**On the automatic audio analysis and classification of cry for infant pain assessment**

**International Journal of Speech Technology manuscript No.**
(will be inserted by the editor)

# On the automatic audio analysis and classification of cry for infant pain assessment

Received: date / Accepted: date

**Abstract** The effectiveness of pain management relies on the choice and the correct use of suitable pain assessment tools. In the case of newborns, some of the most common tools are human-based and observational, thus affected by subjectivity and methodological problems. Therefore, in the last years there has been an increasing interest in developing an automatic machine-based pain assessment tool.

This research is a preliminary investigation towards the inclusion of a scoring system for the vocal expression of the infant into an automatic tool. To this aim we present a method to compute three correlated indicators which measure three distress-related features of the cry: duration, dysphonantion and fundamental frequency of the first cry. In particular, we propose a new method to measure the dysphonantion of the cry via spectral entropy analysis, resulting in an indicator that identifies three well separated levels of distress in the vocal expression. These levels provide a classification that is highly correlated with the human-based assessment of the cry.

**Keywords** Infant cry analysis · machine-based infant pain assessment tool · spectral entropy analysis

D. Ricossa
Dipartimento di Matematica "G. Peano", Università degli Studi di Torino, Via C. Alberto 10, Torino, Italy
E-mail: davide.ricossa@edu.unito.it

E. Baccaglini
MLW, Istituto Superiore Mario Boella, Via P. C. Boggio 61, Torino, Italy
E-mail: baccaglini@ismb.it

E. Di Nardo
Dipartimento di Matematica "G. Peano", Università degli Studi di Torino, Via C. Alberto 10, Torino, Italy
E-mail: elvira.dinardo@unito.it

E. Parodi
SC di Pediatria e Neonatologia, AO Ordine Mauriziano, Largo F. Turati 62, Torino, Italy
E-mail: emilia.parodi@unito.it

R. Scopigno
MLW, Istituto Superiore Mario Boella, Via P. C. Boggio 61, Torino, Italy
E-mail: scopigno@ismb.it

## 1 Introduction

Until the '80s, due to the lack of scientific studies, there was just a set of assumptions about infant pain, which resulted in a common undertreatment of it. Among these assumptions, the major one was that infants do not experience pain due to their neurological immaturity. This assumption was later proven to be incorrect [3]. Moreover, the number of painful events increases with the most immature infants. This fact, jointly with the awareness of the short- and long-time adverse sequelae of the exposure to repeated painful stimuli in early life [18], has made the infant pain assessment a real issue. Nowadays there are numerous neonatal pain scales, that use some observable indicators as surrogate of the patient's self-evaluation. An example is the Douleur Aiguë du Nouveau-né (DAN) scale [10], reported in Table 1. Thus, due to the observational nature of these scales, it is difficult to identify a peak of pain in an acute pain experience and a continuous assessment is not applicable for chronic pain. Besides these methodological problems, human-based tools can also be affected by subjectivity problems. For instance, Bellieni et al. [7] observed a significant difference between three groups of operators (O1, O2 and O3) using some of the most common tools to assess the pain of infants undergoing a routine heel prick procedure as follows: O1 scored after performing the actual heel prick, O2 scored as an observer who was free to watch the procedure closely, O3 recorded the procedure through a video camera and gave the score later by watching the video more than once if necessary. Because pain is subjective [20], a second degree of subjectivity

**Table 1** Douleur Aiguë du Nouveau-né

| Indicators | Score |
|---|---|
| **Facial expression:** *eye squeeze, brow bulge, nasolabial fold;* | |
| Calm | 0 |
| Snivels and alternates gentle eye opening and closing | 1 |
| Mild, intermittent with return to calm | 2 |
| Moderate | 3 |
| Very pronounced, continuous | 4 |
| **Limb movements:** *pedals, toe spread, legs tensed and pulled up, agitation of arms, withdrawal reaction;* | |
| Calm or gentle | 0 |
| Mild, intermittent with return to calm | 1 |
| Moderate | 2 |
| Very pronounced, continuous | 3 |
| **Vocal expression** | |
| No complaints | 0 |
| Moans briefly | 1 |
| Intermittent crying | 2 |
| Long-lasting crying, continuous howl | 3 |

is added and so the bias of a human observer could really compromise the reliability of the pain assessment process. This is why in the past several years there has been an increasing interest for an entirely machine-based pain assessment tool [33]: a way to monitor automatically the various pain indicators and evaluate them continuously and consistently with a minimum bias.

## 2 Background

### 2.1 Infant cry analysis

Crying is the earliest form of communication which constitutes the major part of the infant's vocalization: it is the way the newborn expresses his/her physical and emotional state and needs. Today's research in infant cry analysis was initiated by a team of Scandinavian researchers in the '60s who proposed spectrographic analysis [30] as one of the first approaches. Later, with the development of high-speed computer technology the study of cry has been subject to significant improvement. Over the years, the investigation of the main characteristics of infant cry both in time and in frequency domain has brought to light important insights related to the cry generation process and some models have been proposed [16, 13]. Nevertheless, our knowledge of cry generation is still limited. Therefore, nowadays infant cry is mainly studied from the processing [22]. Being the product of a human's vocal apparatus, although an immature one, it can be considered a particular case of human voice. Studying infant cry using speech signal processing/recognition techniques can thus be a promising approach. These techniques have led to identify some time-frequency patterns[1], or crying features, that are correlated with the context and, therefore, are considered meaningful. For instance, following a linguistic approach, in 1993 Xie et al. [31] first defined a set of 10 cry phonemes which provide the basis of the major part of the time-frequency pattern of variation in infant cry. Then, they analysed the correlation between the permanence time in each cry mode and the level of distress (LOD) perceived by the parents. It was observed [31] that amongst the phonemes, the dysphonation shows the most consistent positive correlation with the perceived LOD. An other example is the fundamental frequency, which is considered an outstanding characteristic [4]. Indeed, it has been observed that the first cry produced in response to an invasive pain stimulus displays a higher fundamental frequency and greater variability in the fundamental frequency during the cry episode [23].

On the other hand, cry interpretation is still a difficult task. In fact, different crying features give information about the LOD of the infant, rather than reflecting the exact reason for crying (e.g. hunger, discomfort, loneliness, pain, colic pain etc.) which is contextual. So, an observed cry helps to identify a set of possible causes or stimuli, but each of them has to be taken as uncertain and presumed without more information about the scene [8]. Over the last 30 years, several models for crying classification have been proposed, with various results (e.g. Hidden Markov [32] and Gaussian Mixture-Universal Background [6] models, Bayesian [5] and Random Forest [22] classifiers). Moreover, because of the absence of an actual ground truth for cry interpretation (quite often the context, e.g. "distressful" or "not distressful", is used in place of it), it is still unclear which one is the best performing approach (see Section 6). Some proposed classifications are strictly related to the processing technique used to analyse the cry, which sometimes relies on the use of a non-fully-accessible (non-free or experimental) software, thus making the comparison even more difficult.

### 2.2 Cry Analysis for Infant Pain Assessment

A preliminary step towards the inclusion of the vocal expression into a machine-based pain assessment tool is to understand if the acoustic features of a painful cry fit some kind of scoring system. To this end, in this paper we present a method to evaluate three indicators of some distress-related features of the cry. Then, we analyze the indicators' correlation when calculated for a dataset of infants subjected to procedural pain. Finally, we compare the results with the sample mode of the human-based assessment of the infant's vocal expression.

The rest of the paper is structured as follows. Section 3 describes the experimental setting and the dataset used in the experiment. Section 4 presents the pre-processing and the proposed method to calculate the indicators. Section 5 resports the experimental results, which are later discussed in Section 6.

[1] Though, some of these patterns have been just described verbally in literature, rather than defined numerically.

## 3 Design

### 3.1 Project

This paper is part of a preliminary investigation commissioned by AO Ordine Mauriziano di Torino (Italy) to Istituto Superiore Mario Boella. Given a small amount of data, the aim of the project was to explore machine-based methods for infant pain-assessment in order to study the feasibility of an automatic pain assessment tool. This is a mandatory step in order to ask for the ethical committee to approve a clinical trial involving a bigger cohort of subjects.

### 3.2 Subjects and data acquisition

This study is based on the analysis of a cohort of 31 healthy term infants who underwent heel lance for neonatal screening. The heel lance is a compulsory procedure (L. 104/92 - art. 6 [1]) that must be performed on every newborn between the first 48 and 72 hours after birth before the hospital discharge. During the procedure, the heel of the newborn is first lanced and then is gently squeezed in order to soak a blood sample into pre-printed collection cards. A monitoring system[2] recorded the reaction of each newborn in the course of the procedure. For every infant in the dataset a parent has given formal consent for the audio-video recording and every recommended measure has been taken so as to minimize the intervention-related stress and pain.

## 4 Methods

### 4.1 Pre-processing

After being extracted from the video, every audio track has been cut manually: for every infant, we considered the $30s$ after the painful stimulus [13]. In what follows we will refer to each of these $30s$-long signals as cry signal ($C_i$ in Fig. 1). First, we checked the dataset for audio tracks containing non-stationary background noises (e.g. speech signals, other crying infants etc...) that may interfere with the vocal expression and removed them. Then, in order to initialise the method, we inspected the original audio records for an interval of at least $1s$ of stationary background noise ($N_i$ in Fig. 1) located as close as possible to the occurrence of the painful stimulus. Because of bad recording environment, this has not been always possible and all those cry signals without an associated background noise have been discarded. The final dataset consists of 14 cry signals coupled with 14 background noise samples: $(C_i, N_i)$, with $i = 1, \dots 14$.

[2] An AXIS M1034-W network camera fixed to the wall through a mounting bracket and connected to a PC. Quality HDTV 720p/1 MP. The resolution varies from 1280x800 to 320x240 pixels and the frame rate is 25/30fps. The sampling rate of the audio signal is 16 kHz.

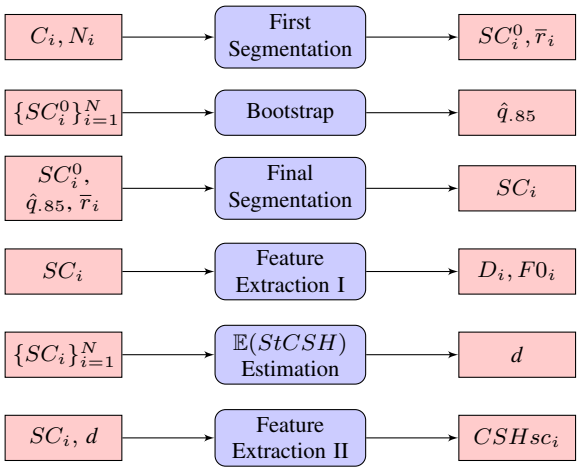


**Fig. 1** Block scheme of the proposed method. The $1^{\text{st}}$, $3^{\text{rd}}$, $4^{\text{th}}$ and $5^{\text{th}}$ block are intended to be embedded in a for loop over the elements of the dataset (i.e. $i = 1, \dots N$). $SC_i^0$ is the first segmentation of the signal given the $i^{\text{th}}$ couple cry-noise $(C_i, N_i)$; $\overline{r}$ is the lower confidence limit for the mean of the spectral entropy of $N_i$; $\hat{q}_{.85}$ is the upper basic bootstrap confidence limit for the $0.85$-quantile of the permanence time of spectral entropy of $C_i$ under the threshold $\overline{r}$; $SC_i$ is the final segmentation of the $i^{\text{th}}$ cry signal; $d$ is the lower 0.95-confidence limit for the mean of the random variable *StCSH*, defined in (1); $D_i, F0_i$ and $CSHsc_i$ are the proposed LOD indicators. In Section 4 we do not use the *i* indexes for the sake of simplicity.

These coupled data have been used as input to an R [24] script. The required audio analysis tools are part of the package 'seewave' [27] and 'tuneR' [19]. The procedure consists of six blocks (Fig. 1): the first three of them concur in the segmentation procedure and the last three blocks perform the feature extraction.

### 4.2 Segmentation

An automatic segmentation algorithm is a primary step towards infant cry analysis. In fact, in order to evaluate the features of a given cry signal, we need to first be able to detect every continuous interval of time in which a vocalization (i.e. a product of the infant's vocal apparatus different from the ihalation-related sounds) occurs. We will refer to these intervals as cry units, although the precise meaning of this term will be defined numerically at the end of this section.

We have treated this detection problem by considering the continuous spectral entropy (CSH) of the signal [28](Appendix). Given a couple of signals in the dataset, let us call *HN* the CSH of the stationary noise sample and *HC* the CSH of the cry. By considering the lower confidence limit for the mean of *HN* (let us call it $\overline{r}$), we have observed that (Fig. 2):

- The entropy of the inhalation related sounds is above $\overline{r}$;

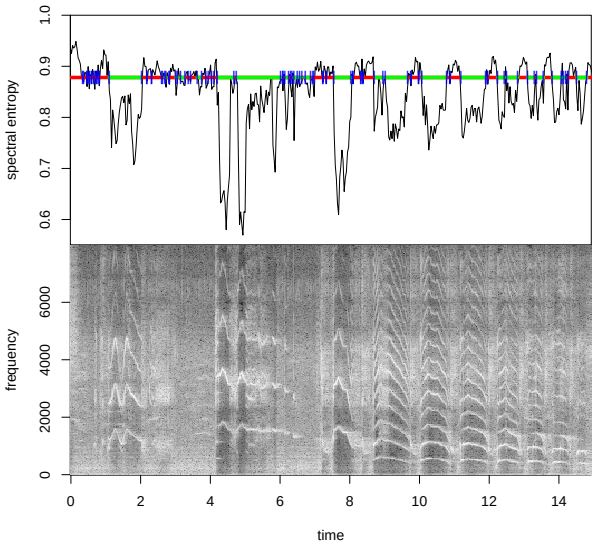


**Fig. 2** A comparison between *HC* and the spectrogram of the cry signal. The horizontal line is the corresponding noise threshold $\overline{r}$. The green intervals are classified as cry units in the first segmentation and the vertical bars | represent the respective cut-points.

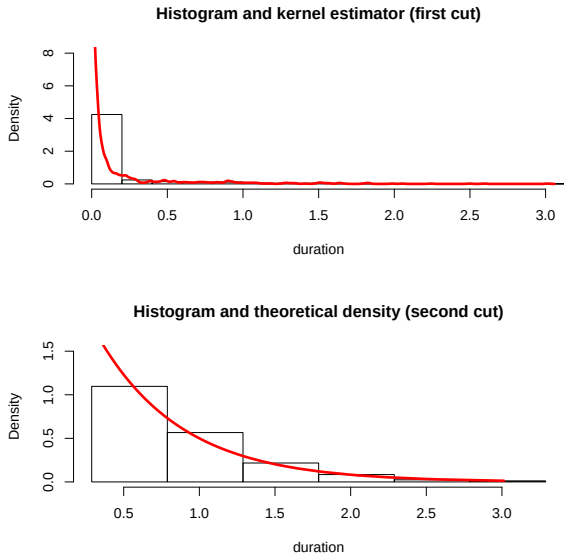


**Fig. 3** The histograms of the cry unit duration after the first (top) and the second (bottom) segmentation of the cry. The second cut considers only those intervals longer than $\hat{q}_{.85}$, that is the upper tail of the firs cut.

– When *HC* is lower than $\overline{r}$, it describes some U-shaped patterns of variation, and the most relevant of them corresponds to a voiced pattern in the spectrogram.

Thus, by considering all those time intervals such that $HC < \overline{r}$, we obtained a first segmentation of the cry signal ($SC_i^0$ in Fig. 1). Now, if performed over the whole dataset, this procedure returns 1352 time intervals with a mean duration of about $0.13s$. This value is far too small from a psychoacoustics point of view: indeed, the human perception of sounds starts to change dramatically under a duration threshold equal to $512ms$ [15]. To increase this value, we need to remove from this segmentation all those time intervals whose duration is not significantly long. From a stochastic point of view, this means to find some kind of boundary for the permanence time of *HC* under the threshold $\overline{r}$. Let us call $\tau$ this permanence time, which, in this method, represents also the random variable "duration of a cry unit". The first segmentation returns 1352 realizations (call them $\hat{\tau}$) of $\tau$. Looking at the empirical density function of $\hat{\tau}$, we noticed that the duration of more than the $90\%$ of the intervals is less than $0.5s$ (Fig. 3). Moreover, we observed that many of these brief time intervals correspond to the occurrence of some non-stationary or transient noise in the original signal, while the rare long-lasting ones correspond to exceptionally long cry units. In particular, for those records in which there is just a brief moan or no cry at all, this first cut fragments the signal in a set of very short segments, returning more noisy intervals than cry units. So, we obtained a second and final segmentation of the signal by removing from the first one all those intervals whose duration is less than the upper basic bootstrap confidence limit [9] for the $0.85$-quantile[3] of $\tau$, which determines a cutoff $\hat{q}_{.85}$ of about $288ms$.

In the resulting segmentation, those signals containing just brief moans become almost silent and all the most significant cry units are preserved in all the cry episodes. Moreover, among all the 166 intervals (with average duration of about $844ms$) identified in this way, just one contained pure noise and the others were exact cry units. The duration of these time intervals can be modeled (K-S statistic $D = 7.8 \cdot 10^{-2}$, $p$-value$= 0.2$) as $X + \hat{q}_{.85}$, where $X$ is an exponential with maximum likelihood estimated [12] rate parameter 1.8 (Fig. 3).

### 4.3 Feature Extraction

In this second part of the process, we use the segmented cry (*SC*) to measure some distress-related features of the original cry. In particular: duration, fundamental frequency and dysphonation.

Duration is undoubtedly an interesting characteristic of the cry signal. In fact, observe that the vocal expression item of the DAN scale [10] is actually an evaluation of the duration of the cry. Fundamental frequency, and in particular the fundamental frequency of the first cry after a painful stimulus is pointed as an oustanding feature in the literature [4, 23]. Finally, we would also give a measure of the dysphona-

[3] The more common 0.9-quantile results in a cutoff that exceeds the $0.5s$ empirical threshold of approximately $43ms$.

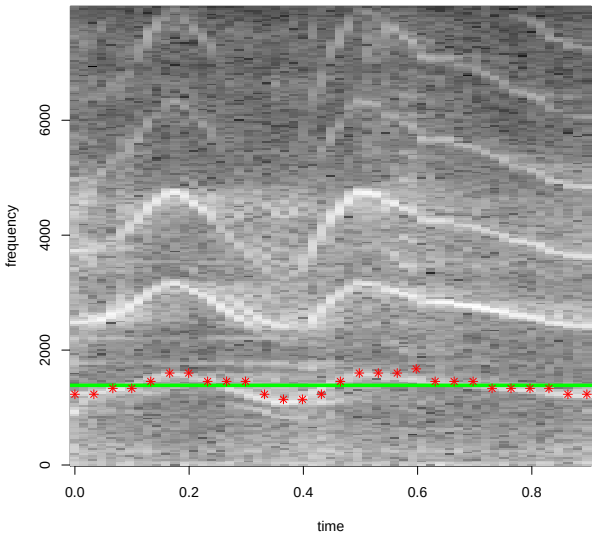


**Fig. 4** Spectrogram of the first significant cry unit, *FC*. Each $*$ represents the fundamental frequency in the respective window estimated by the corresponding high-frequency ripple in the cepstrogram of *FC*. The green line is the mean fundamental frequency of the first cry *F0*.

tion of the cry, because of its correlation with the perceived LOD [31].

As we said, some cry signals can result in a *SC* which is totally silent. We set to 0 all the feature-related scores for these signals. In what follows we denote with $M$ the number of cry units in the *SC* and suppose $M \geq 1$.

#### *4.3.1 Duration*

Let us call *D* the duration of the most significant cry units in the $30s$ after the painful stimulus. If we denote with **s** and **e** the $M$-dimensional vectors recording the cry units' starting and ending points respectively, then:

$$D = \sum_{i=1}^{M} (e_i - s_i).$$

So, *D* can be calculated just after the final segmentation (fourth block in Fig. 1).

#### *4.3.2 Fundamental frequency of the first cry*

The *SC* provides us not the exact first cry but the first significant cry unit after the heel lance, which is the signal in the time interval $[s_1, e_1]$. Let us call first cry (*FC*) this signal. Now, by considering the high-frequency ripples [21] in the cepstrogram (Appendix) of *FC* we can provide a series (with sample frequency equal to the ratio between the window length, 512 in our model, and the sample frequency) of estimates of the fundamental frequency. Thus, we will approximate the fundamental frequency with the sample mean of this series (Fig. 4): we will call *F0* this quantity, which is also calculated after the final segmentation (fourth block in Fig. 1).

#### *4.3.3 CSH score*

The dysphonation is characterized by an unstructured energy distribution [31] in the spectrogram. Now, the energy of a signal is as unstructured as it is dispersed on a larger range of frequencies [25] i.e. as its spectral entropy is near to 1. Clearly the CSH of the *SC* will never be equal to 1, because it is bounded by the threshold $\overline{r}$ (that depends on the noise with which each cry is coupled) that we used to obtain the first cut. Thus, in order to construct a scoring system equal for every cry signal, we classified a certain time window in a cry unit as "dysphonated" if, in that time window, its CSH is significantly close to the noise threshold $\overline{r}$. The specification of how much the distance between these two quantities has to be near to 0 is a thresholding problem involving the random variable (r.v.) *StCSH*, defined as:

$$StCSH = \overline{r} - H(\text{windowed cry unit}), \tag{1}$$

where $H$ denotes the spectral entropy (Appendix). Now, let us suppose that we have $N$ cry signals: $C_1, \dots C_N$ such that all of them give us a $SC_i$ with $i = 1, \dots N$ which is non totally silent. For $i = 1, \dots N$, let $\mathbf{s}_i$ and $\mathbf{e}_i$ be the $M_i$-dimensional vector containing the starting and ending points of the cry units in $SC_i$ respectively. We denote with $\overline{r}_1, \dots \overline{r}_N$ the respective noise thresholds in the CSH. Let us consider a Hanning's window

$$w(t) = \left[\frac{1}{2} + \frac{1}{2}\cos\left(\frac{2\pi}{a}t\right)\right] \mathbb{1}_{[-\frac{a}{2},\frac{a}{2}]}(t),$$

where $a > 0$ is arbitrary but fixed and $\mathbb{1}_{[-\frac{a}{2},\frac{a}{2}]}$ is the indicator function. By using the traslation operator $\vartheta_s : w(t) \mapsto \vartheta_s w(t) = w(t-s)$ to slide the window, then we obtain a realization

$$StCSH_{ij}(s) = \overline{r}_i - H(\vartheta_s w \; C_i \; \mathbb{1}_{[s_{ij}, e_{ij}]})$$

of the r.v. *StCSH* for every $s$ in the support of $\mathbb{1}_{[s_{ij},e_{ij}]}$, for every $j = 1, \dots M_i$ and for every $i = 1, \dots N$. The discrete nature of the signal (and the Heisenberg-Pauli-Weyl uncertainty inequality [17]) forces us to consider only a finite set of equispaced instants of time. So, given a cry unit $C_i \; \mathbb{1}_{[s_{ij},e_{ij}]}$, we considered only the the realizations of *StCSH* corresponding to the points $t_1, \dots t_{K_{ij}}$:

$$s_{ij} < t_1 = s_{ij} + \frac{a}{2} < t_2 = t_1 + a < \dots$$
$$\dots < t_{K_{ij}-1} < t_{K_{ij}} = s_{ij} + aK_{ij} \leq b$$

where $K_{ij}$ is the integer part of $(e_{ij} - s_{ij})/a$. We used those realizations to give an estimate $d$ of the lower 0.95-confidence limit for the mean of *StCSH*.

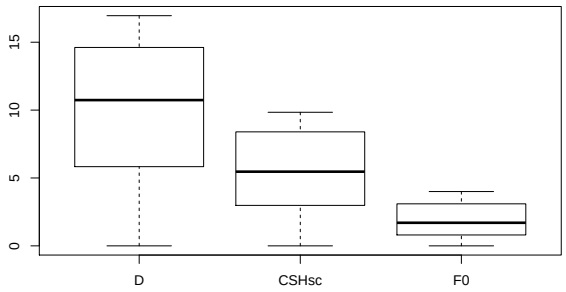


**Fig. 5** Box-plots of the proposed indicators *D*, *CSHsc* and *F0*.

Fixed the threshold $d$, we can assign a dysphonation score to every cry by counting all the time windows in which the distance between the CSH of the *SC* and the corresponding $\bar{r}$ is less than $d$ (the last two blocks in Fig. 1). We named CSH score (*CSHsc*) this quantity multiplied for the length of the window. More formally, we gave the following:

*Definition:* Given a cry signal $C_i$ and a segmentation $SC_i = \{[s_{ij}, e_{ij}]\}_{j=1,\dots M_i}$, chosen a Hanning's window of length $a$, we define:

$$CSHsc_i = \begin{cases} a \sum_{j=1}^{M_i} \sum_{h=1}^{K_{ij}} \mathbb{1}_{\{StCSH_{ij} < d\}}(t_h) & \text{if } M_i \geq 1, \\ 0 & \text{if } M_i = 0. \end{cases}$$

As we said, the final segmentation gives us 166 cry units for a total duration of approximately $137s$. Then, a Hanning's window with $a = 32ms$, by sliding along these cry units, produces about 4281 realizations (that we assume to be independent) of *StCSH*, with estimate $d \approx 77 \cdot 10^{-3}$ of the lower 0.95-confidence limit for its mean.

## 5 Results

In this section we analyze the output of the overall algorithm for the given dataset and in particular the three extracted features/scores: *D*, *CSHsc* and *F0* (box-plots in Fig. 5).

### 5.1 Variables' correlation

The couple *D-CSHsc* is correlated (estimated Pearson's correlation coefficient $\varrho = 0.84$, $p$-value$= 1.4 \cdot 10^{-4}$). This result is in agreement with the fact that both the duration and *CSHsc* have been constructed on the CSH-derived segmentation of the cry, resulting in an inner common dependency by the CSH of the signal and its noise threshold $\bar{r}$. A more interesting fact is that *F0* is correlated with both *D* ($\varrho = 0.74$, $p$-value$= 2.3 \cdot 10^{-3}$) and *CSHsc* ($\varrho = 0.82$, $p$-value$= 3.3 \cdot 10^{-4}$), and that this latter correlation is actually greater than the former one.

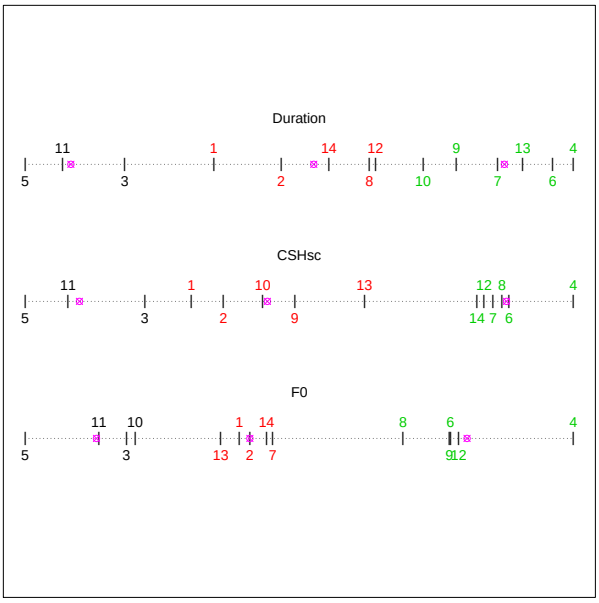


**Fig. 6** A representation of the 3-means clusters for the proposed indicators once standardized. The elements of the dataset are labeled with numbers whose color identifies the cluster. The clusters' centers are denoted by ⊗.

Now, because the intended use of these indicators is to construct a machine-based pain evaluation tool, a preliminary step is to check if their values highlight the presence of the three levels considered by the most common pain assessment scales (i.e. "Mild", "Moderate", "Severe"). Thus, we have turned the continuous scores in categorical data via unidimensional 3-means clustering [29] (Fig. 6). We have compared the resulting classifications by considering their contingency tables and performing the Pearson's chi-squared test on each of them. The results are reported in Table 2. The Spearman's $\rho$ rank correlation helps us to understand if the correlation between the continuous indicators is still present in the paired classifications. The null hypothesis of independence is rejected for all the couples, even though only *CSHsc* and *F0* are strongly correlated.

**Table 2** Pearson's $\chi^2$ test outcomes and Spearman's $\rho$ rank correlation estimaes

| Couple | $\chi^2$ | $p$-value | $\rho$ | $p$-value |
|---|---|---|---|---|
| *D-CSHsc* | 14.14 | $0.6 \cdot 10^{-2}$ | 0.54 | $4.5 \cdot 10^{-2}$ |
| *CSHsc-F0* | 12.13 | $1.6 \cdot 10^{-2}$ | 0.77 | $0.1 \cdot 10^{-2}$ |
| *D-F0* | 10.43 | $3.4 \cdot 10^{-2}$ | 0.5 | $6.5 \cdot 10^{-2}$ |

### 5.2 Human-based assessment

Because the expected receiver of the infant's cry is a human listener, we considered 6 human-based assessments of the

same dataset. The scorers were two near-graduate students ($S_a$ and $S_b$) in pediatric nursing, who repeated the assessment twice ($t_0$ and two months later, $t_1$) and two experienced pediatric nurses ($GS_a$ and $GS_b$). Each of them was provided with the audio-video record of the heel lance and was asked to assess the pain by using the DAN scale (Table 1). We only considered the score of the "Vocal expression" item. Moreover, in order to make the comparison feasible with the 3-mean cluster classifications (Fig. 6), we identified the first two scores of the "Vocal expression" item (i.e. both "No complaints" and "Moans briefly" are labelled with 1). At the beginning we grouped the scorers as follows:

- Group I: $S_a^{t_0}$, $S_a^{t_1}$;
- Group II: $S_b^{t_0}$, $S_b^{t_1}$;
- Group III: $GS_a$, $GS_b$.

To quantify the correlation of the evaluations, we considered the respective contingency tables for each group and performed the Pearson's chi-squared test. The null hypothesis of independence is rejected for all the couples ($p$-value$<0.04$). The reiterated evaluations are the most correlated ($\rho \approx 0.9, 0.89$ for $S_a$ and $S_b$ respectively), while $\rho \approx 0.66$ for the couple $GS_a$- $GS_b$.

We carried out a between-groups analysis in order to understand if it is suitable to use an experienced observer as gold standard for pain assessment. First we have turned both $GS_a$ and $GS_b$ in binary classificators by partitioning the possible outcomes in equal or strictly lower than 3. Then, one at a time, the resulting Boolean variables have been used as correct classification to construct one confusion matrix for every other observer in the dataset. We evaluated the performance of each classification with the ROC curve and in particular the area under the curve (AUC) [26]. The results of this analysis indicate that the assessment of the experienced observers ($0.67 \leq$ AUC $\leq 0.98$) is not meaningfully different from the scores of the inexperienced ones ($0.65 \leq$ AUC $\leq 0.95$).

We performed the same kind of analysis on the final DAN score (Table 1). This analysis required to fix a cutoff for the DAN scale in order to turn the assessments of the experienced observers in Boolean variables. We tried different values: the resulting AUC did not show any kind of improvement or worsening pattern in dependence by the choice of this cutoff.

### 5.3 Correlation between human-based assessment and output variables

The values of the AUC do not identify a difference in the performance of the three groups of human observers. So, to choose a scorer and use its assesment as correct classification seems quite arbitrary in this scenario. Therefore, we considered all the 6 human-based assessments as realizations of the variable "Human Scorer" aiming to compare the human-based assessment of the infant's vocal expression to the values of the output variables. After excluding all those cases (two in the dataset, Table 3) which are not unimodal, we considered the sample mode of these 6 human-based assessments of the vocal expression (*MoH*).

**Table 3** Human observers-Indicators

| | Score Frequecy | | | Indicators | | |
|---|---|---|---|---|---|---|
| Label | I | II | III | D | CSHsc | F0 |
| 1 | 0 | 4 | 2 | 5.83 | 2.98 | 1.56 |
| 2 | 3 | 2 | 1 | 7.92 | 3.56 | 1.64 |
| 3 | 4 | 2 | 0 | 3.08 | 2.15 | 0.74 |
| 4 | 0 | 0 | 6 | 16.96 | 9.84 | 4 |
| 5 | 6 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 6 | 16.31 | 8.69 | 3.1 |
| 7 | 0 | 2 | 4 | 14.62 | 8.4 | 1.81 |
| 8 | 0 | 3 | 3 | 10.64 | 8.56 | 2.76 |
| 9 | 0 | 3 | 3 | 13.33 | 4.84 | 3.1 |
| 10 | 0 | 4 | 2 | 12.31 | 4.26 | 0.8 |
| 11 | 4 | 2 | 0 | 1.15 | 0.77 | 0.54 |
| 12 | 0 | 0 | 6 | 10.83 | 8.24 | 3.16 |
| 13 | 0 | 5 | 1 | 15.38 | 6.09 | 1.43 |
| 14 | 0 | 0 | 6 | 9.39 | 8.11 | 1.76 |

Thus, we constructed the contingency tables between the 3-mean clusters of the proposed indicators and *MoH* for the unimodal cases. Again we performed the Pearson's chi-squared test on each of them and calculated the Spearman's $\rho$ rank correlation of every couple of classifications. The best result is given by the *CSHsc* classification for both Pearson's chi-squared test ($\chi^2 = 18.75$, $p$-value $= 8.8 \cdot 10^{-4}$) and Spearman's rank correlation ($\rho = 0.96$), while for the others the null hypothesis of Pearson's chi-squared test is not rejected ($p$-values $= 7 \cdot 10^{-2}$ for both *D* and *F0*). By interpreting the class of *MoH* as levels, we can try to fit them with a linear model of the form

$$MoH = \alpha_1 D + \alpha_2 CSHsc + \alpha_3 F0 + \epsilon. \quad (2)$$

Estimating the coefficients in (2) for the given dataset, *CSHsc* results to be not only the most relevant indicator ($\alpha_2 = 0.34$, std. error $= 0.9$, $t$-value $= 3.7$, $p$-value $= 5.5 \cdot 10^{-3}$), but also the only one with a coefficient significantly different from 0 ($p$-value $> 0.36$ for both *D* and *F0*).

### 5.4 The StCSH variable

Let us consider *StCSH* defined by (1). Because *CSH* is standardized by subtracting the corresponding noise threshold $\overline{r}$, it is reasonable to think that the lowest values of *StCSH* contain outliers. Thus, we considered the lower $5\%$-trimmed empirical distribution of *StCSH* for the given dataset. In particular, we observed that a beta r.v. with maximum likelihood estimated parameters $2.18$ and $24.24$ fits the observed values of *StCSH* (K-S statistic D $= 1.9 \cdot 10^{-2}$, $p$-value$= 9.5 \cdot 10^{-2}$, Fig. 7).

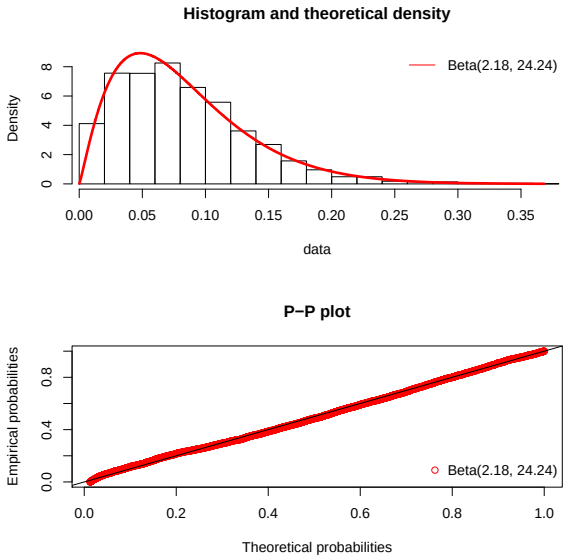


**Fig. 7** A comparison between the theoretical and empirical densities and cumulative density functions of the $5\%$-trimmed variable *StCSH*.

## 6 Discussion and conclusions

In the context of a feasibility study on the development of an automatic pain assessment tool, given a small dataset of cry-noise coupled signals, we have provided a basic method to compute three correlated LOD indicators.

All the estimates were calculated with non-parametric setting. Moreover, differently from other infant cry analysis procedures, the proposed method is entirely implemented on the R [24] free software, making all its steps completely specifiable by an accessible source code and therefore reproducible.

The preliminary nature of the study forced us to operate without a big dataset of infant cry, which is one of the novelty of the proposed method. In fact, as far as we know, the majority of the methods in the literature relies on the use of models whose parameters have to be trained (e.g. Hidden Markov [32], [2], Random Forest [22]), therefore requiring a big amount of data. Besides the peculiar context of this study, to assemble a database of infant cry is not an easy task as there are multiple difficult aspects to take into account to develop of such database (see [11] for details about the ideal characteristics of an infant cry corpus):

- Technical: install and use an audio acquisition framework in a neonatal unit, which is a noisy and uncontrolled environment;
- Legal: confidentiality, parental consensus and privacy;
- Standardization: once acquired, each cry has to be labeled by the context (e.g. distressfull, painful etc.). Moreover, the acoustic features have a great variability with age, weight and gestational age etc.

Each of these aspects becomes even more difficult in the case of the most immature and sick infants, whose pain has to be reliably assessed in order to be menaged. From this point of view, a method based on the use of a big dataset of cry signals could be very impractical, unless it leads to outstanding performances.

On the contrary, the proposed method relies only on the statistical properties of the spectral entropy of the cry. Observe that the use of windows, sliding across a small amount of signals, provided us with statistically significant samples to estimate these properties. So, we have built a method despite the small amout of data, which is applicable even with a dataset of just one cry coupled with a stationary noise. However, once we are able to collect a bigger dataset of both recordings and evaluators, our next objective will be the comparative study of the proposed method with the existing ones in terms of output and performances.

Among the proposed LOD indicators, *D* and *F0* are well-known: the duration and the fundamental frequency of the first cry are considered meaningful in the literature [4] and their computation is easy, once the cry segmentation is given. It is worth to say that these two features have been selected among a greater set of possible distress-related characteristics of the cry. Other indicators suggested in the literature (e.g. the variance in the cry units duration [5] or in the fundamental frequency during the cry [23], the root mean square of the time wave [8]) displayed poor correlation between the other extracted features (estimated Pearson's correlation coefficient $\varrho < 0.2$) and therefore their evaluation has been removed from the method.

The CSH score, as far as we know, provides a new way to measure the dysphonation of the cry by tracking the presence of unstructured energy in it. The "dysphonation phoneme" was introduced by Xie et al. [32] as state in a Hidden Markov Model (HMM). In particular, it was observed [31] that the permanence time in the dysphonation state shows the most consistent positive correlation with the perceived LOD. The dysphonation phoneme is characterized by "an unstructured energy distribution over all the frequency range, sometimes with a tendency of higher concentration over the middle to high (1-5 kHz) frequency range or an unstructured energy distribution imposing on or in between the barely distinguishable harmonics"[31]. This is the definition of just one out of the 10 states of the HMM proposed in [32], each of them is a phoneme analogously described by a time-frequency pattern of variation. The training of a such HMM requires a lot of data and effort. So, instead of applying this HMM to find an estimate of the permanence time in just one of the 10 states of the model, we preferred to track and measure the occurrence of the dysphonation phoneme in the segmented cry by monitoring the presence of unstructured energy in it. Besides this operative practicality, the *CSHsc* is higly correlated with both *D* and *F0*. Moreover, it can be

modeled as the permanence time of a process with known distribution under a threshold, making this new indicator particularly interesting for further investigations

From our results, despite the unquestionable relevance of the feature, the duration $D$ of the cry does not appear to significantly affect the evaluation of a human observer. This is in agreement with the recent literature addressing this feature as not sufficient to bias the human assessment. In particular, the 3-mean clusters of $D$ (see Fig. 6) does not evidence the presence of the levels described in the DAN scale. Only the first and the second are well-separated, thereby suggesting that the duration can only distinguish a "Mild" reaction from a "Moderate" one. Moreover, when the human scorers are not unanimous, six out nine times the scores are distributed among 2 and 3, including the two not-unimodal cases. Thus, more sophisticated features need to be investigated in order to capture the complexity of the human evaluation, which is one of the goals of this paper. Indeed, the fundamental frequency *F0* and the new score *CSHsc* provide better clusterizations. In particular, the 3-mean clusters of *CSHsc* are highly correlated with the sample mode of the human scorers, making it a candidate predictor in a hypothetical model for the human-based assessment of LOD of the infant's vocal expression.

Because of the significant correlation of the proposed indicators when considered as continuous variables, our purpose would be using them as predictors of the human-based assessment of the LOD via a general linear model (an ordered logit would be suitable, in our opinion). Clearly this validation process requires a bigger dataset of both recordings and evaluations, therefore more data are needed.

## Appendix

### Continuous Spectral Entropy

Let $X$ be a discrete random variable (d.r.v.) such that

$$\mathbb{P}(X = n) = \mathbb{P}(A_n) = p_n, \qquad n \in \mathbb{N}.$$

The *spectral entropy* of $X$ [25] is defined as:

$$H(X) = -\sum_{n\in\mathbb{N}} p_n \log p_n.$$

Given a discrete signal $\mathbf{y} \in \mathbb{C}^N$, let us denote with

$$\{F_n\mathbf{y}\}_{n=0,\dots N-1}$$

its discrete Fourer transform. Then, thanks to the discrete Plancherel's equality [14], we can define the d.r.v. $S_\mathbf{y}$ such that:

$$\mathbb{P}(S_\mathbf{y} = n) = \begin{cases} \frac{|F_n\mathbf{y}|^2}{N\|\mathbf{y}\|^2} & \text{if } n = 0,\dots N-1; \\ 0 & \text{if } n \geq N. \end{cases}$$

The *spectral entropy* of $\mathbf{y}\in \mathbb{C}^N$ is defined as:

$$H(S_\mathbf{y}) = -\sum_{n=0}^{N-1} \frac{|F_n\mathbf{y}|^2}{N\|\mathbf{y}\|^2} \log \frac{|F_n\mathbf{y}|^2}{N\|\mathbf{y}\|^2}.$$

By calculating $H$ for every element in the spectrogram of $\mathbf{y}$, i.e. for $S_{\mathbf{y}\mathbf{w}_0}, \dots S_{\mathbf{y}\mathbf{w}_M}$ where $\mathbf{w}_j$ is a discrete sliding window, we get the *continuos spectral entropy* of $\mathbf{y}$ [28].

### Cepstrogram

Given a discrete signal $\mathbf{y} \in \mathbb{C}^N$, let us define:

$$\mathbf{z} = (\log|F_0\mathbf{y}|, \dots \log|F_{N-1}\mathbf{y}|).$$

Then the *discrete cepstrum* [21] $C\mathbf{y} \in \mathbb{R}^N$ of $\mathbf{y}$ is defined as

$$C_k\mathbf{y} = \Re(F_k^{-1}\mathbf{z}) \qquad k = 0, \dots N-1.$$

In analogy with the spectrogram, the *cepstrogram* of $\mathbf{y}$ is defined as the cepstrum of the windowed signal: $\{C\mathbf{y}\mathbf{w}_j\}_{j=1,\dots M}$ where $\mathbf{w}_j$ is a discrete window.

## References


1. Legge 5 febbraio 1992, n. 104. URL http://www.gazzettaufficiale.it/eli/id/1992/02/17/092G0108/sg
2. Abou-Abbas, L., Montazeri, L., Gargour, C., Tadj, C.: On the use of emd for automatic newborn cry segmentation. In: Advances in Biomedical Engineering (ICABME), 2015 International Conference on, pp. 262–265. IEEE (2015)
3. Anand, K., Hickey, P.: Pain and its effects in the human neonate and fetus. N Engl j Med **317**(21), 1321–1329 (1987)
4. Baeck, H.E., Souza, M.N.: Study of acoustic features of newborn cries that correlate with the context. In: Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, vol. 3, pp. 2174–2177. IEEE (2001)
5. Baeck, H.E., Souza, M.N.: A bayesian classifier for baby's cry in pain and non-pain contexts. In: Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE, vol. 3, pp. 2944–2946. IEEE (2003)
6. Bănică, I., Cucu, H., Buzo, A., Burileanu, D., Burileanu, C.: Automatic methods for infant cry classification. In: Communications (COMM), 2016 International Conference on, pp. 51–54. IEEE (2016)
7. Bellieni, C.V., Cordelli, D.M., Caliani, C., Palazzi, C., Franci, N., Perrone, S., Bagnoli, F., Buonocore, G.: Inter-observer reliability of two pain scales for newborns. Early human development **83**(8), 549–552 (2007)
8. Bellieni, C.V., Sisto, R., Cordelli, D.M., Buonocore, G.: Cry features reflect pain intensity in term newborns: an alarm threshold. Pediatric research **55**(1), 142–146 (2004)
9. Canty, A., Ripley, B.: boot: Bootstrap r (s-plus) functions. R package version **1**(7) (2012)
10. Carbajal, R., Paupe, A., Hoenn, E., Lenclen, R., Olivier-Martin, M.: Dan: une échelle comportementale d'évaluation de la douleur aiguë du nouveau-né. Archives de pédiatrie **4**(7), 623–628 (1997)
11. Chittora, A., Patil, H.A.: Data collection of infant cries for research and analysis. Journal of Voice **31**(2), 252–e15 (2017)
12. Delignette-Muller, M.L., Dutang, C.: fitdistrplus: An R package for fitting distributions. Journal of Statistical Software **64**(4), 1–34 (2015). URL http://www.jstatsoft.org/v64/i04/
13. Facchini, A., Bellieni, C.V., Marchettini, N., Pulselli, F.M., Tiezzi, E.B.P.: Relating pain intensity of newborns to onset of nonlinear phenomena in cry recordings. Physics Letters A **338**(3), 332–337 (2005)

14. Gasquet, C., Witomski, P.: Fourier analysis and applications: filtering, numerical computation, wavelets, vol. 30. Springer Science & Business Media (2013)
15. Gelfand, S.A.: Hearing: An introduction to psychological and physiological acoustics. CRC Press (2016)
16. Golub, H.L., Corwin, M.J.: A physioacoustic model of the infant cry. In: Infant crying, pp. 59–82. Springer (1985)
17. Gröchenig, K.: Foundations of time-frequency analysis. Springer Science & Business Media (2013)
18. Hermann, C., Hohmeister, J., Demirakça, S., Zohsel, K., Flor, H.: Long-term alteration of pain sensitivity in school-aged children with early pain experiences. Pain **125**(3), 278–285 (2006)
19. Ligges, U., Krey, S., Mersmann, O., Schnackenberg, S.: tuneR: Analysis of music (2016). URL http://r-forge.r-project.org/projects/tuner/
20. Merskey, H., Bogduk, N.: Classification of chronic pain, iasp task force on taxonomy. Seattle, WA: International Association for the Study of Pain Press (Also available online at www. iasp-painorg) (1994)
21. Noll, M.A.: Cepstrum pitch determination. The journal of the acoustical society of America **41**(2), 293–309 (1967)
22. 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 (2015)
23. Porter, F.L., Miller, R.H., Marshall, R.E.: Neonatal pain cries: effect of circumcision on acoustic features and perceived urgency. Child development pp. 790–802 (1986)
24. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015). URL https://www.R-project.org/
25. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review **5**(1), 3–55 (2001)
26. Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T.: Rocr: visualizing classifier performance in r. Bioinformatics **21**(20), 7881 (2005). URL http://rocr.bioinf.mpi-sb.mpg.de
27. Sueur, J., Aubin, T., Simonis, C.: Seewave: a free modular tool for sound analysis and synthesis. Bioacoustics **18**, 213–226 (2008). URL http://isyeb.mnhn.fr/IMG/pdf/sueuretal_bioacoustics_2008.pdf
28. Toh, A.M., Togneri, R., Nordholm, S.: Spectral entropy as speech features for speech recognition. Proceedings of PEECS **1, 2005** (2005)
29. Wang, H., Song, M.: Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming. The R journal **3**(2), 29 (2011)
30. Wasz-Höckert, O., Michelsson, K., Lind, J.: Twenty-five years of scandinavian cry research. In: Infant crying, pp. 83–104. Springer (1985)
31. Xie, Q., Ward, R.K., Laszlo, C.A.: Determining normal infants' level-of-distress from cry sounds. In: Electrical and Computer Engineering, 1993. Canadian Conference on, pp. 1094–1096. IEEE (1993)
32. Xie, Q., Ward, R.K., Laszlo, C.A.: Automatic assessment of infants' levels-of-distress from the cry signals. IEEE transactions on speech and audio processing **4**(4), 253 (1996)
33. Zamzmi, G., Pai, C., Goldgof, D., Kasturi, R., Sun, Y., Ashmeade, T.: Machine-based multimodal pain assessment tool for infants: A review. arXiv preprint arXiv:1607.00331 (2016)