Artificial Intelligence for Human-Robot Interaction
AAAI Technical Report FS-17-01

Platform for Assessment and Monitoring of Infant Comfort

Aomar Osmani, Massinissa Hamidi
Laboratoire LIPN-UMR CNRS 7030
PRES Sorbonne Paris Cité, France
{ao,hamidi}@lipn.univ-paris13.fr

Abdelghani Chibani
Laboratoire LISSI
Université Paris Est Créteil, France
chibani@u-pec.fr

Abstract

Latest advances in Internet of Things (IoT) technologies, signal processing and machine learning can help in developing valuable solutions for supporting parents as well as caregivers in better taking care of the babies in particular. The pediatric studies that have been conducted since the sixties demonstrated that there is a universal language that hides behind the infant crying. This paper describes an IoT platform for ambient assisted living, which is dedicated to the automatic assessment of new born discomfort situations and providing assistive services to enhance infant as well as parent wellbeing. The present paper focuses on the description of a promising approach relying on the machine learning analysis of pre-cry signals to filter best signals. Obtained results show that learning rate of baby discomfort state is close to or better than results using cries only. This important result gives the opportunity to develop new baby monitors able to anticipate the infants needs and opens up perspectives to experiment and model mother/caregiver-infant like interactions in a human-robot interaction (HRI) context to enrich emotional support provided by robots to infants.

Introduction

When infant sleeps, especially during the night, the whole family members appreciate the serenity of that moment. This situation can change very quickly for many reasons: illness, hunger, intestinal pain, colics or discomfort. In many cases, parents are frustrated and complain about their inability to calm their infant despite all the actions they are taking to this end (Scott and Richards 1990). Studies reported a high prevalence of bedtime problems and frequent night wakings (Hiscock and Wake 2001) which corresponds to more than 40 million newborn infants that are affected worldwide each year on the regular basis of 145 million births according to the 2015 world population data sheet published by the Population Reference Bureau.

Crying is the instinctive way for an infant to express his needs as well as his internal state (Lieberman et al. 1971). It is a main vector and a universal language, through which the newborn can convey his reactions, desires and intentions, despite the poverty of this language in terms of information (Wasz-Hockert et al. 1963; Wasz-Höckert, Michelsson, and Lind 1985). For instance, all newborn infants cry to express the needs for food because of hunger or to sleeping due to fatigue. Moreover, they also cry with mostly the same way to react to discomforting feelings such as dirty diaper, feeling too cold or too hot, etc. (Lieberman et al. 1971; Wermke et al. 2002). According to pediatricians, in most cases, crying is related to discomfort and affect up to 30% of infant (Cook et al. 2012). Unlike adults, infants do not have the ability to soothe themselves and express their discomfort through cryings.

From pediatrics domain perspective, it is widely recognized that we can distinguish among cry types, and many studies started to focus on this aspect since the sixties (Wasz-Hockert et al. 1963). For example, in (Tsukamoto and Tohkura 1990), authors were interested in three categories of cry, namely, the hunger cry, the call cry and the anger cry. In (Chang et al. 2016), attention was drawn toward the automatic classification of infant cries into hunger, pain as well as sleepness categories while in (Petroni et al. 1995), authors were interested in categorizing anger, pain and fear cries. The approach used by all these studies is to analyze audio signals corresponding mainly to crying episodes that are generally longer and more difficult to soothe. However, more recently, some studies attempt to understand how infant behavioral patterns as well as specific vocalizations that precedes the onset of the longer and unsoothable crying episods can give some valuable indications about the internal state of the infant (Dunstan 2009). Actually, in addition to vocalizations, behavioral patterns such as movements are sometimes dismissed as unintentional or purposeless however, studies, for example (Weggemann et al. 1987), suggest that such movements are significantly related to the internal state of the infant and convey valuable cues that could be also exploited to detect discomfort situations.

Developing new machine learning approach which is able to analyze these infant vocalizations, or what we can call infant pre-cries, in order to recognize accurately the crying reason is the main challenge addressed in the present work. This paper shows clearly how the selected features can help in recognizing accurately the discomfort-related cries and identifying the reasons behind. This approach can be enhanced in a second time in order to help in refine crying analysis and provide pre-diagnostics related to pain expressed through cryings for infant health monitoring.

The proposed cry recognition mechanism is integrated in

an ambient assisted living (AAL) platform, which when it recognizes a discomfort-related cry, it can trigger automatically some actions to soothe the baby or notify parents or caregivers and suggest for them solutions if the infant crying persist. Amongst the actions that can be automatically triggered, thanks to the current advances in social robotics and IoT technologies: playing a song or a lullaby that is appreciated by the infant in that context, start to emit, adaptively, lighting effects or moreover. If the discomfort cry is due to sleep disturbance, a light spot can be switched on with an adequate color and luminosity can be gradually adapted until the infant fall asleep again. According to some research (Varendi et al. 1998), the design of the AAL platform is enhanced in order to include also actuators that can spread some synthetic odors that are recognized to calm the infant.

The remainder of the paper is organized as follows: Section 2 presents an overview of infant cry and comfort studies. Then, after reflecting the novelty of the proposed approach and its relevance to HRI, Section 4 deals with our global platform, it presents the overall architecture which is then detailed in Section 5 from the crying signal acquisition until selection of an appropriate action to reduce infant discomfort including response model. To have a clear picture of the process of data acquisition and cry classification, Section 6 gives an overview of the results obtained on a real world dataset. Section 7 concludes the paper.

Related work

The main problem with infant cry is to establish the right diagnostic and to understand the cry causes. Numerous studies have focused on the one hand on the psychological and development aspects underlying newborn infants sleep and cry problems and on the other hand on pathological aspect (Scott and Richards 1990; St James-Roberts 2007). Different characterizations or patterns have been provided in studies for infant sleep and crying problems as well as factors influencing these problems; for example, in (Sadeh 1994), healthy infants aged 9-24 months were included to be assessed for sleep disturbances if they were solely subject to 3-8 wakings each night or having prolonged night wakings. In the other hand, (Hiscock and Wake 2001) characterization includes infant sleeping in the parent's bed, being nursed to sleep, taking longer to fall asleep, waking more often and for longer periods overnight, and taking shorter naps. (St James-Roberts 2007), for his part, states that not all babies know how to put themselves to sleep and how to resettle after night waking.

Even if the first industrial product to deal with infant cry is the first "babyphone" proposed by *Zenith's Radio Nurse* in 1937 —it encompasses a transmitter to be plugged in by the child's crib and a receiver, to be located alongside the parents—, the categorization of crying types began in the sixties. In (Fraiberg 1950), different factors were summarized as being associated with night waking. These include children's anxieties and separation fears, but a common claim is that night waking is a learned response, usually to parental over-attention. According to (Wasz-Hockert et al. 1963), it is known that there is a strong correlation between baby cry and his specific needs. Several works have refined this discovery (Wermke et al. 2002; Lieberman et al. 1971) and some formal interpretations are proposed (Bănică et al. 2016). Technically, the most successful solution is probably the one which is proposed in (Chang et al. 2016) but the existing baby monitor (Rincon et al. 2013) and application (Chang, Hsiao, and Chen 2015) use only baby cry to identify baby's state. More recently, an interesting and deep work is proposed in (Dunstan 2009). The author found that crying is the final stage of the expression of need, when baby is upset after having tried to reduce the discomfort by himself. Several pre-cry signals are produced before crying, it is a set of automatic reflexes including phonetic sounds, movements, back-arching, knees flexion, and so on (Douglas and Hiscock 2010). In (Dunstan 2009), a set of primitive vocalizations are defined as a universal set of automatic reflexes that newborn babies make. In this paper, we propose to improve the discomfort learning rate by analyzing pre-cry period.

Proposed Approach and relevance to HRI

Infant manifestation of high level-of-distress, *i.e.* intense cries, require parents to be notified in order to come and soothe the cause of this manifestation. Consider the simple scenario when infant wake up during the night, typically, the waking is followed in a large majority of cases by intense, prolonged and difficult-to-soothe cries. These cries are, however, preceeded by low level-of-distress vocalizations that can be thought of as signs by which the baby is trying to soothe himself or get the attention of the caregiver. In fact, studies show that these wakings are caused essentially by discomfort and a large number of infant have the ability to resettle themselves back to sleep with little help (St James-Roberts 2007).

It is within this short, but important, timeframe, between the first signs of discomfort-related vocalizations and the onset of the more intense cries, that our value proposition lies. In fact, analyzing pre-cry cues opens up a wide variety of work that can be performed in the interaction aspect of systems such the one we are proposing in this paper and simulate mother/caregiver-infant like interactions via the different actuators and the response selection module (see Section 3). Modelisation of such human-robot interactions would be difficult if we rely only on the analysis of high level-of-distress cries which are more difficult to soothe and require parents intervention.

System Overview

The main functions of the AAL platform, which can be embodied by using a social robot and baby furnitures, consist in cry analysis and reduction of infant discomfort. It also provides other capabilities including sleep phases tracking, detecton of strong agitation that is caused mainly by overstimutlation in noisy environments. It also gives indications

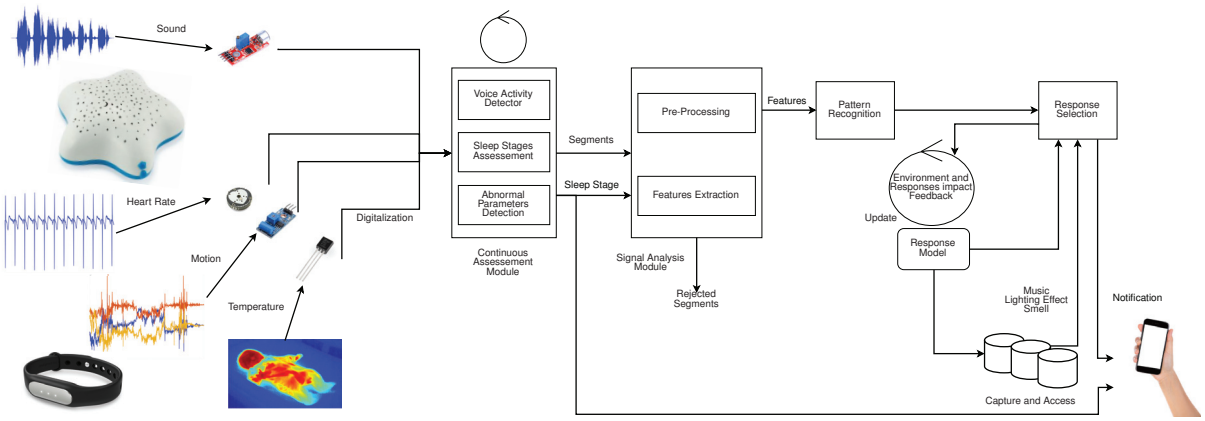


Figure 1: Architecture of the proposed smart baby monitor

about vital parameters (body temperature, heart rate, oxygen saturation, body movements, respiration rythm). It includes sound detection with a microphone, speakers, light emitter source, odor diffuser and sensors for local environmental parameters (temperature, humidity, noise, air quality, brightness). Figure 1 gives an overview of the general architecture of the AAL platform.

Continuous Assessment Module

The personal assistant will include, in particular, a voice activity detector (VAD) (Sohn, Kim, and Sung 1999) which will signal any vocal activity emanating from parents or from the babies themselves. This component will eliminate a lot of false alarms which could be caused by various surrounding noise. In conjunction with the sleep stages assessment and abnormal parameters detection components, the AAL will build a rich representation of its surrounding; It will deduce, for example, the presence of a parent alongside the baby's cradle or if the baby is sleeping, playing or trying to express a desire, etc. Using this representation, the system will adapt its behavior.

Signal Analysis Module

After the VAD has confirmed the occurrence of any vocal activity, the signal analysis module will begin processing the signals in order to extract relevant segments that correspond to vocalizations of the infant (see pre-cry units detection Section 5). These segments are then processed frame by frame during the preprocessing step, in order to extract, in a second time, a set of features (Section 6) on the basis of which, the prediction of the state of the infant will be made.

Cry Recognition

The function of recognizing the current situation of the infant is based on the signal's fundamental features (temporal, spectral, prosodic and cepstral features). Typically, the output of this step will be one of the crying classes that were enumerated by (Dunstan 2009). The author discovered that at a certain point of time, before the onset of crying, babies use universal phonetic sounds summarized in five classes as follows: (1) *Eairh* which stands for flatulence or the accumulation of gas in the alimentary canal, (2) *Eh* for eructation, burping *i.e.* the release of gas from the digestive tract through the mouth, (3) *Heh* indicating discomfort (cold, hot, wet diaper, change position, etc.), (4) *Neh* which indicates that baby is hungry, and finally (5) *Owh* which refers to tiredness. These sounds will be exploited in order to improve the overall situation recognition process of our system. Our experiments are done according to this classification (In the rest of the paper, we refer to cry to indicate cry, pre-cry or phonetic sound).

Response Selection

The response selection module implements a retroaction loop, which is responsible of taking decision and providing a solution towards soothing the infant and improving his wellbeing accordingly. These decisions range from displaying lighting effects, music, lullabies, parents' pre-recorded messages as well as different smells or an appropriate combination of these actions.

The main issue at this stage is whereas infants express their internal state by a set of universal sounds and movement patterns which are identical across all situations, it remains that each infant react differently to soothing. Therefore, the system has to come up with an adaptative policy which will react specifically. In other words, the system will evaluate the impact of the solution it provided according to the infant state; getting, for example, an important reward each time it guesses correctly the infant's needs through cryings and successfully soothing the infant or resettling him back to sleep.

Notification and Reporting

During all these steps, notifications are sent to parents when it is necessary through a mobile application. The application will handle discomfort notifications as well as notifications that require parental intervention and those which correspond to the variations of the vital parameters.

In addition, parents will have access to a detailed report of their infant's day through an application. In the same

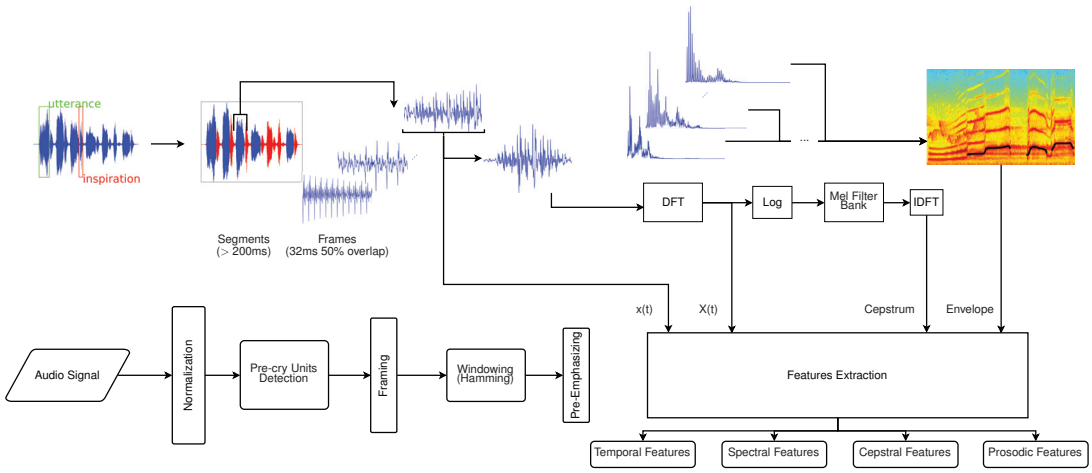


Figure 2: Block diagram of data preprocessing and features extraction

style as capture and access systems (Chang et al. 2016; Rincon et al. 2013), the application will summarize interest points that parents can navigate through in the form of a time-line. These interest points represent moments of the day when important variations of some biophysiological parameters were detected by the system.

Implementation Details

Data Preprocessing

Temporal, spectral, cepstral and prosodic representations are exploited to describe cry signal characteristics. According to the dynamic nature of sound as well as real-time requirements, the signal analysis is carried out on overlapping short-term frames with 32ms being a common length. We process these frames with an overlapping of 50% (Abdulaziz and Ahmad 2010; Petroni et al. 1995). In addition, Fourier analysis of the signal is performed to obtain the spectral representation. However, the necessary framing part generates a *spectral leakage* phenomenon. In order to overcome this issue, we apply on each frame a Hamming window in order to flatten the signal at the edges.

Some features that are widely used in speech recognition such as Mel frequency cepstral coefficients, linear predictive cepstral coefficients rely on the source-filter model presented in (Fant 1960) which describes formally the human speech production process as a combination of a vocal source, producing sounds (namely vocal cords), and a vocal tract (laryngeal and oral cavity) which acts, with its particular shape and positioning, as a filter. The source carries the fundamental frequency (f_0), whose variations encode prosodic speech information, however, the most useful information for (phone detection) is the exact position of the vocal tract (*i.e.* the filter) which models the formants (higher frequencies). The nature of the glottal pulse (source) causes energy to drop across frequencies resulting in less powerful formants and less information in the acoustic model. This phenomenon is called *spectral tilt*. To cop with it, high-frequencies energy has to be boosted or pre-emphasized using a high-pass filter. Figure 2 show the flow of audio signals through the pre-processing steps. At the end of these filtering and transformation steps, the resulting signals are ready to be processed in order to extract appropriate signal features.

Pre-cry units detection

A typical infant pre-crying episode encompasses different parts; utterances, or expiratory sounds, are the most relevant as they are the product of an intentional movement consisting in an airflow passing through infant vocal tract. In addition to silent parts, we can distinguish the presence of short (but powerful) regions corresponding to the sound produced by the inspiration phase of the vocalization and sometime by coughing. It is upon the same distinction that (Tsukamoto and Tohkura 1990; Baeck and Souza 2001) divided the whole cry episodes into single segment units in their experiments. The purpose of this step is then to extract only the relevant segments using an end-point detection method relying only on simple temporal features. In (Díaz et al. 2012), authors used, first, a threshold based on short-time energy so as to differentiate loud windows from the silent ones. Then, segments which have a duration less than 200ms are removed. With this second step, inspiratory sounds are eliminated. Same procedure was conducted in (Várallyay 2007) for the purpose of crying segmentation and in (Rabiner and Sambur 1975), zero-crossing rate is used in conjunction with short-time energy so as to detect end-points of speech utterances.

Features Extraction

Different features are extracted after the segmentation, transformation and filtering steps. In this work, temporal, spectral, prosodic and cepstral features are extracted and provide more than 40 real attributes for the dataset. Including respectively *root mean square* (RMS) for the loudness and *Zero crossing rate* (ZCR), *i.e.* the number of times the audio signal changes the sign, the timber to distinguish sounds of same loudness and pitch (fundamental frequency), spectral

centroid and spread.
Fourier analysis of the signal that have a power-of-two number of samples per frame is carried via the Fast Fourier Transform Algorithm. The result of this analysis is a spectrum which describes the frequency components of the signal. For a frame of size 256 samples the resulting spectrum will be of size 256 bins (which are complex numbers, so we take their magnitude). After performing the Fourier analysis, we extract *spectral Centroid*, *roll-off frequency*, *bandwidth* for each frame as well as *peaks*, *valleys* and *spectral contrast* (Jiang et al. 2002) after dividing the resulting spectrum into different octave-based subbands.
Variation in duration, pitch and intensity are the three acoustic cues that carry prosodic or suprasegmental information about infant cry and speech in general. These features describe the cry envelope or the shape that is formed by the successive frames. Infant crying fundamental frequency is characterized by its high pitch (250-700 Hz). This feature is then widely used in the detection of infant crying and shows good results in adverse conditions (Cohen and Lavner 2012). In the recognition part, it is used to describe the shape, the contour or the melody of the crying across frames(Várallyay 2007).

In this research study, the pitch is computed using an autocorrelation-based method (Boersma and Weenik 1996) which is provided in `Praat` [1]. This method shows good tracking performances outperforming other techniques according to (Babacan et al. 2013). While other techniques rely on the cepstrum, for example (Noll 1967), all of these have to cope with the high-pitched nature of crying which preclude a good tracking of fundamental frequency through time. Formants, linear predictive cepstral coefficients and Mel frequency cepstral coefficients of the cepstrum analysis (inverse Fourier transform of the logarithm of the spectrum of an audio signal) rely on the source-filter model developed in (Fant 1960). They are actively used in speech recognition systems as they have the ability to detect the phonemes (*i.e.* the shape of the vocal tract when the phoneme is pronounced).

Linear prediction cepstral coefficients (LPCC) represent an all-pole filter, according to the source-filter model, that captures the spectral envelope of a speech signal (formants or spectral resonances of the vocal tract), and have been extensively used for speech coding and recognition applications as well as in the context of infant cry recognition (Abdulaziz and Ahmad 2010).

The *Mel frequency cepstral coefficients* (MFCC) are a widely used metric for describing timbral characteristics based on the Mel scale. In our experiments, we used a 40 bands filter-bank and retained the first 13 coefficients as these will represent information merely about the vocal tract filter, discarding information about the glottal source. In (Abdulaziz and Ahmad 2010), authors used for their part, 12 coefficients from 50 ms and 100 ms frames in the same time.

[1] `Praat` (Boersma and Weenik 1996) is a software that provides tools for speech analysis and synthesis. It is commonly used in phonetics and speech research.

Cry type	Encoding Phoneme	Number of records	Number of examples
Eructation	Eh	59	1001
Flatulence	Eairh	19	2303
Discomfort	Heh	12	776
Hunger	Neh	131	15331
Tiredness	Owh	67	3656

Table 1: Number of records per class and the corresponding encoding phonemes (Dunstan 2009)

Experimental results

From the previous features extraction steps, a machine learning dataset is built from real data of infant pre-cry vocalizations. The obtained dataset is characterized by a high degree of imbalance in addition to a neighborhood bias (Hammerla and Plötz 2015) which is introduced by the framing step of the pre-processing and requires us to a careful interpretation of performance results obtained under cross-validation. The unbalanced nature of the dataset requires us to choose an appropriate evaluation metric so as to determine which classifier performs best. However, as outlined in (Forman and Scholz 2010), there is a wide disparity in performance arising from how these metrics are exactly calculated. Regarding the neighborhood bias, we experimented different validation methods under the balanced nature of our dataset, including a modified cross-validation/partitioning method (Hammerla and Plötz 2015).

Dataset description

The real dataset contain a total of 288 recordings of infant pre-cry vocalizations and cries produced by more than 30 babies of different ethnicities. The work-flow corresponding to the achievement of the different steps of the proposed machine learning process is presented along the different results yielding from each stage.

In the used dataset, the length of the recordings ranges from 0.3 s to 6.3 s, some vocalizations came in an isolated manner while others lasted for a long period. The different needs were encoded with a set of phonemes, which sound similar with the newborns utterance, and labelled on the basis of the perceived sound. Five cry labels are defined as shown in table 1 (Dunstan 2009). For example, the *Owh* vocalization which corresponds to tiredness reflects infant yawning. It encodes in some manner the oval-shaped mouth associated to it.

After the application of the pre-processing step [2] as defined in Section 3, we obtain a set of frames which repartition amongst considered classes is presented in Table 1.

Table 2 summarizes some results about the resulting set of frames from the pre-processing stage which rely mainly

[2] In this implementation, we used some functions provided by the `MIRtoolbox` (Lartillot and Toiviainen 2007) in both pre-processing and features extraction steps.

	All recordings
Total number of analysis window	55209
Retained windows (frames)	23027 (41.7%)
Retained windows without a corresponding f_0	656 (2.85%)
Total recordings	288
Total duration	890.08(s)
Number of segments	504

Table 2: Proportion of relevant frames

on the cry units detection phase. Results obtained after this phase show a large amount of dropped analysis windows from the recordings. Some of these windows correspond, after manual verification, to relevant parts that could possibly be taken into account. In the other hand, some retained frames do not have a corresponding fundamental frequency. We remove these frames which represents a total of 2.85% of the retained windows.

As we can see in Table 1, the dataset is highly unbalanced with a degree of imbalance of 3.4% corresponding to the amount of frames labeled as discomfort whereas the majority class corresponds to hunger. This could lead eventually to a wide disparity of results obtained under cross-validation depending on the method of calculation of the performance metric.

Evaluation metrics

As stated before, because of the highly unbalanced nature of our labelled dataset — 3.4% degree of imbalance being considered a challenging one —, in the following, we focus in the F-measure as, in addition to be the most employed metric in the presence of datasets that suffer from high class imbalance, it is the metric showing the trade-off between precision and recall and that we cosider as a reference so as to choose the right solution for our previously stated goal.

There are several methods used in the literature, and summarized in (Forman and Scholz 2010), to compute performance metrics like the F-measure as well as the area under the roc curve. These subtleties goes disregarded, however, the choice of the right method to calculate the performance metric has a great impact on the results (Forman and Scholz 2010). Here, we list the different methods used in our experiments to compute the F-measure. The i-superscripted measures correspond to measures obtained when the ith fold is used as the test set. Given the usual definition of precision $\mathrm{Pr}^{(i)}$ and recall $\mathrm{Re}^{(i)}$ for the ith fold, the first calculation method of the F-measure is done by averaging the F-measure obtained for each fold.

$$\mathrm{F}_{\mathtt{avg}} = \frac{1}{k} \cdot \sum_{i=1}^{k} \mathrm{F}^{(i)} \tag{1}$$

where $\mathrm{F}^{(i)} = \begin{cases} 2 \cdot \frac{\mathrm{Pr}^{(i)} \cdot \mathrm{Re}^{(i)}}{\mathrm{Pr}^{(i)} + \mathrm{Re}^{(i)}}, & \text{if both } \mathrm{Pr}^{(i)} \text{ and } \mathrm{Re}^{(i)} \text{ are defined} \\ 0, & \text{otherwise} \end{cases}$

This first method of computation, used under meta-segmented cross-validation, shows a significant drop at a segment legnth of 100 (Figure 3). As we can see in Figure 6,

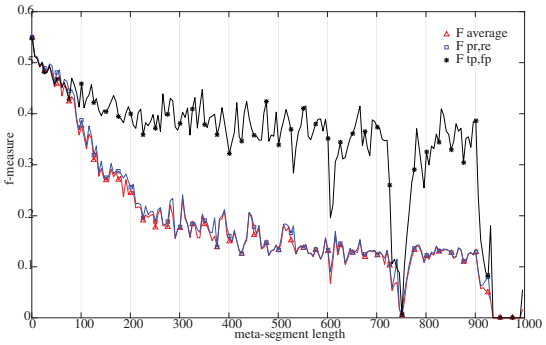


Figure 3: F-measures of a trained classifier (boosted trees) as a function of meta-segment length.

minority class is absolutely not represented in fold 7. This is worse for bigger segment lengths.

The second method computed as follow:

$$\mathrm{F}_{\mathtt{pr,re}} = 2 \cdot \frac{\mathrm{Pr}_{\mathtt{avg}} \cdot \mathrm{Re}_{\mathtt{avg}}}{\mathrm{Pr}_{\mathtt{avg}} + \mathrm{Re}_{\mathtt{avg}}} \tag{2}$$

where $\mathrm{Pr}_{\mathtt{avg}} = \frac{1}{k} \cdot \sum_{i=1}^{k} \mathrm{Pr}^{(i)}$ and $\mathrm{Re}_{\mathtt{avg}} = \frac{1}{k} \cdot \sum_{i=1}^{k} \mathrm{Re}^{(i)}$ and replacing each $\mathrm{Pr}^{(i)}$ and $\mathrm{Re}^{(i)}$ that are undefined by 0, follow the same trajectory as the first one with slightly little variations. However, F-mesure computed with the finale method;

$$\mathrm{F}_{\mathtt{tp,fp}} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

where TP, FP and FN are the total number of true positives, false positives and false negatives, respectively, does not show an abrupt drop, compared to the former methods, and stay relatively constant as the meta-segment length grows.

Results

Several ML strategies are applied on the dataset in order to find out the appropriate classifier that can provide significant classification rate to better recognize the infant cry. We consider the following parameters: a cubic kernel was used for the SVM. The Bagging and Boosted Trees were trained with 30 trees and a maximum depth of 20. As for the Decision Tree, the maximum number of splits is set to 1000 and Gini's split criterion is used. Table 3 summarizes performance results of the first part of our experiments obtained after performing a regular 10-fold cross-validation. Highest F-measure is obtained with bagged trees.

Unlike some representative related works (Chang et al. 2016), in our second part of experiments, each classifier was evaluated with a 10-fold meta-segmented cross validation to avoid the problem of overestimation of the quality of results induced by standard cross validation process (Hammerla and Plötz 2015). This technique relies on a modified partitioning procedure that alleviate the neighborhood bias, which results from the high probability that adjacent (moreover, overlapping) frames fall into training and test-set at the same time.

Figure 4 shows the prediction performances, namely the F-measure computed using Equation 3, of different classifiers with an increasing segment length. The greater the size

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
SVM	81.11	91.46	37.33	52.98
Bagged Trees	94.43	93.89	72.00	81.45
Boosted Trees	83.55	41.66	82.93	55.26
Decision Tree	86.52	61.98	39.07	47.95
kNN	97.78	91.37	36.27	51.91
Subspace kNN	92.23	81.67	70.40	75.54

Table 3: Recognition performances of different classifiers for the discomfort/non-discomfort pre-cryings obtained with a regular cross-validation process. The F-measure use equation 3.

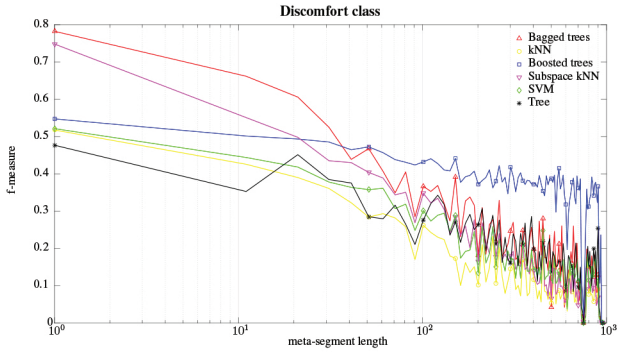


Figure 4: Prediction performances (F-measure computed with Equation 3) of several machine learning algorithms for discomfort class as a function of meta-segment length used during partitioning of the dataset. x-axis grows in a logarithmic scale.

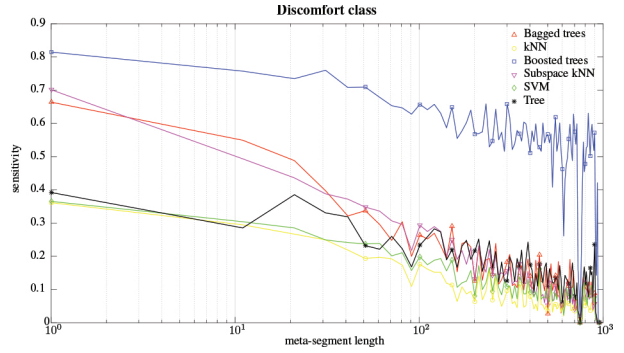


Figure 5: Sensitivity (or recall) as a function of meta-segment length. X-axis grows in a logarithmic scale.

of the segments, the less the adjacent frames are found in different folds, which reduces the bias associated with an overestimation of the results.

Ensemble methods, namely bagged trees and subspace kNN, show good performance results followed closely by boosted trees. Performances of these classifiers intersect at approximately a segment length of 25. As stated before, the goal is to reduce false negatives making the system act as a filter, and for this, highest sensitivity (or recall) is achieved by the boosted trees (Figure 5).

We notice that our modeling process for pre-processing stages and characteristics selection allow us to get good classification performances, particularly for the discomfort class, as shown by its area under the ROC curve in Figure 8, representing a good result considering the high degree of imbalance and the nature of observations at hand. Obtained results, without using hot topic ML algorithms including deep learning, validate, at least on our used dataset, the global analysis process of infant cries and confirms the studies conducted by (Dunstan 2009).

Conclusion

This paper presents an AAL system for infant monitoring that relies on various IoT sensing and actuation technologies. The main capability of the system, which is not supported in the proposals of the state of the art, is its ability of recognizing and analyzing infant (pre-)cries vocalizations, in particular those due to discomfort. This is done whithin an important timeframe — starting at the first manifestation of discomfort related vocalizations and extending to the prolonged and more difficult to soothe cries — where mother/caregiver-infant like interactions can be studied in the HRI context. Once the cry reason is identified at certain extent, the AAL system is able then to trigger the reactive calming actions by broadcasting, for instance, an adequate music, turning on the light spot with suitable color and luminosity, etc. The proposed mechanism for cry recognition and analysis is based on machine learning process, which exploits successfully pre-cry vocalizations of the infant in order to improve the overall performance. The ongoing works concerns the improvement of the reaction by learning from the response of the newborn to the previous actions; for instance taking into account the parameters of the actions that have failed or those that have contributed successfully in calming the infant. The future works concerns the validation of the targeted response adaptation mechanism on a dataset that will be collected in real conditions.

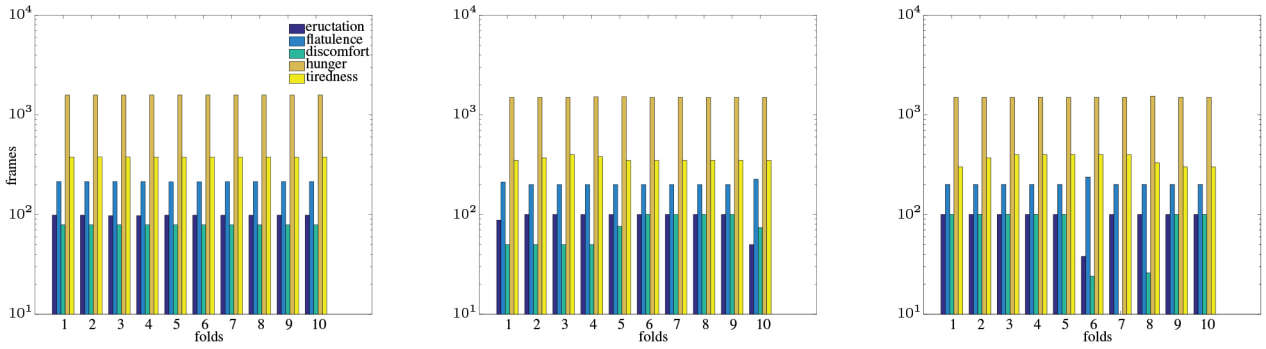
References

Abdulaziz, Y., and Ahmad, S. M. S. 2010. Infant cry recognition system: A comparison of system performance based on mel frequency and linear prediction cepstral coefficients. In *International Conference on Information Retrieval & Knowledge Management,(CAMP)*, 260–263. IEEE.

Babacan, O.; Drugman, T.; d'Alessandro, N.; Henrich, N.; and Dutoit, T. 2013. A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7815–7819. IEEE.

Baeck, H., and Souza, M. 2001. Study of acoustic features of newborn cries that correlate with the context. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, 2174–2177. IEEE.

Boersma, P., and Weenik, D. 1996. Praat: a system for doing phonetics by computer. report of the institute of phonetic sciences of the university of amsterdam. *Amsterdam: University of Amsterdam*.



(a) segment length 25 (b) segment length 50 (c) segment length 100

Figure 6: Distribution of classes over folds for an increasing segment length using a meta-segmented partitioning. Resulting fold distribution with a segment length of 1 (regular cross-validation partitioning) and 25 are the same. Note that fold 7 in (c) obtained with a segment length of 100 does not contain any sample from the minority class. Y-axis grow in a logarithmic scale

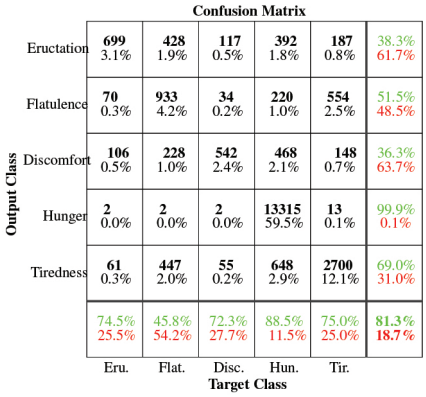


Figure 7: Confusion matrix for boosted trees classifier and a segment length of 25 frames.

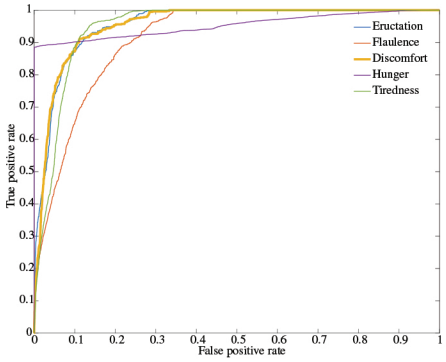


Figure 8: Receiver Operator Characteristics curve with all considered classes obtained with a segment length of 25. [best viewed in color]

Bănică, I.-A.; Cucu, H.; Buzo, A.; Burileanu, D.; and Burileanu, C. 2016. Automatic methods for infant cry classification. In *2016 International Conference on Communications (COMM)*, 51–54. IEEE.

Chang, C.-Y.; Chang, C.-W.; Kathiravan, S.; Lin, C.; and Chen, S.-T. 2016. DAG-SVM based infant cry classification system using sequential forward floating feature selection. *Multidimensional Systems and Signal Processing* 1–16.

Chang, C.-Y.; Hsiao, Y.-C.; and Chen, S.-T. 2015. Application of Incremental SVM Learning for Infant Cries Recognition. In *18th International Conference on Network-Based Information Systems (NBiS)*, 607–610. IEEE.

Cohen, R., and Lavner, Y. 2012. Infant cry analysis and detection. In *IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, 1–5. IEEE.

Cook, F.; Bayer, J.; Le, H. N.; Mensah, F.; Cann, W.; and Hiscock, H. 2012. Baby business: a randomised controlled trial of a universal parenting program that aims to prevent early infant sleep and cry problems and associated parental depression. *BMC pediatrics* 12(1):13.

Díaz, M. A. R.; García, C. A. R.; Robles, L. C. A.; Altamirano, J. E. X.; and Mendoza, A. V. 2012. Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis. *Biomedical Signal Processing and Control* 7(1):43–49.

Douglas, P. S., and Hiscock, H. 2010. The unsettled baby: crying out for an integrated, multidisciplinary primary care approach. *Med J Aust* 193(9):533–6.

Dunstan, P. 2009. *Child Sense: From Birth to Age 5, how to Use the 5 Senses to Make Sleeping, Eating, Dressing, and Other Everyday Activities Easier While Strengthening Your Bond with Your Child*. Bantam.

Fant, G. 1960. Acoustic theory of speech production.

Forman, G., and Scholz, M. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter* 12(1):49–57.

Fraiberg, S. 1950. On the sleep disturbances of early childhood. *The psychoanalytic study of the child* 5(1):285–309.

Hammerla, N. Y., and Plötz, T. 2015. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the ACM international joint*

conference on pervasive and ubiquitous computing, 1041–1051. ACM.

Hiscock, H., and Wake, M. 2001. Infant sleep problems and postnatal depression: a community-based study. *Pediatrics* 107(6):1317–1322.

Jiang, D.-N.; Lu, L.; Zhang, H.-J.; Tao, J.-H.; and Cai, L.-H. 2002. Music type classification by spectral contrast feature. In *Proceedings of IEEE International Conference on Multimedia and Expo ICME*, volume 1, 113–116. IEEE.

Lartillot, O., and Toiviainen, P. 2007. A Matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, 237–244.

Lieberman, P.; Harris, K. S.; Wolff, P.; and Russell, L. H. 1971. Newborn infant cry and nonhuman primate vocalization. *Journal of Speech, Language, and Hearing Research* 14(4):718–727.

Noll, A. M. 1967. Cepstrum pitch determination. *The journal of the acoustical society of America* 41(2):293–309.

Petroni, M.; Malowany, A.; Johnston, C.; and Stevens, B. 1995. A comparison of neural network architectures for the classification of three types of infant cry vocalizations. In *17th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, 821–822. IEEE.

Rabiner, L. R., and Sambur, M. R. 1975. An algorithm for determining the endpoints of isolated utterances. *Bell Labs Technical Journal* 54(2):297–315.

Rincon, E.; Beltran, J.; Tentori, M.; Favela, J.; and Chavez, E. 2013. A Context-Aware Baby Monitor for the Automatic Selective Archiving of the Language of Infants. 60–67. IEEE.

Sadeh, A. 1994. Assessment of intervention for infant night waking: parental reports and activity-based home monitoring. *Journal of consulting and clinical psychology* 62(1):63.

Scott, G., and Richards, M. 1990. Night waking in infants: effects of providing advice and support for parents. *Journal of Child Psychology and Psychiatry* 31(4):551–567.

Sohn, J.; Kim, N. S.; and Sung, W. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters* 6(1):1–3.

St James-Roberts, I. 2007. Helping parents to manage infant crying and sleeping: A review of the evidence and its implications for services. *Child Abuse Review* 16(1):47–69.

Tsukamoto, T., and Tohkura, Y. 1990. Perceptual units of the infant cry. *Early Child Development and Care* 65(1):167–178.

Várallyay, G. 2007. The melody of crying. *international journal of pediatric otorhinolaryngology* 71(11):1699–1708.

Varendi, H.; Christensson, K.; Porter, R. H.; and Winberg, J. 1998. Soothing effect of amniotic fluid smell in newborn infants. *Early human development* 51(1):47–55.

Wasz-Hockert, O.; Valanne, E.; Vuorenkoski, V.; Michelsson, K.; and Sovijarvi, A. 1963. Analysis of some types of vocalization in the newborn and in early infancy. In *Annales Paediatriae Fenniae*, volume 9, 1.

Wasz-Höckert, O.; Michelsson, K.; and Lind, J. 1985. Twenty-five years of Scandinavian cry research. In *Infant crying*. Springer. 83–104.

Weggemann, T.; Brown, J.; Fulford, G.; and Minns, R. 1987. A study of normal baby movements. *Child: care, health and development* 13(1):41–58.

Wermke, K.; Mende, W.; Manfredi, C.; and Bruscaglioni, P. 2002. Developmental aspects of infant's cry melody and formants. *Medical engineering & physics* 24(7):501–514.