A Self-Training Automatic Infant Cry Detector

Gianpaolo Coro[1*], Serena Bardelli[2], Armando Cuttano[2,3], Rosa T. Scaramuzzo[2,3] and Massimiliano Ciantelli[2,3]

[1]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via Moruzzi 1, Pisa, 56124, Italy.
[2]Centro di Formazione e Simulazione Neonatale NINA, Dipartimento Materno-Infantile, AOUP, Via Roma 67, Pisa, 56126, Italy.
[3]Unitá Operativa Neonatologia, Dipartimento Materno-Infantile, AOUP, Via Roma 67, Pisa, 56126, Italy.

*Corresponding author(s). E-mail(s): gianpaolo.coro@cnr.it;

Abstract

Infant cry is one of the first distinctive and informative life signals observed after birth. Neonatologists and automatic assistive systems can analyse infant cry to early-detect pathologies. These analyses extensively use reference expert-curated databases containing annotated infant-cry audio samples. However, these databases are not publicly accessible because of their sensitive data. Moreover, the recorded data can under-represent specific phenomena or the operational conditions required by other medical teams. Additionally, building these databases requires significant investments that few hospitals can afford. This paper describes an open-source workflow for infant-cry detection, which identifies audio segments containing high-quality infant-cry samples with no other overlapping audio events (e.g., machine noise or adult speech). It requires minimal training because it trains an LSTM-with-self-attention model on infant-cry samples automatically detected from the recorded audio through cluster analysis and HMM classification. The audio signal processing uses energy and intonation acoustic features from 100 ms segments to improve spectral robustness to noise. The workflow annotates the input audio with intervals containing infant-cry samples suited for populating a database for neonatological and early diagnosis studies. On 16 minutes of hospital phone-audio recordings, it reached

sufficient infant-cry detection accuracy in 3 neonatal care environments (nursery-69%, sub-intensive-82%, intensive-77%) involving 20 infants subject to heterogeneous cry stimuli, and had substantial agreement with an expert's annotation. Our workflow is a cost-effective solution, particularly suited for a sub-intensive care environment, scalable to monitor from one to many infants. It allows a hospital to build and populate an extensive high-quality infant-cry database with minimal investment.

Keywords: Artificial Intelligence, Neonatology, Infant Cry Detection, Audio Processing, Machine Learning, Early Diagnosis

1 Introduction

Despite the advances in health technology and early diagnosis techniques, infant mortality is still high, with ∼2.4 million babies dying within the first month of life every year [1]. Newborns (infants with less than 28 days) represent ∼47% of the total deaths of children under five years. However, most newborn deaths could be early diagnosed and avoided if proper techniques were implemented [2]. Infant cry is among the most promising newborns' communication signals to study for early diagnosis. It is the first distinctive and informative life signal in born-at-term and often in preterm infants [3–7]. From a pediatric perspective, infant cry is the reflection of complex neurophysiological functions that can allow assessing a newborn's psychological and clinical status [8]. The functional multiplicity of cry has been studied in several scientific disciplines. Through cry, infants express emotional needs, physical pain, discomfort, and needs to caregivers [9]. Moreover, cry induces an internal stress signal in caregivers that triggers instinctive responses [10, 11]. From a psychological perspective, infant cry - as a social interaction modality - contains the core of emotional growth and long-term social skill development [8, 9]. Specific audio-signal spectral characteristics of infant cry are associated with emotional states, health status, gender, and gestational development conditions [12, 13]. The existence of pathology-related spectral characteristics is also largely documented [14–21]. In the last decades, several studies have adopted Artificial Intelligence methodologies to analyse infant cry and assist medical experts in the early diagnosis of neonatal pathological status. These studies have targeted several critical pathologies, such as deafness [22], asphyxia [23, 24], hypothyroidism [25–27], cleft palate [28, 29], brain damage [30, 31], autism [32–35], cri-du chat [36], respiratory distress syndrome [37], and other pathologies [38–41].

Most automatic approaches use databases that contain expert-curated audio annotations of infant cry recordings, which include spontaneous and induced cry [42]. The database annotations concern the type of cry, the infant identity, and the recording conditions (e.g., the microphone used, the noise level, and the people present) for each infant monitored. The automatic approaches commonly process acoustic features extracted from these databases and might also use

para-linguistic information and emotional cues [43–45]. The most common approaches use a four-step workflow [42]. First, the audio is normalised, silence and low frequencies are eliminated, and only high-energy parts are retained. Second, acoustic features are extracted from ∼10 ms audio segments that address the phonetic structure of the cry signal. Third, the features are post-processed to maximise information representation and reduce feature vector dimensionality. Fourth, a pattern recognition model is used to detect infant cry or detect pre-defined pathologies.

An infant-cry database can act as a knowledge base supporting neonatologists for studying, diagnosing, and annotating infant clinical status. A database-monitored infant has historical data recorded continuously, which a neonatologist can consult and compare in the case of a pathology onset evidence [46]. The medic can also add annotations in the database for future comparisons and share them with colleagues.

Recently, infant-cry databases have been crucial for expert studies that discovered unexpected features. For example, some studies have highlighted that infant-cry prosody (the modulations of cry pitch) depends on the mother's native language [47]. Other studies have demonstrated that prosodic acoustic stimuli are memorised during the third trimester of pregnancy and can be detected in a newborn's cry-signal spectrum [48, 49]. Moreover, newborns' prosodic-pattern learning greatly increases in the first months of life and adapts to the family's native-language prosody [50].

These observations demonstrate the importance of infant-cry databases for automatic and expert analyses and the role of prosody in infant cry decoding. However, because of the sensitive data they contain, infant-cry databases are publicly unavailable. Therefore, hospitals that need experimental data, practically turn to building their own databases through long-term data collection and annotation plans, which soon or later discourage most investments. Moreover, building an infant-cry database includes intrinsic difficulties that reduce the suitability of the experimental conditions for early-diagnosis experiments. For example, spontaneous cry is less frequent than induced cry and is usually under-represented [42]. Moreover, samples tend to include 1-2 month infants, who have better physiological and anatomical structures for cry production and vocalisation control than newborns [51].

This paper describes an open-source infant cry detection workflow that identifies *prominent* infant cry from audio recordings in real-life noisy neonatal care environments. A *prominent* infant-cry sample is an audio segment including one infant's cry sample with clear energy and intonation conditions and no other acoustic event overlapped. Our workflow is an easy and cost-effective solution to building an infant-cry database for early diagnosis by medics and automatic systems. Differently from other systems [52–54], it adapts to the particular noise, infants, medical operators' speech, and other sound sources present in the recording environment. The software is entirely open source. It can work with the audio recorded by cellular phones and thus does not require additional costs to the hospital and the medical operators. The workflow is a

machine-learning-based system, mostly unsupervised, which requires a minimal annotated training set (21 s) and self-trains on the analysed audio recording. It uses acoustic features extracted from ∼100 ms segments - a much larger window than the phonetic-scale window used by other approaches [22, 24, 52–55] - whose spectral structure is more robust to noise. The workflow embeds three machine learning models: an unsupervised model (cluster analysis), a minimally-supervised model (a Hidden Markov Model), and a self-training model (a Long Short Term Memory model with a self-attention layer). Its output is an annotation file accompanying the audio recording, indicating the intervals that contain the detected *prominent* infant cry.

The novelty of our workflow is that it is directly re-usable with new data and by other hospitals because it adapts itself to new recording conditions through self-training. Performance optimisation can be achieved using recording equipment that filters out complex interference sources like machine beeping sounds and human speech (Section 4). Our approach is innovative and cost-effective, considering the difficulty of accessing infant-cry databases and representing the large variability of infant speech through huge and expensive annotated databases. We demonstrate that our workflow performance is comparable with that of a reference supervised infant-cry detection system [52], which would require re-calibration in new operational conditions. Moreover, we use syllabic-scale acoustic features, which are more robust to noise [56–58] than the phonetic-scale features used by other approaches [52–54]. Finally, the fact that our software is completely open-source is uncommon in this context but highly valuable for hospitals.

In summary, our workflow aims to produce clear infant-cry samples to help early diagnosis. These samples are the indispensable basis for analyses by medical experts and automatic systems to discover new correlations between infant pathologies and cry, thus transforming observed illness signs into clearly reported symptoms.

2 Methods

Our infant cry detection workflow was entirely developed in Java to improve its potential embedding in small devices [59]. It is also entirely open source ("Code availability" statement). The workflow is constituted by a sequence of 5 computational *modules* (Figure 1):

- The "signal segmentation" module, which divides the signal into smaller units with acoustic characteristics that indicate the potential presence of infant cry (Section 2.2);
- The "energy and pitch extraction" module, which estimates energy and intonation features to robustly represent infant-cry related audio characteristics in noisy operational conditions (Section 2.3);
- The "cluster analysis" module, which optimally clusters features to potentially distinguish between infant cry and other audio types (Section 2.4);

- The "infant cry cluster identification" module, which identifies the clusters that probably contain infant cry (Section 2.5);
- The "infant cry detector" modules, which first train a complex machine learning model on the cry and non-cry cluster samples and then annotate the original audio file. A final module ("consecutive segment merging") merges consecutive audio segments containing infant cry (Section 2.6).

The workflow aims to detect audio segments containing *prominent* infant cry, i.e., audio intervals containing the clear cry of one infant at a time, without machine noise, medical operators' speech, and other audio events. These segments constitute valuable samples of infant cries to store for later analyses and studies.

2.1 Study cases

Nine study cases were selected from the audio recordings collected by the "Centro di Simulazione e Formazione Neonatale" (Centro NINA) in three neonatal care environments (Table 1): 3 recordings were conducted in *nursery* care (where healthy newborns are held); 4 in *sub-intensive* care (where pre-term newborns and infants coming from intensive care are monitored); and 2 in *intensive care* (where newborns with severe clinical states are monitored; normally abbreviated as ICU). These are the principal environments where infants are monitored by the neonatal care staff of the "Dipartimento Materno Infantile" of the Azienda Ospedaliero-Universitaria Pisana (Pisa, Italy). Neonatologists voluntarily conducted audio recording sessions under complete anonymity and lowest invasiveness constraints. The audio was recorded through cellular phone microphones at a 44,100 Hz sampling frequency by placing the phone in the middle of the room. The voices of the medical staff captured by the microphone were entirely anonymous, with no pseudo-labelling included. The identity of the infants present in the room was unknown too. Cellular phone usage was already allowed in neonatal environments and thus did not violate the hospital's policies. The medical staff annotated each audio recording by indicating the neonatal care environment and the most frequent cry stimuli (e.g., examination, hunger, physical needs, spontaneous). The study cases' recording conditions, infant cry types, and the number of different infants and operators recorded are summarised in Table 1 and Figure 2.

The average signal-to-noise ratio (SNR) was generally low across the environments (i.e., the noise level was high). It ranged between 9.44 and 16.86 in nursery care, 3.13 and 20.01 in sub-intensive care, and 4.40 and 5.05 in intensive care. The self-voluntary nature of the experiment limited the number of recordings. Nevertheless, 16 minutes of audio was collected, containing 3.3 minutes of prominent infant-cry samples from ∼20 infants. This material was suitable for conducting an inspectional experiment like the one presented in this paper, which indeed has a higher or comparable length to the test sets used by other infant-cry detection studies [52, 54]. Additional recordings were taken by placing the microphone close to crying infants to collect a specific corpus of

infant-cry samples with low surrounding noise. Given the constant and dynamic noise present in the environments, this operation was complex and resulted in collecting 21 s of cry samples from 8 infants. We used these 21 s audio samples of *prominent* infant cry as the minimal training set of our workflow.

A qualitative environmental stress level classification was associated to the study cases (Table 1 and Figure 2): We classified a crowded environment with several people talking loudly, a high noise level, and multiple infants crying as a *high stress-level* condition. Conversely, we classified an environment with few people talking and an averagely low-volume noise as a *medium stress-level* condition. The collected recordings did not include quiet conditions with low noise and few people talking (*low stress-level* condition). The recording conditions were sufficient to test our system in a real-life context where the stress level is averagely high or medium.

All recordings were manually annotated by an expert speech annotator to mark the segments containing prominent infant-cry samples. These annotations (3.3 minutes in total) represented gold-reference samples suited for populating an infant-cry database for automatic and medical analyses and diagnoses. Therefore, we measured how much our infant cry detector was able to capture these segments (Section 3.3).

The study cases had overall the advantage of forcing the workflow to achieve good performance in the critical, noisy, and real-life operating scenarios of neonatal care, with minimal burden on the medical staff.

2.2 Signal Segmentation

Our workflow input is an audio signal recorded by a cellular phone's microphone at 44,100 Hz. The high noise level of our study case recordings prevented using Automatic Speech Recognition (ASR) software to detect and remove human speech from the audio, because the phonetic structure of the speech was compromised in such conditions [58]. Therefore, to simplify audio processing and remove pure-noise audio segments, the first workflow module ("signal segmentation") divides the complete audio signal into smaller portions of coherent intonation that may include cry (Figure 1). These smaller segments allow the workflow to focus the analysis on audio "islands" with stable spectral characteristics [60]. We identified these segments as *tone units*, i.e., audio-signal portions with a high and continuous energy level. Energy is here intended as the squared sum of the samples of an audio-signal segment divided by the number of signal-samples (signal-segment *power*). Tone units have been used in spoken dialogue processing and to improve ASR performance because they mostly contain complete sentences [61, 62]. Our signal segmentation module uses a fast algorithm for tone unit detection [61], which requires setting a *tone unit window length* (in ms) parameter to calculate segment energy and find tone unit boundaries:

Algorithm 1 Signal Segmentation

Calculate the energy (e) of sequential signal segments with *tone unit window length* over the entire signal
For each e_i in the energy sequence:

Calculate the 1st-order sample derivative $d_i = e_i - e_{i-1}$
If $d_i < 0$ and $e_i <$ *tone unit energy threshold* $\rightarrow$ mark the window start time as a tone unit end.

The algorithm identifies a tone unit as the end of a sequence of high-energy signal segments. The *tone unit energy threshold* is an adaptive and iterative value initially set to a minimal value (0.001 dB), which doubles until at least three units are found. The *tone unit window length* parameter requires optimisation to maximise the workflow performance (Section 3.2). It should be set to discard noise-only segments and include segments that also contain infant cry. In our data, prominent infant cry was always included in high-energy tone units. As the output, the process divides the initial audio recording into shorter recordings. Tone units with lengths under 1s are discarded as containing either noise or sound bursts [62].

2.3 Energy and Pitch Extraction

Tone unit segmentation removes short islands of noise, low-energy signal segments, and sound bursts. The segments left can contain continuous adult speech and infant cry, overlapped with machine noise within a constant or increasing energy profile [61, 62]. As the next step, our workflow analyses each tone unit independently of the other. With our SNR levels, the phonetic-scale (10-20 ms) structure of the audio is corrupted, i.e., the short-windowed spectrogram contains too much noise to distinguish between speech, noise, and infant cry automatically [63]. Therefore, the analysis would be better conducted at a syllabic scale (100-250 ms), where the spectrum is more robust to high noise levels and the study of energy modulations is more reliable [62, 64, 65].

Our "energy and pitch extraction" module extracts syllabic-scale acoustic features of energy and prosody, which are particularly suited for infant cry identification [5, 66–68]. Syllabic-scale energy is the power of a 100-250 ms signal segment. Syllabic-scale pitch is the frequency of a *tone* associable with a syllable, and its time series represents the musicality and intonation of the signal. Our module calculates pitch through the Boersma's sound-to-pitch algorithm [69] - which uses signal autocorrelation - setting a cut-off frequency band between 50 and 500 Hz. It requires estimating the optimal *energy and pitch window length* to maximise the workflow performance (Section 3.2). This analysis window is passed over the signal to produce aligned energy and pitch time series for each tone unit. These time series constitute the basis of all further analyses in the workflow. Undefined pitch values from non-auto-correlated signal segments is set to 0 to perfectly align the energy and pitch time series.

In summary, the "energy and pitch extraction" module implements the following algorithm:

Algorithm 2 Energy and Pitch Extraction

For each tone unit:

Calculate the energy of consecutive signal segments with *energy and pitch window length*
Calculate the pitch of consecutive signal segments with *energy and pitch window length*. If pitch is undefined report 0

Store the tone unit energy and pitch sequences.

As a result, the algorithm associates each tone unit with two aligned time series of energy and pitch. Therefore, it defines a sequence of two-feature vectors over the signal, whose modulations can be analysed to detect infant cry.

2.4 Cluster Analysis

A first selection of the potential vectors referring to infant cry can be made by applying cluster analysis to groups of energy-pitch feature vectors. Indeed, an infant-cry signal segment corresponds to a group of energy-pitch feature vectors. Our "cluster analysis" module passes an analysis window over the signal while shifting it by some milliseconds (*analysis-window length* and *analysis-window shift* parameters). These parameters were optimised through cross validation (Section 3.2). The feature vectors falling within an analysis window are concatenated and then clustered. The module uses the K-means clustering algorithm [70]. This unsupervised model assigns the elements of a vector space to a fixed number of clusters without prior knowledge of the vector distribution. K-means uses an iterative process that optimally assigns the vectors to K clusters organised around K centroids. It assigns a vector to the nearest cluster based on its Euclidean distance from the centroid. In our workflow, K is an unknown parameter that is optimised for each tone unit. For each tone unit separately, our module iteratively uses K-means with different K values to find the optimal K value. Each clustering is evaluated through the Bayesian Information Criterion (BIC) under identical spherical Gaussian assumption. The highest BIC value corresponds to the optimal K value [71].

2.4.1 BIC selection criterion

The BIC selection criterion assumes that the optimal model among a set of candidate models (each corresponding to a different K value) is the one with the highest data likelihood, minus a penalty that increases with the number of parameters and data involved (to discourage overfitting). Mathematically, BIC is defined as

$$BIC(M_K) = ML(D|M_K) - \frac{NPar_{M_K}}{2} \cdot NData \quad (1)$$

Where M_K is the model using K clusters; $ML(D|M_K)$ is the maximum likelihood of the model; $NPar_{M_K}$ is the number of parameters involved in the model (i.e., the number of clusters plus the number of features of all cluster centroids); and $NData$ is the number of data clustered. The spherical Gaussian assumption is used to calculate the maximum likelihood by hypothesising that the data distributions around the centroids resemble independent normal distributions [71].

2.4.2 Clustering algorithm

The minimum number of clusters to test (*minimum clusters*) is an unknown workflow parameter to be optimised (Section 3.2). It represents the minimum number of different energy-pitch combinations that should be distinguishable within a tone unit. If infant-cry segments were included in the tone unit, they would naturally fall in the same cluster as infant-cry-like signals (e.g., machine beeping sounds). Adult speech and white noise would instead fall in other clusters. The cluster analysis process can be summarised as follows:

Algorithm 3 Cluster Analysis with Multi K-means

For each tone unit signal:

Set a signal window with *analysis-window length* at the tone unit begin
While the window falls withing the tone unit boundaries:

Concatenate the energy and pitch features falling in the window
Shift the analysis window of *analysis-window shift* milliseconds and define a new window

For K between *minimum clusters* and total number of analysis windows:

Execute K-Means on the concatenated features
Calculate the Bayesian Information Criterion (BIC) under identical spherical Gaussian assumption

Select the optimal K^* as the K value corresponding to the highest BIC.

The result of this algorithm is the labelling of signal segments with anonymous cluster indexes.

As an alternative to Multi K-means, we tested X-Means (an optimised Multi K-means algorithm) [72] and DBScan (a density-based algorithm) [73], but they achieved a lower performance in our workflow.

2.5 Infant Cry Cluster Identification

Our cluster analysis module uses feature vector concatenations within analysis windows. However, these vectors are small time series of energy-pitch vectors whose modulations can characterise infant cry. A model able to detect the modulations associated with infant cry could identify which cluster likely contains infant-cry samples and thus produce a training set out of this. This statistical task would not necessarily require a model trained on a large corpus.

2.5.1 Hidden Markov Models

Hidden Markov Models (HMMs) are the most common choice for acoustic modelling based on a few training samples [58, 74]. An HMM is made up of sequentially connected states $H = \{h_1, .., h_g\}$ (Figure 1-middle frame). Given a sequence of acoustic feature vectors X, an HMM estimates the conditional probability distribution $p(X|S)$ of X given a sequence (S) of state occurrences $S = s_1, s_2, .., s_T$, with each s_i belonging to H. A decoding algorithm [75] estimates $p(X|S^*)$, with S^* being the sequence of states that maximises $p(X|S)$. The HMM learns $p(X|S^*)$ based on the X vectors samples associated with the modelled phenomenon. For example, an HMM modelling a specific syllable is trained only on feature vectors of that syllable. In HMM-based syllable classifiers, one HMM is trained for each syllable [58]. When an unknown syllables' vectors are input to all HMMs, the one modelling that syllable will output the highest $p(X|S^*)$ value. If only one HMM is available, its output will likely be high when X corresponds to the modelled phenomenon. Unlike cluster analysis, HMMs explicitly model the temporal relations and evolution of the feature vectors.

2.5.2 Infant cry cluster identification algorithm

An HMM trained on infant-cry samples will more frequently return a higher score on infant-cry features than on white noise or adult speech features. Therefore it can be used to statistically assess which cluster likely contains infant cry. Our "infant cry cluster identification" module uses an HMM trained on 21 s of prominent infant cry (Section 2.1). As a pre-processing step, it transforms the concatenated acoustic features of the previous module into time series. It then applies the HMM to the time series of each cluster of a tone unit. The scores that overcome an HMM *high-likelihood* threshold, are averaged to approximate a cluster likelihood. If the cluster with the highest likelihood also overcomes a *minimum-likelihood* threshold, its time series are marked as *potential infant cry*. These thresholds were optimised through cross validation (Section 3.2). The other clusters' time series are instead marked as *probable non infant cry*. If the *minimum-likelihood* threshold is never overcome, the tone unit under analysis is marked as not containing infant cry and discarded from further processing. Such type of tone unit would not show a clear separation between infant-cry and non-infant-cry clusters and could confound the next workflow modules. Conducting the analysis by tone unit thus reduces the

entropy of the cluster identification operation by specialising and simplifying the task and enhancing the identification reliability [58, 76, 77].

The infant cry cluster identification process can be summarised as follows:

Algorithm 4 Infant Cry Cluster Identification

Load a Hidden Markov Model (HMM) pre-trained on prominent infant-cry time series
For each tone unit:

For each cluster c:

For each feature vector V in c:

Represent the feature vector V as a time series X
Calculate the HMM likelihood score L to X
If $L >$ *high-likelihood threshold*, record L

Calculate the average likelihood $\bar{L}(c)$ of c based on the recorded L scores

Select the optimal candidate cluster $c^* = argmax(\bar{L}(c))$
If $max(\bar{L}(c)) >$ *minimum-likelihood threshold*, classify the high-likelihood time series of c^* as *potential infant cry* and the time series in the other clusters as *probable non infant cry*
If $max(\bar{L}(c)) \leq$ *minimum-likelihood threshold* discard the tone unit.

The output of this algorithm is a set of time series across several tone units, each with *analysis-window length*, marked as either *potential infant cry* or *probable non infant cry*.

2.6 Infant Cry Detector

In automatic speech recognition, speaker-dependent systems generally achieve a higher performance than speaker-independent systems [57, 74]. Training an ASR on the specific acoustic characteristics of one speaker and the surrounding noise improves recognition accuracy. Similarly, our "infant cry detector training" module adapts a new model to the *potential infant cry* and *probable non infant* samples marked in the tone units by the previous module. Although this training set incompletely represents all infant-cry spectral characteristics, it corresponds to the reality of the recording conditions and is suited for operating within these conditions. Our module trains a Long Short Term Memory model (LSTM) using the *potential infant cry* time series, from all tone units, as positive cases (target output = 1), and the *probable non infant cry* time series as negative cases (target output = 0).

2.6.1 Long Short Term Memory model with self-attention layer

LSTMs are suited to build classifiers for observation vector time series [78]. They consist of one computational unit (Figure 1-lowest frame) that iteratively processes all vectors of an input time series. The computational unit comprises three *gates* that process one vector at a time together with information extracted from the previous vectors. All gates are one-layer artificial neural networks with the same number of output neurons (*hidden layer length*) and tanh or sigmoid activation functions. The unit receives an input vector (cell state, c_t) from the previous computational step, which stores a long-term memory representation of the data. This vector is first updated through the *forget* gate, which selects the elements to retain. The selected elements are updated by processing the unit's current input vector x_t through a sigmoid-activated neural network (*input* gate) and a tanh-activated neural network. The two network outputs are multiplied and then added to the filtered cell state to produce a new cell state. This updated cell state is passed to the next LSTM processing step. The output cell state is also re-used internally to the LSTM unit after tanh value scaling to [-1,1]. This scaled vector is multiplied with the output of another sigmoid-activated neural network (*output* gate) whose input is the unit's input vector x_t. This operation combines the previous long-term information summarised by the cell state with the short-term information learned by the *output* gate to produce the unit's output vector h_t. The last output vector h_T represents the final LSTM output. To produce a final 0/1 classification function, with 1 indicating infant cry and 0 non infant cry, we added two fully-connected layers to the LSTM. The first layer ("classification" layer) consists of a neural network with two output neurons that produce [1,0] for infant-cry time series and [0,1] for non-infant-cry time series. The second layer ("decision" layer) returns 0 or 1 depending on the highest score between the two classification neurons. It classifies the input time series as *infant cry* if the first neuron's score is higher than the second neuron's score.

To further enhance the LSTM performance, we added a self-attention layer [79] between the LSTM and the classification layers. Self-attention combines all input-sequence elements (i.e., the LSTM outputs, in our case) to determine which one has the highest importance (attention weight) to enhance the classification performance. Self-attention was introduced in deep learning to mimic cognitive attention because it relates one vector with all the previous and subsequent vectors. Therefore, it processes the LSTM output vectors forward and backwards in time while giving all vectors appropriate weights at each processing step. A self-attention layer produces a new time series, where each vector is a position-wise weighted combination of the other vectors. The weights simulate the "attention" that the surrounding vectors should be given when processing one vector. During the training process, the self-attention layer optimises three matrices, named query, keys, and values, that are used to produce the attention weights [79–81]. We used a self-attention layer to improve the detection of infant-cry segment boundaries. Detecting the onset of a cry signal indeed requires looking ahead in the time series to see if more

energetic and better-classifiable segments are present. Similarly, the terminal part of a cry segment requires information from the previous segments for better classification. Hereafter, we will refer to our combined LSTM/self-attention model as LSTM+A.

2.6.2 Infant cry detection algorithm

After the training session, the LSTM+A model is used to classify a signal segment as containing infant cry or non infant cry ("infant cry detector annotation" module). As the final post-processing step, another module joins the adjacent or overlapping infant-cry segments to produce continuous annotations ("consecutive segment merging" module). This module saves the annotations to a LAB-formatted file [82], which contains textual lines with start and end seconds and an associated comment, e.g., "12.5 15.9 Infant Cry". LAB is an easy-to-parse format that allows importing our workflow output into other speech analysis tools (e.g., WaveSurfer and Praat).

Using an LSTM+A model requires optimising several parameters such as (i) the neural network weight initialisation modality, (ii) the training algorithm to use, (iii) the learning loss function to measure training progress, (iv) the number of training samples after which the weights should be updated (training batch size), (v) the number of complete passes through the training data (training epochs), and (vi) the hidden layer length. After optimisation on the training set (Section 3.2) these parameters were fixed for all applications. In summary, the "infant cry detector" modules implement the following processing steps:

Algorithm 5 Infant Cry Detector

Collect all *potential infant cry* (*pic*) and *probable non infant cry* (*pnic*) time series across the tone units
Train a Long Short Term Memory model with self-attention (LSTM+A). Use the *pic* series as positive classification cases (output = 1) and the *pnic* series as negative cases (output = 0)
Set a window with *analysis-window length* over the audio signal start
While the window falls within over the signal boundaries:

- Extract the window-associated energy and pitch acoustic feature time series
- Classify the time series through LSTM+A
- Shift the analysis window of *analysis-window shift* milliseconds

Merge adjacent windows classified as containing infant cry
Use these concatenated windows' temporal boundaries to annotate continuous audio segments containing infant cry
Produce an annotation file in the LAB format to report the detected infant-cry segments.

The output of this module is the final annotation of the entire recording (in LAB format), with the indication of the audio segments containing prominent infant cry.

3 Results

3.1 Evaluation metrics

Nine study cases were used to test our workflow performance (Section 2.1), distributed in nursery (3 cases), sub-intensive (4 cases), and intensive (2 cases) neonatal care environments. The expert's annotations of prominent infant cry were used as the gold standard reference. The annotations covered 16 minutes of audio recordings and identified 3.3 minutes of prominent infant-cry audio, corresponding to 8,624,522 signal-samples over the 27,248,593 total signal-samples. We used signal-sample-wise classification to precisely calculate the matching between the expert's and the automatic annotations. The signal-samples that fell in a correctly detected infant-cry segment were considered *true positives* (TP) and *false positives* otherwise (FP). The signal-samples correctly classified as non infant cry were considered *true negatives* (TN) and *false negatives* (FN) otherwise. Based on these definitions, we used the following standard metrics for our workflow performance:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{(Precision + Recall)}$$

Accuracy measures the total portion of correctly classified signal-samples; precision is the fraction of correctly classified infant-cry signal samples; recall measures the model's sensitivity to classifying infant cry; F1 is the harmonic mean of precision and recall and indicates how balanced the workflow is between these two measurements. Finally, we used Cohen's kappa [83] to measure the agreement between the expert's annotations and the automatic classifications with respect to chance agreement. Interpretations of kappa by Landis and Koch [84] and Fleiss [85] were used to interpret the values.

3.2 Workflow optimisation and parameter selection

A set of 21 s of audio samples containing only prominent infant cry was used to train the HMM for infant-cry cluster identification. The HMM number of states was the principal unknown parameter of the model. Its optimal value was found through a cross validation, in which 80% of the input was used to train the model and 20% to test it, repeated 10 times. Eventually, the number

of states resulting in the highest average likelihood was selected, which was equal to 4 (Table 2). This session also estimated optimal values for the high- and medium-likelihood thresholds (Section 2.5).

A 10-fold cross validation was used to find the other optimal workflow parameters before assessing the performance on the entire data set (Section 3.3). This validation operation tested all combinations of the following parameters' ranges:

- *Tone unit window length* between 100 and 500 ms;
- *Energy and pitch window lengths* between 50 and 200 ms;
- *Analysis window length* between 10 and 300 ms and *shift* between 0 and 300 ms;
- *Minimum clusters* between 1 and 5;
- *LSTM hidden layer length* between 1 and 100;

The optimal parameter set is reported in Table 2. We also optimised the other LSTM parameters, i.e., sigmoid function for uniform weight initialisation, Adam optimiser [86] with cross-entropy loss function as the training procedure, batch size equal to 150, and 2 training epochs. Moreover, to reduce overfitting risk, we enabled a *dropout* neuron selection strategy [87] in the LSTM gates, which statistically excluded - with a 0.2 probability - each node and its weights from each training session.

The selected parameters are meant to be valid also for other workflow applications, although they can be recalculated through the tools provided with the workflow. In fact, they constitute general parameters that help the workflow detect segments on which the LSTM+A model self-trains to adapt to new operational conditions. The performance measurements clarify the advantages and limitations of this choice (Section 3.3). On the one hand, this makes the workflow ready to be re-used in new cases, also by other hospitals. On the other hand, it requires meeting specific working conditions to improve performance.

3.3 Performance

Performance measurements revealed the advantages and optimal operational conditions of our workflow (Table 3 and Figure 3-a-e). The optimal performance was reached in the sub-intensive care environment (82% accuracy, 80% F1), with a *substantial* agreement with the expert's annotations. Accuracy in the intensive and nursery care environments were averagely comparable (76.6% and 69.3% respectively). Information retrieval performance was higher in the nursery care (70% F1) than in the intensive care environments (50% F1). Our workflow had an overall *moderate* agreement with the expert's annotations across all cases, with a 76.4% accuracy and 70% F1. These results, compared to other systems [52, 62], indicate that it was reasonably good considering the SNR ranging between 2.4 and 20.

The nursery cases had the most extensive accuracy variation (26%, between 63.6 and 89.6%) due to high conditions' variability across the cases, e.g., heterogeneous infant densities, frequent adult speech, and interfering noise with

rich spectrum. Among the nursery care study cases, the lowest accuracy (63.6%, in Nursery-3) was measured with relatively low-noise conditions (16.9 SNR) but in a crowded environment with multiple infants crying and adults talking (Table 1). In sub-intensive care, accuracy had a 24.8% range, but was skewed towards ∼80%. In this environment, people talk loudly but less frequently. Adults have distinguishable energy and prosody profiles, the infant presence density is low, and the environmental noise falls within frequency bands (<1 kHz and >10 kHz) outside of the ranges of infant cry (1-10.5 kHz). This was the main reason for classification accuracy improvement of our workflow in this environment. The agreement with the expert's annotations was overall *substantial*, apart from study case Sub-intensive-4 that contained far more adult speech than infant cry. In this case, infant cry was a rare event; thus, the LSTM+A classification performance decreased. In the intensive care study cases, there was a large accuracy discrepancy between the study cases (56.8% vs 82.0%), although the noise levels were comparable (4.4 vs 5.05 SNR). However, the F1 measures (65.5% and 32.2%), and the fair/slight agreements with the expert, indicate that the performance was overall low in this environment. The performance was indeed influenced by a large amount of noise concentrated in the same frequency range of infant cry, which was sometimes indistinguishable from infant cry within a 100-200 ms audio segment (even for a human ear).

Due to heterogeneous noise sources, SNR had a broader range across the sub-intensive and nursery cases. In the intensive care cases, it was principally related to one source (i.e., monitoring machines) (Figure 3-f). However, the distribution of the performance measurements across the SNR values showed no evident correlation with noise level (Figure 4). This property is a significant consequence of using energy and pitch features for classification, which are particularly robust to noise level but may depend on noise type.

4 Discussion and Conclusions

We have presented a workflow to detect infant-cry audio, which is suited for populating a database for neonatal diagnoses and analyses. The advantage of our workflow is that it requires minimal training and uses a self-training strategy to improve classification performance and adapt to the specific environmental noise and infants present in the recording. This adaptive approach is innovative, considering that infant-cry databases are hardly accessible and infant-cry spectral characteristics depend on the mother's native language. The presented approach is cost-effective from the point of view of recording session organisation and realisation, since it works even with cellular phone recordings. Therefore, its realisation would be affordable for many hospitals. Our present solution has a large applicability range at the expense of a lower precision and accuracy in some operational conditions. It reached good-level performance in a sub-intensive care environment where it could be already operational. An 82% accuracy can be considered sufficient given the high level of noise (∼11.5 SNR) and the minimal training set used (21 s). In fact, the frequently cited supervised

system for infant-cry detection by Cohen and Lavner [52] gains between 70 and 80% detection accuracy with our SNR levels. With respect to this system, our workflow (i) uses syllabic-scale acoustic features - instead of phonetic-scale features (∼10 ms) - that are more robust to noise, (ii) is nearly unsupervised and thus does not require large annotated corpora for model training, (iii) can be easily re-used in new operational environments than those of the training set (thanks to the self-training strategy), (iv) is completely open-source, which is uncommon for this type of systems.

The higher performance in the sub-intensive care environment likely depends on the particular noise type, a low density of infants in the room, and a less frequent mutually-induced cry. In this environment, machine noise was concentrated in high- and low-frequency ranges outside of the infant-cry frequency range (1 kHz-10.5 kHz). In nursery care, the more significant infant density created interfering and degraded signals with complex and rich spectra, also overlapping with frequent adult speech. Therefore, it was difficult to detect prominent infant cry. In intensive care, the lower performance was mainly due to machine noise concentrated in the 1-10.5 kHz frequency band, which sometimes was indistinguishable from infant cry. Overall, these results indicate that our workflow was more influenced by noise type than by noise level. In summary, a real-world application scenario would require preventing the microphone from capturing machine beeping sounds and human speech, while positioning the microphone close to the infant. A directional microphone or an incubator-mounted microphone in a sub-intensive care environment would therefore represent the optimal condition for our workflow.

The near-unsupervised nature of our workflow makes it directly usable with new data or by other hospitals with their data. Open and direct re-usability was one of our main goals. Re-using our workflow on new data would not necessarily require re-training the models because our infant-cry detector rather depends on the self-trained LSTM+A model when processing new data and environments. Attention should be therefore paid to guarantee optimal working conditions through appropriate recording equipment and environment.

One planned enhancement to generalise our workflow is to reduce false positives by more precisely separating articulated machine noise from infant cry and discarding tone units that do not contain infant cry. This phase will require enhancing the "infant cry cluster identification" module by (i) extending the training set, (ii) using additional acoustic features, (iii) using feature transformation to improve input data quality and informativeness [39], and (iv) using more powerful classifiers. For example, the HMM could be substituted with Conditional Random Fields [88] to improve classification performance by modelling long-term temporal relations in the training data [89]. Moreover, adding more syllabic-scale or perceptually motivated spectral information (e.g., the Modulation Spectrogram [56]) could enhance the separation between infant cry, machine noise, and adult speech [90]. One alternative to infant cry cluster identification could be to use a pre-trained automatic speech recogniser based on syllabic-scale features [58] to build a preliminary filtering stage identifying the

signal segments potentially containing infant cry. A transfer-learning strategy could be used to re-adapt such ASR to this scope.

As a near-future application of our workflow, we will create an infant-cry database for the sub-intensive care environment of the "Dipartimento Materno Infantile" of the Azienda Ospedaliero-Universitaria Pisana to help neonatologists study pathological and normal cry and possibly conduct early diagnoses. We will use infant-specific equipment (a directional microphone or an incubator-mounted microphone) to optimise the workflow operational conditions by removing confounding noise. This installation will also require interfacing with ethical and privacy commissions to guarantee non-invasive and privacy-safe solutions. Such a system will be a permanent monitoring system generating infant-specific data flows for the entire neonatal care environment.

Supplementary information. The workflow code and related modules and models are entirely Java-based and available as an open source software on the GitHub at https://github.com/cybprojects65/DeepCry. The repository also includes the acoustic features of the extracted tone units.

Acknowledgments. The authors wish to thank all involved AOUP medical staff for their voluntary participation to this experiments.

Declarations

- Funding information: The paper is the result of a self-funded research initiative within the neonatology unit's activities of the Santa Chiara Hospital of Pisa, Italy.
- Conflict of interest/Competing interests: The authors declare no conflict of interest.
- Ethics approval: Not applicable.
- Code availability/Data availability: The workflow code and related modules and models are entirely Java-based and available as an open source software on the GitHub at https://github.com/cybprojects65/DeepCry. The source code contains some sample data from the training and test sets. Original full training and test data are property of AOUP and are only available after request to the authors. However, the repository includes the acoustic features of the extracted and analysed tone units.
- Authors' contributions: CRediT author statement:
  Gianpaolo Coro: Conceptualization, Methodology, Validation, Visualization, Writing, Software
  Serena Bardelli: Investigation, Supervision, Validation, Writing
  Armando Cuttano: Supervision, Resources, Validation
  Rosa T. Scaramuzzo: Validation
  Massimiliano Ciantelli: Validation

References

[1] World Health Organization: Newborns: improving survival and well-being. https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality (2020)

[2] World Health Organization: Every newborn: an action plan to end preventable deaths. World Health Organization. https://www.who.int/initiatives/every-newborn-action-plan (2014)

[3] Golub, H.L., Corwin, M.J.: Infant cry: a clue to diagnosis. Pediatrics **69**(2), 197–201 (1982)

[4] Messaoud, A., Tadj, C.: A cry-based babies identification system. In: International Conference on Image and Signal Processing, pp. 192–199 (2010). Springer

[5] Vempada, R.R., Kumar, B.S.A., Rao, K.S.: Characterization of infant cries using spectral and prosodic features. In: 2012 National Conference on Communications (NCC), pp. 1–5 (2012). IEEE

[6] Ntalampiras, S.: Audio pattern recognition of baby crying sound events. Journal of the Audio Engineering Society **63**(5), 358–369 (2015)

[7] Orlandi, S., Garcia, C.A.R., Bandini, A., Donzelli, G., Manfredi, C.: Application of pattern recognition techniques to the classification of full-term and preterm infant cry. Journal of Voice **30**(6), 656–663 (2016)

[8] Ainsworth, M.D.S., Blehar, M.C., Waters, E., Wall, S.N.: Patterns of Attachment: A Psychological Study of the Strange Situation. Psychology Press, London, England, United Kingdom (2015)

[9] Zeifman, D.M.: An ethological analysis of human infant crying: answering tinbergen's four questions. Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology **39**(4), 265–285 (2001)

[10] Furlow, F.B.: Human neonatal cry quality as an honest signal of fitness. Evolution and Human Behavior **18**(3), 175–193 (1997)

[11] Bornstein, M.H., Putnick, D.L., Rigo, P., Esposito, G., Swain, J.E., Suwalsky, J.T., Su, X., Du, X., Zhang, K., Cote, L.R., *et al.*: Neurobiology of culturally common maternal responses to infant cry. Proceedings of the National Academy of Sciences **114**(45), 9465–9473 (2017)

[12] Patil, H.A.: "cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry. In: Advances in Speech Recognition,

pp. 323–348. Springer, New York City, U.S.A. (2010)

[13] Liang, Y.-C., Wijaya, I., Yang, M.-T., Cuevas Juarez, J.R., Chang, H.-T.: Deep learning for infant cry recognition. International Journal of Environmental Research and Public Health **19**(10), 6311 (2022)

[14] Wasz-Hockert, O.: The infant cry: A spectrographic and auditory analysis. Clinics in developmental medicine, 1–42 (1968)

[15] Wasz-Höckert, O., Michelsson, K., Lind, J.: Twenty-five years of scandinavian cry research. In: Infant Crying, pp. 83–104. Springer, New York City, U.S.A. (1985)

[16] Johnston, C.C., Stevens, B., Craig, K.D., Grunau, R.V.: Developmental changes in pain expression in premature, full-term, two-and four-month-old infants. Pain **52**(2), 201–208 (1993)

[17] Mima, Y., Arakawa, K.: Cause estimation of younger babies' cries from the frequency analyses of the voice-classification of hunger, sleepiness, and discomfort. In: 2006 International Symposium on Intelligent Signal Processing and Communications, pp. 29–32 (2006). IEEE

[18] Benson, J.B., Haith, M.M.: Social and Emotional Development in Infancy and Early Childhood. Academic Press, Cambridge, Massachusetts, U.S.A. (2010)

[19] Bănică, I.-A., Cucu, H., Buzo, A., Burileanu, D., Burileanu, C.: Automatic methods for infant cry classification. In: 2016 International Conference on Communications (COMM), pp. 51–54 (2016). IEEE

[20] Chang, C.-Y., Chang, C.-W., Kathiravan, S., Lin, C., Chen, S.-T.: Dag-svm based infant cry classification system using sequential forward floating feature selection. Multidimensional Systems and Signal Processing **28**(3), 961–976 (2017)

[21] Lawford, H.L., Sazon, H., Richard, C., Robb, M.P., Bora, S.: Acoustic cry characteristics of infants as a marker of neurological dysfunction: A systematic review and meta-analysis. Pediatric Neurology (2021)

[22] Garcia, J.O., Garcia, C.R.: Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks. In: Proceedings of the International Joint Conference on Neural Networks, 2003., vol. 4, pp. 3140–3145 (2003). IEEE

[23] Reyes-Galaviz, O.F., Tirado, E.A., Reyes-Garcia, C.A.: Classification of infant crying to identify pathologies in recently born babies with anfis. In: International Conference on Computers for Handicapped Persons, pp.

408–415 (2004). Springer

[24] Galaviz, O.F.R., García, C.A.R.: Infant cry classification to identify hypo acoustics and asphyxia comparing an evolutionary-neural system with a neural network system. In: Mexican International Conference on Artificial Intelligence, pp. 949–958 (2005). Springer

[25] Zabidi, A., Mansor, W., Khuan, L.Y., Sahak, R., Rahman, F.: Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In: 2009 5th International Colloquium on Signal Processing & Its Applications, pp. 204–208 (2009). IEEE

[26] Zabidi, A., Mansor, W., Khuan, L.Y., Yassin, I.M., Sahak, R.: Classification of infant cries with hypothyroidism using multilayer perceptron neural network. In: 2009 IEEE International Conference on Signal and Image Processing Applications, pp. 246–251 (2009). IEEE

[27] Lenti Boero, D., Weber, G., Vigone, M.C., Lenti, C.: Crying abnormalities in congenital hypothyroidism: preliminary spectrographic study. Journal of child neurology **15**(9), 603–608 (2000)

[28] Wermke, K., Hauser, C., Komposch, G., Stellzig, A.: Spectral analysis of prespeech sounds (spontaneous cries) in infants with unilateral cleft lip and palate (uclp): a pilot study. The Cleft palate-craniofacial journal **39**(3), 285–294 (2002)

[29] Lederman, D., Zmora, E., Hauschildt, S., Stellzig-Eisenhauer, A., Wermke, K.: Classification of cries of infants with cleft-palate using parallel hidden markov models. Medical & biological engineering & computing **46**(10), 965–975 (2008)

[30] LaGasse, L.L., Neal, A.R., Lester, B.M.: Assessment of infant cry: acoustic cry analysis and parental perception. Mental retardation and developmental disabilities research reviews **11**(1), 83–93 (2005)

[31] Alaie, H.F., Abou-Abbas, L., Tadj, C.: Cry-based infant pathology classification using gmms. Speech communication **77**, 28–52 (2016)

[32] Esposito, G., Venuti, P.: Comparative analysis of crying in children with autism, developmental delays, and typical development. Focus on Autism and Other Developmental Disabilities **24**(4), 240–247 (2009)

[33] Esposito, G., Venuti, P.: Developmental changes in the fundamental frequency (f0) of infants' cries: a study of children with autism spectrum disorder. Early Child Development and Care **180**(8), 1093–1102 (2010)

[34] Esposito, G., Hiroi, N., Scattoni, M.L.: Cry, baby, cry: expression of distress

as a biomarker and modulator in autism spectrum disorder. International Journal of Neuropsychopharmacology **20**(6), 498–503 (2017)

[35] Orlandi, S., Manfredi, C., Bocchi, L., Scattoni, M.L.: Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2953–2956 (2012). IEEE

[36] Aucouturier, J.-J., Nonaka, Y., Katahira, K., Okanoya, K.: Segmentation of expiratory and inspiratory sounds in baby cry audio recordings using hidden markov models. The Journal of the Acoustical Society of America **130**(5), 2969–2977 (2011)

[37] Lederman, D., Cohen, A., Zmora, E., Wermke, K., Hauschildt, S., Stellzig-Eisenhauer, A.: On the use of hidden markov models in infants' cry classification. In: The 22nd Convention on Electrical and Electronics Engineers in Israel, 2002., pp. 350–352 (2002). IEEE

[38] Kheddache, Y., Tadj, C.: Newborn's pathological cry identification system. In: 2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA), pp. 1024–1029 (2012). IEEE

[39] Jeyaraman, S., Muthusamy, H., Khairunizam, W., Jeyaraman, S., Nadarajaw, T., Yaacob, S., Nisha, S.: A review: survey on automatic infant cry analysis and classification. Health and Technology **8**(5), 391–404 (2018)

[40] Cohen, R., Ruinskiy, D., Zickfeld, J., IJzerman, H., Lavner, Y.: Baby cry detection: deep learning and classical approaches. In: Development and Analysis of Deep Learning Architectures, pp. 171–196. Springer, New York City, U.S.A. (2020)

[41] Ji, C., Mudiyanselage, T.B., Gao, Y., Pan, Y.: A review of infant cry analysis and classification. EURASIP Journal on Audio, Speech, and Music Processing **2021**(1), 1–17 (2021)

[42] Saraswathy, J., Hariharan, M., Yaacob, S., Khairunizam, W.: Automatic classification of infant cry: A review. In: 2012 International Conference on Biomedical Engineering (ICoBE), pp. 543–548 (2012). IEEE

[43] Reyes-Galaviz, O.F., Cano-Ortiz, S.D., Reyes-García, C.A.: Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In: 2008 Seventh Mexican International Conference on Artificial Intelligence, pp. 330–335 (2008). IEEE

[44] Tuduce, R.I., Cucu, H., Burileanu, C.: Why is my baby crying? an in-depth analysis of paralinguistic features and classical machine learning algorithms for baby cry classification. In: 2018 41st International Conference on

Telecommunications and Signal Processing (TSP), pp. 1–4 (2018). IEEE

[45] Sun, Y., Kommers, D., Wang, W., Joshi, R., Shan, C., Tan, T., Aarts, R.M., van Pul, C., Andriessen, P., de With, P.H.: Automatic and continuous discomfort detection for premature infants in a nicu using video-based motion analysis. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5995–5999 (2019). IEEE

[46] Chittora, A., Patil, H.A.: Data collection of infant cries for research and analysis. Journal of Voice **31**(2), 252–15 (2017)

[47] Wermke, K., Teiser, J., Yovsi, E., Kohlenberg, P.J., Wermke, P., Robb, M., Keller, H., Lamm, B.: Fundamental frequency variation within neonatal crying: Does ambient language matter? Speech, Language and Hearing **19**(4), 211–217 (2016)

[48] Mampe, B., Friederici, A.D., Christophe, A., Wermke, K.: Newborns' cry melody is shaped by their native language. Current biology **19**(23), 1994–1997 (2009)

[49] Wermke, K., Ruan, Y., Feng, Y., Dobnig, D., Stephan, S., Wermke, P., Ma, L., Chang, H., Liu, Y., Hesse, V., *et al.*: Fundamental frequency variation in crying of mandarin and german neonates. Journal of Voice **31**(2), 255–25 (2017)

[50] Wermke, K., Robb, M.P., Schluter, P.J.: Melody complexity of infants' cry and non-cry vocalisations increases across the first six months. Scientific reports **11**(1), 1–11 (2021)

[51] Kheddache, Y., Tadj, C., *et al.*: Characterization of pathologic cries of newborns based on fundamental frequency estimation. Engineering **5**(10), 272 (2013)

[52] Cohen, R., Lavner, Y.: Infant cry analysis and detection. In: 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, pp. 1–5 (2012). IEEE

[53] Lavner, Y., Cohen, R., Ruinskiy, D., IJzerman, H.: Baby cry detection in domestic environment using deep learning. In: 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), pp. 1–5 (2016). IEEE

[54] Liu, L., Li, Y., Kuo, K.: Infant cry signal detection, pattern extraction and recognition. In: 2018 International Conference on Information and Computer Technologies (ICICT), pp. 159–163 (2018). IEEE

[55] García, J.O., García, C.A.R.: Acoustic features analysis for recognition of normal and hypoacustic infant cry based on neural networks. In: International Work-Conference on Artificial Neural Networks, pp. 615–622 (2003). Springer

[56] Greenberg, S., Kingsbury, B.E.: The modulation spectrogram: In pursuit of an invariant representation of speech. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3. Munich, Germany, pp. 1647–1650 (1997). IEEE

[57] Wu, S.-L., Kingsbury, E., Morgan, N., Greenberg, S.: Incorporating information from syllable-length time scales into automatic speech recognition. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 2, pp. 721–724 (1998). IEEE

[58] Coro, G., Massoli, F.V., Origlia, A., Cutugno, F.: Psycho-acoustics inspired automatic speech recognition. Computers & Electrical Engineering **93**, 107238 (2021)

[59] Coro, G., Walsh, M.B.: An intelligent and cost-effective remote underwater video device for fish size monitoring. Ecological Informatics **63**, 101311 (2021)

[60] Cutugno, F., D'Anna, L., Petrillo, M., Zovato, E.: Apa: Towards an automatic tool for prosodic analysis. In: Speech Prosody 2002, International Conference, pp. 231–234 (2002)

[61] D'Anna, L., Petrillo, M.: Sistemi automatici per la segmentazione in unità tonali. In: Atti delle XIII Giornate di Studio del Gruppo di Fonetica Sperimentale (GFS), pp. 285–290 (2003)

[62] Coro, G., Bardelli, S., Cuttano, A., Fossati, N.: Automatic detection of potentially ineffective verbal communication for training through simulation in neonatology. Education and Information Technologies, 1–23 (2022)

[63] Chittora, A., Patil, H.A.: Spectral analysis of infant cries and adult speech. International Journal of Speech Technology **19**(4), 841–856 (2016)

[64] Cutugno, F., Coro, G., Petrillo, M.: Multigranular scale speech recognizers: Technological and cognitive view. In: Congress of the Italian Association for Artificial Intelligence, pp. 327–330 (2005). Springer

[65] Cutugno, F., Leone, E., Ludusan, B., Origlia, A.: Investigating syllabic prominence with conditional random fields and latent-dynamic conditional random fields. In: Thirteenth Annual Conference of the International

Speech Communication Association, pp. 2402–2405 (2012)

[66] Osmani, A., Hamidi, M., Chibani, A.: Machine learning approach for infant cry interpretation. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 182–186 (2017). IEEE

[67] Ji, C., Xiao, X., Basodi, S., Pan, Y.: Deep learning for asphyxiated infant cry classification based on acoustic features and weighted prosodic features. In: 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 1233–1240 (2019). IEEE

[68] Matikolaie, F.S., Kheddache, Y., Tadj, C.: Automated newborn cry diagnostic system using machine learning approach. Biomedical Signal Processing and Control **73**, 103434 (2022)

[69] Boersma, P., *et al.*: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the Institute of Phonetic Sciences, vol. 17, pp. 97–110 (1993). Citeseer

[70] MacQueen, J., *et al.*: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967). Oakland, CA, USA

[71] Pelleg, D., Moore, A.: X-means: Extending k-means with efficient estimation of the number of clusters. In: In Proceedings of the 17th International Conf. on Machine Learning, pp. 727–734. Morgan Kaufmann, Burlington, Massachusetts, U.S.A. (2000)

[72] Pelleg, D., Moore, A.W., *et al.*: X-means: Extending k-means with efficient estimation of the number of clusters. In: Icml, vol. 1, pp. 727–734 (2000)

[73] Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X.: Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Transactions on Database Systems (TODS) **42**(3), 1–21 (2017)

[74] Huang, X., Acero, A., Hon, H.-W., Foreword By-Reddy, R.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice hall PTR, Hoboken, New Jersey, U.S.A. (2001)

[75] Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory **13**(2), 260–269 (1967)

[76] Zwicker, E., Terhardt, E., Paulus, E.: Automatic speech recognition using psychoacoustic models. The Journal of the Acoustical Society of America **65**(2), 487–498 (1979)

[77] Stern, R.M., Morgan, N.: Hearing is believing: Biologically inspired methods for robust automatic speech recognition. IEEE signal processing magazine **29**(6), 34–43 (2012)

[78] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[79] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[80] Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A Decomposable Attention Model for Natural Language Inference. arXiv (2016). https://doi.org/10.48550/ARXIV.1606.01933. https://arxiv.org/abs/1606.01933

[81] Karim, R.: Illustrated: Self-Attention. https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a (2019)

[82] WaveSurfer: Software Guide for L541. https://phonlab.sitehost.iu.edu/wsman157/wsman10.htm (2021)

[83] Cohen, J., *et al.*: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)

[84] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics, 159–174 (1977)

[85] Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)

[86] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[87] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)

[88] Wallach, H.M.: Conditional random fields: An introduction. Technical Reports (CIS), 22 (2004)

[89] Origlia, A., Cutugno, F., Galatà, V.: Continuous emotion recognition with phonetic syllables. Speech Communication **57**, 155–169 (2014)

[90] Baby, D., Hamme, H.V.: Investigating modulation spectrogram features

for deep neural network-based automatic speech recognition. In: Sixteenth Annual Conference of the International Speech Communication Association, pp. 2479–2483 (2015)

Tables

Table 1: Descriptions of the study cases and their recording conditions, duration, stress levels and noise levels, the most frequently recorded cry types, and the number of different infants and operators involved in the recordings.

Study case	Description	Conditions	Stress level	Most frequent cry type	SNR	Duration (s)	N. of infants	N. of operators
NURSERY 1	Infants during examination	People talking loudly, low noise level	HIGH	Examination	9.44	36	2	2
NURSERY 2	Infants in cradle before examination	People talking loudly, low noise level	HIGH	Physical needs	13.66	21	3	2
NURSERY 3	Infants crying in crowded environment	Crowded, several people talking, multiple infant crying	HIGH	Spontaneous	16.86	451	4	3
SUB-INTENSIVE 1	Stable infants crying	People talking loudly, low noise level	MEDIUM	Spontaneous	20.01	152	2	2
SUB-INTENSIVE 2	Hungry infants crying	Periodic machine noise, alarms	MEDIUM	Hunger	16.71	50	2	0
SUB-INTENSIVE 3	Pre-term infant after examination and other infants crying	People talking loudly, high noise level	HIGH	Examination	3.13	21	2	1
SUB-INTENSIVE 4	Hungry infants crying	People talking loudly, high noise level	HIGH	Hunger	6.04	53	1	4
INTENSIVE 1	Pre-term infant crying during care manoeuvres	People talking loudly, high noise level, periodic machine noise	HIGH	Examination	4.40	36	2	3
INTENSIVE 2	Pre-term infant crying inside an incubator	People talking quietly, high noise level, constant machine noise	MEDIUM	Spontaneous	5.05	160	2	2

Table 2: Optimal parameter values of our workflow.

Parameter	**Value**
Tone unit window length (ms)	500
Energy and pitch window lengths (ms)	100
Analysis window length (ms)	300
Analysis window shift (ms)	100
Minimum clusters	5
Hidden Markov Model state length	4
LSTM hidden layer length	3

Table 3: Performance measurements per environment and study case, with the highest-accuracy cases highlighted in bold. The total case with the highest-performance is highlighted in red. Cohen's kappa is reported with its interpretations according to Landis and Koch and Fleiss.

Environment	SNR	Accuracy	Precision	Recall	F1	Kappa	Kappa (Landis/Koch)	Kappa (Fleiss)
NURSERY 1	9.44	**89.6%**	95.2%	84.3%	89.4%	0.79	**Substantial**	**Excellent**
NURSERY 2	13.7	**85.9%**	49.1%	65.7%	56.1%	0.48	**Moderate**	**Good**
NURSERY 3	16.9	63.6%	49.0%	91.6%	63.8%	0.33	Fair	Marginal
SUB-INTENSIVE 1	20	**87.5%**	73.0%	84.5%	78.3%	0.70	**Substantial**	**Good**
SUB-INTENSIVE 2	16.7	**79.6%**	69.9%	97.7%	81.5%	0.60	**Substantial**	**Good**
SUB-INTENSIVE 3	3.13	**87.0%**	84.5%	91.7%	88.0%	0.74	**Substantial**	**Good**
SUB-INTENSIVE 4	6.04	62.7%	63.3%	93.4%	75.5%	0.08	Slight	Marginal
INTENSIVE 1	4.4	56.8%	54.2%	82.2%	65.3%	0.14	Slight	Marginal
INTENSIVE 2	5.05	**82.0%**	19.4%	95.0%	32.2%	0.27	Fair	Marginal
TOTAL	10.6	**76.4%**	58.3%	89.5%	70.6%	0.48	**Moderate**	**Good**
TOTAL-NURSERY	13.3	69.3%	54.3%	89.2%	67.5%	0.42	**Moderate**	**Good**
TOTAL-SUB-INTENSIVE	11.5	**82.0%**	70.4%	90.6%	79.2%	0.64	**Substantial**	**Good**
TOTAL-INTENSIVE	4.73	**76.6%**	36.2%	85.4%	50.8%	0.39	Fair	Marginal

Figures

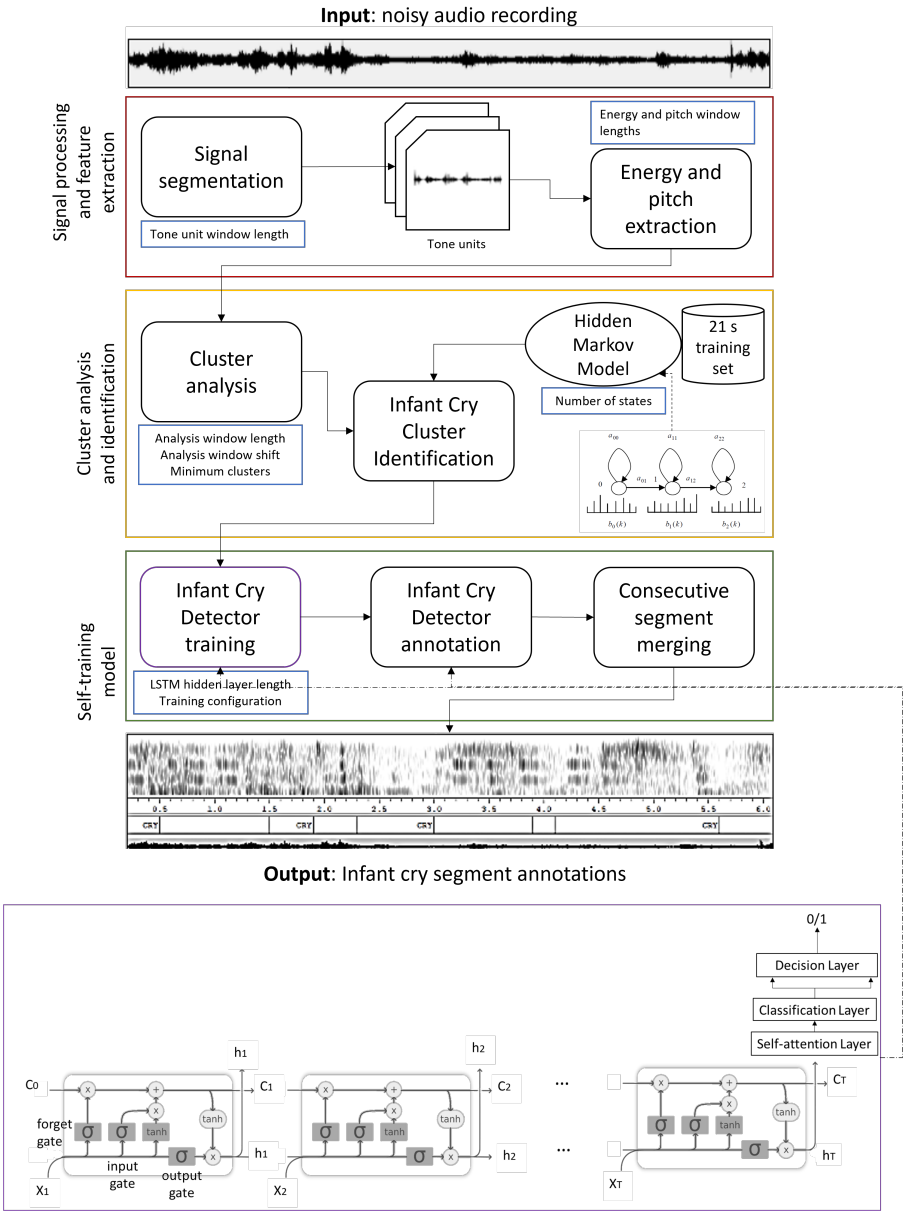

Fig. 1: Schema of the proposed infant cry detection workflow. The lowest frame shows the LSTM+A classification model.

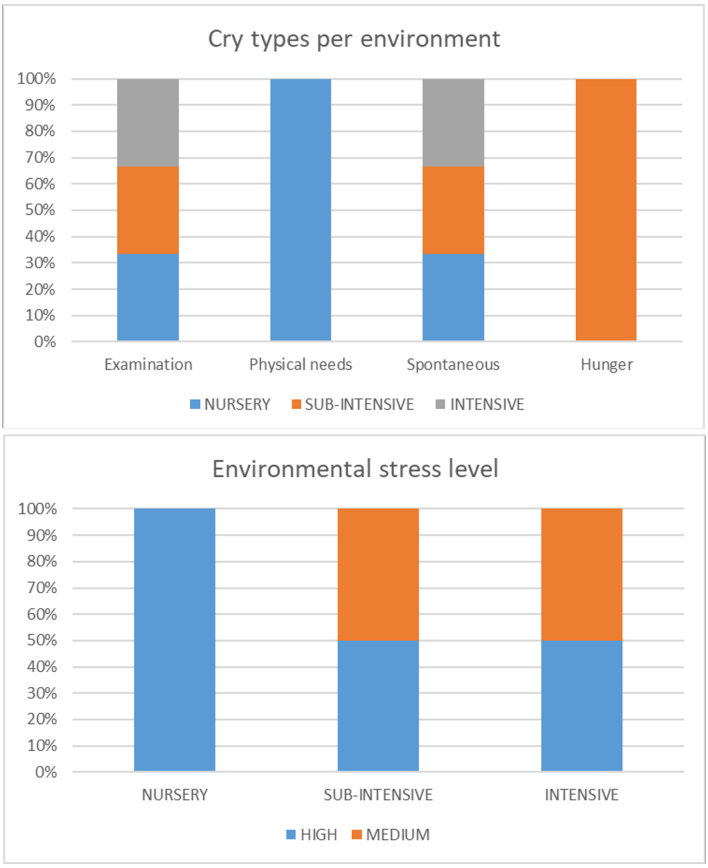

Fig. 2: Distributions of the most frequent infant cry types (upper chart) and environmental stress levels (lower chart) across the analysed neonatal care environments.

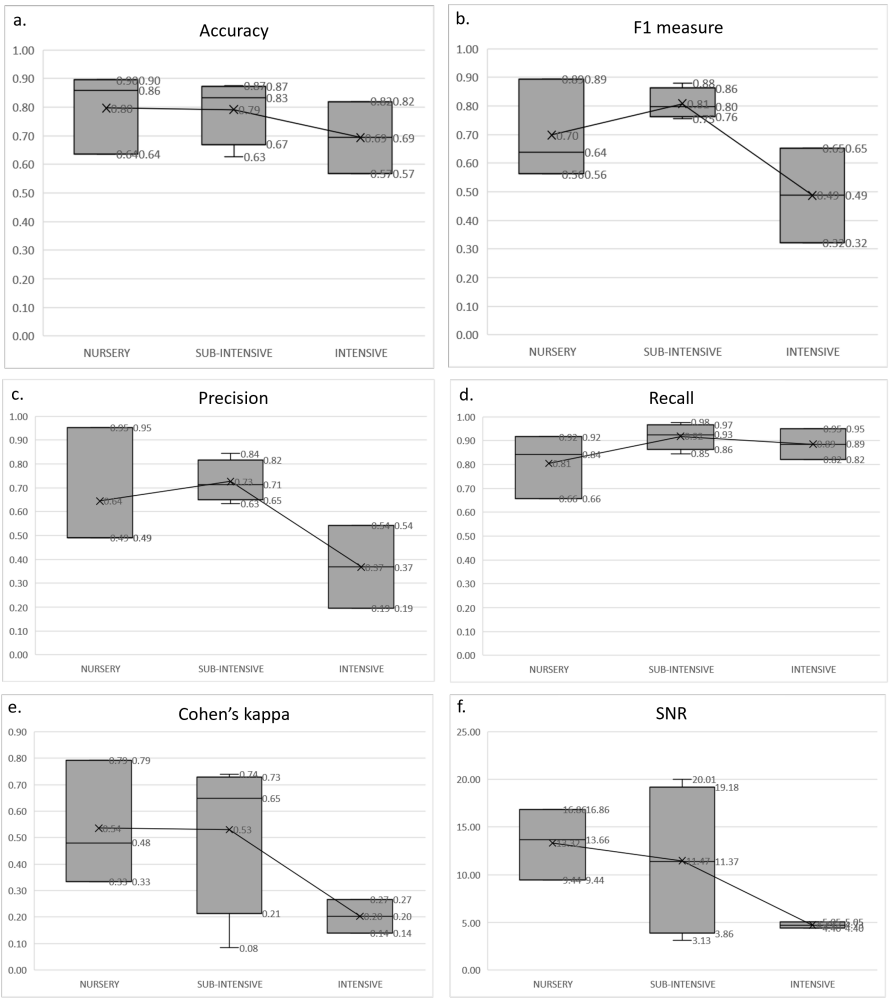

Fig. 3: Box plots showing performance measurements across the study-case environments: (a) accuracy on infant cry detection; (b) F1 measure, (c) precision, and (d) recall of the infant cry detector; (e) Cohen's kappa agreement with the expert's annotations; (f) signal-to-noise ratio (SNR)

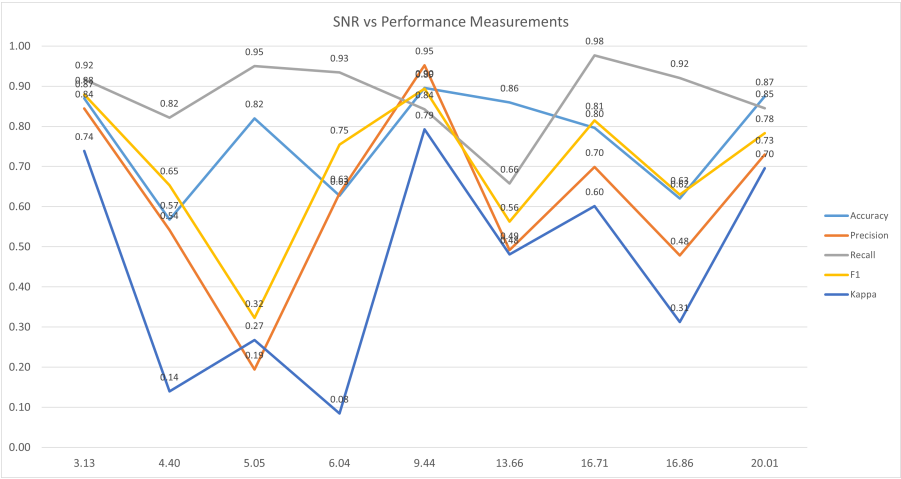

Fig. 4: A chart showing the relation between signal-to-noise ratio (SNR) (on the X axis) and different performance measurements (on the Y axis). The chart demonstrates that our model's performance does not depend on the noise level directly.