presented in the previous section, a single-channel solution has also been evaluated. In this case, instead of having a single network that processes all the microphone signals, D separate networks process individually the signals coming from the cribs. The architecture of single-channel networks is similar to the one shown in Fig. 4: in this case, however, each network is given as input a feature matrix composed of only the Log-Mel coefficients extracted from a single microphone. Moreover, in the output layer, only a single branch is present after the PReLU activation function, since the network outputs a single value.

IV. Experiments

The proposed method has been implemented by using Keras with TensorFlow as back-end and *librosa* [31] for Log-Mel extraction. In the following, we will describe the adopted datasets, the experimental setup, the comparative methods, and discuss the obtained results.

A. Dataset

The experiments have been conducted by using a synthetic dataset similar to the one presented in [21]. Compared with the one presented in [21], here we did not simulate a circular microphone array, but a single omnidirectional microphone. Moreover, in [21] only one crib is equipped with a microphone whereas in this case two target cribs have been considered. As in [21], the synthetic datasets simulate the acoustic environment of the Salesi Hospital NICU (Fig. 1). The impulse responses between an audio source and the microphones located in the blue cribs shown in Fig. 1 have been created by using Pyroomacoustics [32].

In addition to the cries of the target cribs, we have considered the presence of the following sources of interference:

- human speech: it considers the presence of persons (parents, medical staff) in the NICU;
- infant cry: it considers the presence of other infants that are not monitored (gray cribs in Fig. 1);
- "beep" sound: it represents the typical noises of a medical equipment;
- fan noise: it considers the presence of a Heating, Ventilation and Air Conditioning system;

TABLE I
DETAILS OF THE DATASET. FOR EACH SCENARIO, THE TABLE SHOWS THE TYPE OF NOISE, THE LENGTH IN MINUTES OF THE CRY SIGNALS FOR THE TARGET CRIBS, AND THE AMOUNT OF OVERLAP WITH THE NOISE SIGNAL AND BETWEEN THE TWO TARGET CRIBS.

		Cry Length (min)		Overlap (%)	
Scenario	Noise	Crib 1	Crib 2	Noise	Crib 1&2
1	Interferent Cry	29.67	29.71	58.25	66.34
2	Beep	44.57	44.38	43.34	58.94
3	Speech	44.49	44.38	51.57	58.31
4	Fan, Oxygen	44.53	44.61	100.00	58.44

- oxygen concentrator: it considers the presence of an oxygen concentrator.

Human speech, infant cry, beeps are coherent noise sources placed in the positions shown Fig. 1. The speech signals have been extracted from the WSJ0 dataset [33], which contains mono clean speech signals of American English sentences. Infant cries and the other noise sources have been collected from different web sources[1,2]. All signals are sampled at 16 kHz.

The total number of cry recordings is 64, and they belong to 29 different subjects. Based on the acoustic environment depicted in Fig. 1, and the different noise types mentioned above, 4 scenarios at SNR 0 dB and 5 dB have been devised. For each scenario multiple audio sequences of 30 s have been created, each differing for the positions of the noise source in the room and the combination of cry sequences appearing in the target cribs. The sequences have been created so that the simultaneous presence of the same subject in more than one crib is avoided. The resulting 352 sequences have been divided into two sets of 176 elements each. The first half has been used to carry out a 4-fold cross-validation, the second half has been used to test the network performance.

B. Experimental setup

The experiments have been conducted by using a 4-fold cross-validation, and the topology of each architecture and the related hyper-parameters have been determined by using Bayesian optimization with Hyperopt-Keras [24]. Table III shows the hyper-parameters search space used in the single and multi-channel Conv-DetNet Bayesian optimization. After this phase, the determined architectures have been evaluated in the test set of the dataset described previously.

The performance has been evaluated by means of the Area Under the Precision-Recall Curve (PR-AUC) which is calculated as follows:

$$\text{PR-AUC} = \sum_n (R_n - R_{n-1}) \cdot P_n, \tag{5}$$

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:

$$R = \frac{TP}{TP + FN}, \quad P = \frac{TP}{TP + FP}, \tag{6}$$

where the subscript n has been omitted for simplicity.

C. Comparative Methods

The performance of the proposed method has been compared to architectures similar to the ones presented in our previous work [21]. Specifically, we evaluated the Half-band 1Ch-DNN topology identified in [21], and we determined an additional single-channel topology (1Ch-DNN) by using Bayesian optimization. The Half-band 1Ch-DNN architecture is composed of 3 convolutional layers and 3 fully-connected layers (details of the hyper-parameters values are shown in Table II). The 1Ch-DNN architecture was determined by using the search space shown in Table II, and it is composed of 2 convolutional layers, each followed by batch normalization, rectifier linear unit (ReLU) activation function, max-pooling, and dropout. The second part is composed of two fully-connected layers followed by a single neuron with a sigmoid activation function, that outputs the probability of the central frame being a cry. Details on the final hyper-parameters values are shown in Table II.

In addition to single-channel architectures, we evaluated a two-channel network (2Ch-DNN) which processes all the audio signals coming from the target cribs as the proposed multi-channel Conv-DetNet. As the 1Ch-DNN architecture, the hyperparameters of the networks have been determined by performing a Bayesian search [24] (Table II). The final architecture is composed of a separate branch for each audio channel, with each branch composed of 1 convolutional layer, followed by batch normalization, ReLU activation function, max-pooling, and dropout. The output of each branch is then concatenated and processed by a convolutional layer followed by batch normalization, ReLU activation function, max-pooling, and dropout. For each output branch, the second part is composed of two fully-connected layers each followed by a dropout layer and a single neuron with sigmoid activation.

D. Results

Table IV shows the results obtained for the single and multi-channel Conv-DetNet (respectively 1Ch-Conv-DetNet and 2Ch-Conv-DetNet) as well as for the comparative methods in the 4-fold cross-validation phase. As shown in the table, the PRC-AUC of the Half-band 1Ch-DNN presented in [21] is 69.93%, i.e., about 13 percentage points lower than the score achieved over the previous synthetic dataset (see Table 3 in [21]). From this result, it is possible to conclude that cry detection in the current dataset is more challenging. Since in this case we consider 2 target cribs, the overlap between the two targets cries occurs in every sequence, even in scenarios that do not include the interfering cries from not monitored subjects.

The 1Ch-DNN architecture is similar to Half-band 1Ch-DNN, but it has been obtained by performing a Bayesian

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

TABLE II

HYPERPARAMETERS EXPLORED IN THE BAYESIAN SEARCH FOR THE 1CH-DNN AND THE 2CH-DNN ARCHITECTURES. "U": UNIFORM DISTRIBUTION; "$\log U$" UNIFORM DISTRIBUTION IN THE LOG-DOMAIN.

Parameter (Distribution)	Range	1Ch-DNN	2Ch-DNN	Half-band 1Ch-DNN
Batch size (U)	{64, 128, 256}	64	256	512
Learning Rate ($\log U$)	$[5.14 \cdot 10^{-6}, 45.54 \cdot 10^{-53}]$	$2.94 \cdot 10^{-5}$	$3.05 \cdot 10^{-5}$	$2.18 \cdot 10^{-3}$
CNN layers				
Nr. of CNN layers (U)	$\{1, 3\}$	2	1	3
Kernel shape (U)	$[1, 10] \times [1, 10]$	6×8, 1×10	9×6	1×1, 1×1, 1×1
Kernel number ($\log U$)	$[16, 64]$	46, 42	37	63, 18, 19
Strides ($\log U$)	$[1, 10] \times [1, 10]$	4×8, 1×10	6×6	2×4, 5×1, 4×1
Pooling Shape (U)	$[1, 10] \times [1, 10]$	4×3, 1×1	2×7	2×1, 2×2, 2×1
Pooling Strides (U)	$[1, 10] \times [1, 10]$	2×1, 1×1	2×2	1×2, 1×2, 1×2
Dropout Rate (U)	$[0, 0.5]$	0.21, 0.22	0.42	0.1, 0.2, 0.3
Last CNN Layer				
Kernel shape (U)	$[1, 10] \times [1, 10]$	-	1×3	-
Kernel number ($\log U$)	$[16, 64]$	-	53	-
Strides ($\log U$)	$[1, 10] \times [1, 10]$	-	1×3	-
Pooling Shape (U)	$[1, 10] \times [1, 10]$	-	1×4	-
Pooling Strides (U)	$[1, 10] \times [1, 10]$	-	1×2	-
Dropout Rate (U)	$[0, 0.5]$	-	0.31	-
Fully-connected layers				
Nr. of fully-connected layers (U)	$\{1, 3\}$	2	2	3
Units $\log U$	$[100, 1024]$	101, 65	227, 131	154, 113, 107
Dropout Rate (U)	$[0, 0.5]$	0	0.31, 0.48	0.5, 0.5, 0.5
Number of trainable parameters	-	32,831	220,969	49.570

TABLE III

HYPERPARAMETERS EXPLORED IN THE BAYESIAN SEARCH FOR THE 1CH-CONV-DETNET AND THE 2CH-CONV-DETNET ARCHITECTURES. "U": UNIFORM DISTRIBUTION; "$\log U$" UNIFORM DISTRIBUTION IN THE LOG-DOMAIN.

Parameter	Distribution	Range	1Ch-Conv-DetNet	2Ch-Conv-DetNet
Batch size	U	{32, 64, 128}	128	2048
Learning rate	$\log U$	$[1.01 \cdot 10^{-3}, 9.99 \cdot 10^{-3}]$	$4.73 \cdot 10^{-3}$	$4.06 \cdot 10^{-3}$
Hidden channels (H)	U	$[20, 40]$	31	34
Skip channels (Sc)	U	$[20, 40]$	30	30
Kernel size (P)	U	$[1, 5]$	2	4
Dropout rate	U	$[0, 0.5]$	0.12	0.16
Nr. of blocks (M)	U	$[1, 5]$	2	2
Nr. of repeats (R)	U	$[1, 5]$	2	3
Number of trainable parameters			9,749	25,002

TABLE IV

PR-AUC (%) FOR THE PROPOSED AND THE COMPARATIVE METHODS (4-FOLD VALIDATION).

Algorithm	Crib 1	Crib 2	Average
Half-band 1Ch-DNN [21]	67.83	72.03	69.93
1Ch-DNN	70.74	70.17	70.45
2Ch-DNN	78.80	79.46	79.13
1Ch-Conv-DetNet	69.31	73.76	71.54
2Ch-Conv-DetNet	**83.08**	**84.56**	**83.82**

TABLE V

PR-AUC (%) FOR THE PROPOSED AND THE COMPARATIVE METHODS (TEST).

Algorithm	Crib 1	Crib 2	Average
Half-band 1Ch-DNN [21]	55.86	55.25	55.56
1Ch-DNN	54.44	54.30	54.37
2Ch-DNN	81.52	80.06	80.79
1Ch-Conv-DetNet	57.04	55.68	56.36
2Ch-Conv-DetNet	**86.46**	**88.69**	**87.58**

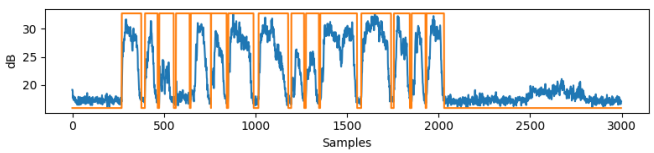
optimization on the dataset used in this paper. The related score is 70.45%, and although higher compared to the Half-band 1Ch-DNN architecture, the improvement is very limited (0.5 percentage points), demonstrating the difficulty of the task for this type of single-channel topologies. In the case of the 1Ch-Conv-DetNet, the achieved score is about 71.54%, i.e., about 1 percentage point above the score of the 1Ch-DNN architecture. In spite of the slight improvement, given the same setup, the 1Ch-Conv-DetNet is a much smaller network than the 1Ch-DNN, hence resulting a much less computationally intensive solution. Nonetheless, it is possible to conclude that with the use of a single microphone, the cry detection task might become more challenging whenever the overlap of cries from close subjects occurs.

On the other hand, the use of multiple input channels provides additional information, which might help to discriminate between the cries of subjects. Concerning this, the 2Ch-DNN configuration scores a PR-AUC of about 79.13%. With respect to the results from the single-channel networks, the score has increased of about 9 percentage points showing that the use of multiple input channels provides an edge toward the discerning abilities of the network. In fact, a single channel network has limited information and, therefore, might lack the ability to discern whether a specific intensity of cry depends on the subject or the distance of the infant from the microphone.

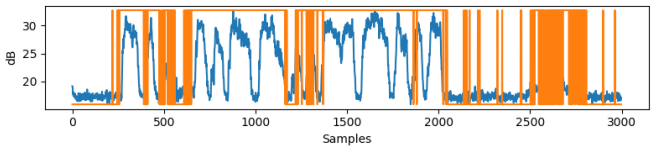
In the same scenario, however, the 2Ch-Conv-DetNet achieves a score of about 83.82%, thus outperforming the 2Ch-

DNN of about 4.69 percentage points, thus proving to be fairly robust against overlapping cries. Additionally, the 2Ch-Conv-DetNet is about 8.8 times smaller than the 2Ch-DNN, since its number of trainable parameters is 25,002 compared to 220,969 of the 2Ch-DNN (see Table II and Table III), resulting in a more computationally efficient approach to the cry diarization task. To exemplify the results for each network over a single sequence, one of the sequences of the set used in the 4-fold cross-validation has been selected. This sequence includes the cries of the two subjects and the fan noise at SNR 5 dB (Scenario 4 in Table I). The energy profile of this sequence has been computed by calculating the sum of the energy values of the 20 mel bands of the filter-bank, thus frequency components below 4 kHz are not present. The profile is shown in Fig. 6 in blue together with the ground-truth and output of each network in orange. Concerning the energy profile, it should be noted that since the spectral components below 4 kHz have been discarded, the noise level is quite low with respect to the energy cry. Observing the plots reveals that in the case of both 1Ch-DNN and 2Ch-DNN, the detected cry sequences are larger than the one present in the ground truth revealing the presence of false positives. Moreover, few fairly large cry sequences are missed around samples 500, 1250. The 1Ch-DNN network detects a false positive after sample 2500. Whenever detection errors occur, the value switches frequently from high to low meaning that the network identifies a few samples but not the entire sequence. In the case of 2Ch-Conv-DetNet and 1Ch-Conv-DetNet the cry detection appears more accurate. The length of cry sequences resembles more the ground truth. In this case, a sequence has been missed around sample 600. In the case of 1Ch-Conv-DetNet non-existent cry detections appear at sample 0, and above sample 2500.

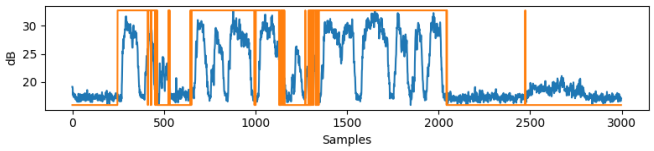
The results on the test set reported in Table V, further confirm that cry overlaps strongly impairs the detection ability of the single-channel configurations. Specifically, the score of the Half-band 1Ch-DNN is 55.56%, whereas 1Ch-DNN and 1Ch-Conv-DetNet achieve 54.37 and 56.36 respectively, hence all the single-channel configurations show a performance drop of roughly about 15 percentage points compared to their respective score in the 4-fold cross-validation. On the other hand, the 2Ch-DNN and the 2Ch-Conv-DetNet show improved performance. Indeed, the 2Ch-DNN configuration achieves a score of about 80.79%, 1.66 percentage points higher than the score of the 4-fold cross-validation. The 2Ch-Conv-DetNet scores 87.58%, hence 3.76 additional percentage points with respect to the 4-fold cross-validation. Overall the 2Ch-Conv-DetNet improves the score of the 2Ch-DNN of about 6.79 percentage points. From a different angle, the test shows an increased performance gap between the scores of single channels networks and 2 channel networks, thus proving the trend shown by the 4-fold cross-validation scores. The behavior of the 2-channels networks can be explained by considering that they are trained by using all the dataset used in the cross-validation phase (i.e., 176 sequences). On the other hand, in the cross-validation phase, single-channel networks seem to overfit on the validation set. The consequence is the poor generalization capabilities when evaluated on the test set.



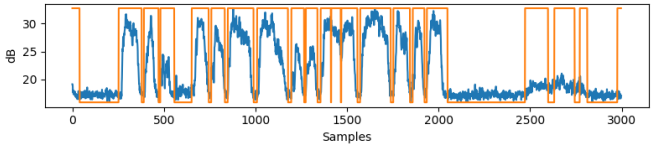
(a) Ground truth.



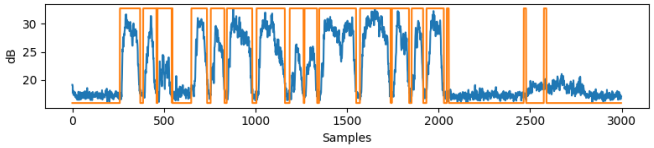
(b) 1Ch-DNN.



(c) 2Ch-DNN.



(d) 1Ch-Conv-DetNet.



(e) 2Ch-Conv-DetNet.

Fig. 6. Cry detections for different network types against cry signal energy from the 4-fold cross-validation training set with fan noise and SNR 5 dB.

V. Conclusion

This paper presented an infant cry diarization method based on dilated fully-convolutional networks for NICU environments. The objective is to determine "who cried when", i.e., the portions of the audio signal where each infant cried in a scenario where multiple cribs are equipped with a dedicated microphone. The proposed method consists of a feature extraction stage that calculates Log-Mel coefficients for all the acquired audio signals and determines the presence of cries by using concurrently the Log-Mels of all signals. The proposed neural network, named Conv-DetNet, is composed of stacked dilated convolutional blocks with increasing dilation factors, and each block comprises pointwise and depthwise convolutional stage that guarantee computational efficiency. Along with the multi-channel Conv-DetNet, a single-channel archi-

tecture operating individually on the microphones of each crib has been evaluated. Conv-DetNet has been also been compared to single and multi-channel architectures composed of standard convolutional and fully-connected layers (respectively 1Ch-DNN and 2Ch-DNN) and to the single-channel architecture presented in our previous work (Half-band 1Ch-DNN) [21]. The experimental evaluation has been conducted on a synthetic dataset that simulates the acoustic environment of the NICU of the Salesi Hospital located in Ancona, Italy. The dataset considers six cribs, two equipped with a microphone and four possible sources of interferent cries. Moreover, the dataset comprised multiple acoustic scenarios where the noises commonly found in NICU have been simulated. The results showed that the multi-channel architectures achieve the highest performance and that the proposed approach outperforms the 2Ch-DNN network by 6.79 percentage points.

Future works will explore different neural network architectures such as attention mechanism [34], and transfer learning methods such as domain adaptation [35] to adapt networks trained on synthetic datasets to real NICU environments.

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] 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 Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, p. 1, Dec. 2018.

[3] 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.

[4] C.-Y. Chang and L.-Y. Tsai, "A CNN-Based Method for Infant Cry Detection and Recognition," in *Proc. of AINA*, Matsue, Japan, Mar. 27-29 2019, pp. 786–792.

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

[6] C. Ji, X. Xiao, S. Basodi, and Y. Pan, "Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features," in *Proc. of the 5th IEEE Int. Conf. on Smart Data*, Atlanta, GA, USA, Jul. 14-17 2019, pp. 1233–1240.

[7] C. C. Onu, J. Lebensold, W. L. Hamilton, and D. Precup, "Neural transfer learning for cry-based diagnosis of perinatal asphyxia," in *Proc. of Interspeech*, Graz, Austria, Sep. 15-19 2019, pp. 3053–3057.

[8] 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. of MICAI*, Atizapán de Zaragoza, Mexico, Oct. 27-31 2008, pp. 330–335.

[9] 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 noise of an information?" in *Proc. of International Congress and Exposition on Noise Control Engineering*, vol. 2004, no. 5, Czech Republic, Prague, 2004, pp. 2774–2781.

[10] V. K. 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.

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

[12] A. Chittora and H. A. Patil, "Newborn infant's cry analysis," *International Journal of Speech Technology*, vol. 19, no. 4, pp. 919–928, 2016.

[13] 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.

[14] 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," *Biomedical Signal Processing and Control*, vol. 7, no. 1, pp. 43–49, Jan. 2012.

[15] B. Reggiannini, S. Sheinkopf, H. Silverman, X. Li, and B. Lester, "A flexible analysis tool for the quantitative acoustic assessment of infant cry," *Journal of Speech Language and Hearing Research*, vol. 56, no. 5, pp. 1416–1428, 2013.

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

[17] L. Abou-Abbas, H. F. Alaie, and C. Tadj, "Automatic detection of the expiratory and inspiratory phases in newborn cry signals," *Biomedical Signal Processing and Control*, vol. 19, pp. 35–43, May 2015.

[18] 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," *Biomedical Signal Processing and Control*, vol. 39, pp. 390–395, 2018.

[19] 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.

[20] 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. of EUSIPCO*, Rome, Italy, Sep. 3-7 2018, pp. 997–1001.

[21] M. Severini, D. Ferretti, E. Principi, and S. Squartini, "Automatic Detection of Cry Sounds in Neonatal Intensive Care Units by Using Deep Learning and Acoustic Scene Simulation," *IEEE Access*, vol. 7, pp. 51 982–51 993, 2019.

[22] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: a comparison of hand crafted features and deep learning approach," in *Proc. of EANN*, Athens, Greece, Aug. 25-27 2017, pp. 168–179.

[23] X. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[24] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Prof. of ICML*, no. PART 1, Atlanta, GA, USA, Jun. 16-21 2013, pp. 115–123.

[25] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[26] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Proc. of the ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, Sep. 13-15 2016, pp. 125–125.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. of the ICCV*, Santiago, Chile, Dec. 11-18 2015, pp. 1026–1034.

[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of CVPR*, Honolulu, HI, USA, Jul. 21-26 2017, pp. 1800–1807.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, San Diego, CA, USA, May 7-9 2015.

[31] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proc. of the 14th Python in Science Conference*, Austin, Texas, USA, 2015, pp. 18–25.

[32] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *Proc. of ICASSP*, Calgary, Canada, Apr. 15-20 2018, pp. 351–355.

[33] 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. of ICASSP*, Detroit, MI, USA, May 9-12 1994, pp. 81–84.

[34] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: Different processes and overlapping neural systems," *Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014.

[35] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, 2016.