On the Monitoring of Infant Discomfort Using Machine Learning Techniques

Massinissa Hamidi

[a] *Université Paris Diderot, France*
[b] *Internship conducted at Laboratoire LIPN-UMR CNRS 7030 PRES Sorbonne Paris Cité France*

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 fourties demonstrated that there is a universal language that hides behind the infant crying. This study describes an IoT platform for ambient assisted living, which is dedicated to the automatic assessment of newborn discomfort situations and providing assistive services to enhance infant as well as parents wellbeing. The present work focuses on the description of a promising approach relying on the machine learning analysis of both cry and 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.

Keywords: Infant cry, automated recognition, machine learning, dataset generation, feature selection, Ambient intelligence

1. Introduction

For decades, parents' main companion to monitor infants is the baby-phone with all the limitations that we know. The rapid advancement of pervasive computing technology in particular for wireless sensors networks and connected things encourages the investigation for developing context aware systems such as new generation baby-phone or any other replacement device that will have cognitive capabilities on which parents can rely. Such kind of systems will pave

*This work is initiated by A. Osmani and done in collaboration with him. Possible extention in a robotic environment should be done in collaboration with A. Chibani.

the way for several Ambient Intelligence (AmI) applications, which can significantly improve the quality of life and wellbeing of both parents and infants. The global scope of the research for innovative AmI applications for infants poses several challenges to develop smart systems that can monitor for instance sleep phases, crying, motion, or monitor vital signs to detect symptoms of infections or monitor some contextual parameters of infants local living environment such as temperature, humidity, noise, air quality, luminosity, etc. However, the detection and analytics of complex events characterizing infants cry by considering audio, video and/or physiological signals is among the most important challenging issues. Therefore, the main research issue that is considered in this work is how to develop an automatic cognitive process allowing a refined analysis and interpretation of the infant cry.

In fact, crying is a universal language used by infants from their earliest days as a biological alarm system to express their basic needs such as food or react to pain or discomfort; and with experience, parents learn how to recognize some patterns of the cry and infer baby's needs. In most cases, crying is related to discomfort, it concerns 25% of the 145 million newborn worldwide on a regular basis according to the 2015 world population data sheet published by the Population Reference Bureau. According to [1], it has been stated that there is a strong correlation between baby cry and his specific needs. More recently, in [2] authors 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 [3].

This work proposes machine learning mechanism as the core component of an overall ambient intelligence system to continuously monitor infant behavior in order to detect and reduce discomfort. We propose a complete machine learning process that includes low-level audio features selection methods from labeled infant pre-cry recordings such as spectral descriptors and Mel frequency cepstral coefficients as well as high-level features characterizing the envelop of the crying. The classification is performed using different machine learning algorithms after the features selection step. To this end, the proposed approach improve the discomfort learning rate by analyzing pre-cry period. Obtained results show that learning rate of baby discomfort state is close to or better than results using cries only. This promising result gives the opportunity to develop new baby monitors, such as the one we presented in [4], able to anticipate the infants needs.

The document is organized in the following manner: Section 2 presents the related work framed in a way that highlights evolution of the field from manual inspection of infant cries to recent works on the automatization of such tasks. In Section 3, we detail the proposed approach and its relevance to the human-robot interaction field which is then followed, in Section 4, by a description of the system that we propose and its architecture. Starting from Section 5, we dig into audio signal analysis with a description of the main steps of preprocessing and filtering. Section 6 gives an overview of the extracted features from the different constructed representations. Concerning the experimentation part, we

start with a quantitative as well as a qualitative description of the real dataset. Obtained results are given in Section 8. We end up with a conclusion and pointers to future work.

2. Related work

We organized the related work into different parts: we first present studies that dealt with categorizing infant crying from both a pediatric and a technical point of view. This is followed by the studies that were interested in the automatization aspects of such tasks which is linked then to a review of recent industrial products.

Infant cry studies. The interest in the study of infant crying started long ago. The first works on this subject appeared in the 1940s [5, 6]. Several active research groups have been formed since then, namely, the scandinavian reseach group [7, 8] and the hungarian research group [9]. These groups focused on the developmental and pathological aspects related to crying. Indeed, authors in [7] have studied the crying of the infant from a medical point of view in order to diagnose, for example, the symptoms of serious illnesses during the first days of the infant through analysis of qualitative features of the crying.

Despite this enthusiasm of the medical corps for the prospects offered by the study of infant cries in the early diagnosis of diseases, scientific work was not confined to the study of this aspect alone. There has also been interest in the categorization of the rasons to which crying is due.

Mentionning categories of cry in studies. To our knowledge one of the first papers to have investigated the proprieties for crying due to different causes is Fairbanks in 1942 [5]. The author has in particular studied the fundamental frequency of the signal corresponding to cries induced experimentally by the sensation of hunger. Following this, other categorizations have emerged and have been studied experimentally. Table 1 summarizes some studies that have been conducted in this sense.

Interest on categorization and distinction amongst classes. 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 [1]. In [1] for example, authors dealt with four types of infant vocalizations: the first birth cry, the hunger cry, the pain cry, and the pleasure cry. Authors came to the conclusion that "*these cry types could be distinguished from each other both auditorily and by means of sound spectrography*". In [10], authors were interested in three categories of cry, namely, the hunger cry, the call cry and the anger cry. In [11], attention was drawn toward the automatic classification of infant cries into hunger, pain as well as sleepness categories while in [12], authors were interested in categorizing anger, pain and fear cries.

In the continuity of the work of categorizing infant cries, one of the most recent and most interesting proposals in this field is certainly that of Dunstan [2].

the author proposes an auditory technique which exploits the vocalizations of all kinds produced by the infant before the more difficult-to-soothe cries appear. This proposition is not fortuitous since a long series of studies were carried out in this sense, starting with that of Irwin [6].

Interest on phoneme distinction in infant cry. Phoneme distinction or vowel descrimination in infant utterances have been studied for the first time in [13]. Author noticed patterns of vocalization indicating emergence of a capacity for speech. this work was continued in [14, 15, 16] with ramifications in the first signs of speech. Indeed, vocalizations that are studied are considered as the precursory signs of communication in infants.

Author-year	# of infants	Age	Categories
Irwin (1948) [14]	62	1^{st} to 30^{th} month	Front vowels Middle vowels Back vowels
Lieberman et al (1971) [17]	20	Birth to 4^{th} day	Birth cry Fussing cry Angry cry Gurgles Hunger cry Shrieks Inspiratory whistles
Papouvsek et al. (1992) [18]	NA *	2 months	Joy Comfort Neutral Discomfort Cry
Hsu et al. (2000) [19]	13	4^{th} to 24^{th} week	Speech quality (syllabic versus vocalic) Melodic complexity (simple versus complex)
Galaviz et al. (2004) [20]	NA *	2 months	Normal Hypo-acoustic (deaf) Asphyxia

* NA: not available.

Table 1: List of some studies concerned with categorization of infant's cries from a developmental and pathological point of view.

Infant cries and sleep disorders. The infant discomfort is an important societal problem especially since in our modern societies both parents work. In addition to the discomforts caused to the child and the whole family, this problem has an important financial impact, as shown by several studies such as the one carried out in England described in [21]. This study focuses on children aged 1-3 months and shows that, in addition to parents efforts, the cost of the professional time spent helping parents to manage crying and sleeping that the financial outlay related to crying is in the order of 65 million pounds per year for the British National Health Service.

Numerous studies have focused on one hand on the psychological and development aspects underlying newborn infants sleep and cry problems and on the other hand on pathological aspect [22, 23] . 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 [24], 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, [25] 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. Author in [23], for his part, states that not all babies know how to put themselves to sleep and how to ressetle after night waking.

Automatic categorization of infant cries. while these studies are based on the manual inspection of experts in the field, recent advances in automatic signal processing and machine learning have opened the prospects for automating the treatment of infant crying to the diagnosis of diseases and more importantly for the detection and monitoring of situations due to infant discomfort.

Earlier studies used mainly variations of the fundamental frequency of infant cries signals. Since then, many other techniques have been borrowed from the field of signal processing and speech recognition in order to detect and distinguish between the different causes to which crying is due. Authors in [12] extracted Mel frequency cepstum coefficients from real database of infant cries in conjunction with a feed-forward neural network for the classification part. In [11], authors used a set of temporal and spectral characteristics of the signal with support vector machines algorithm. These studies show good results in the recognition of emotional state of infants through cries. Table 2 gives a brief sumary of significant works on the categorization of infant cries.

Indeed, discomfort situations appear to be recurrent within families and recent advances in the field of AmI and pervasive technologies can be used to address this problem effectively. Actually, in addition to vocalizations, behavioral patterns such as movements are sometimes dismissed as unintentional or purposeless however, studies, for example [26], 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.

Infant monitoring. From industrial point of view, some products implementing these new generation of baby-phones are proposed. It is the case with *Owlet Baby Care Inc.* which released its first product named Owlet in 2013; Their product takes the form of a sock and measures the newborn's heart rate, as well as blood oxygen saturation levels and skin temperature to assess sleep quality and allow parents to access gathered data via a smartphone app.

Still in the wearable branch *Mimo* [27] is another product that fits into this framework. It is a onesie for babies that, similarly to the Owlet's sock, includes sensors for acquiring baby's respiration, moisture and temperature in order to be further analyzed and track the sleep schedule of the infant to make predictions on waking and sleep patterns.

The best commercial product of this new babyphone generation is probably the *Sproutling* product presented by the company of the same name before being bought by Mattel company for $21 million in february 2016. We note that mattle's goal is to be recognized leader in play, learning and development worldwide and that this company prepares a new smart baby monitor platform

Author-year	# of infants	Feature Extraction	Classifiers	Performance results
Petroni M. et al. [12] (1995)	16	Mel frequency cepstral coefficients	artificial neural network	90.4%
Messaoud A. et al. [29] (2010)	13	Mel frequency cepstral coefficients	artificial neural networks	71.4%
Abdulaziz Y. et al. [30] (2010)	NA *	Mel frequency cepstral coefficients and linear prediction cepstral coefficients	feed forward neural network architecture	76.2%
Chang C. Y. et al. [11] (2016)	37	temporal and spectral features followed by a feature selection step	support vector machines	93.75%

* NA: not available.

Table 2: List of some recent works selected to reflect the different techniques used for the categorization of infant's emotions and for building a complete automated system that could recognize the different cry types.

named *Aristotle* [28]. This product exploit the capabilities of *Microsoft's Cortana* to initiate interactions with the infant in likeness with what is being done by *Amazon's Alexa* or *Google's Home*. It is mainly designed with the goal of entertaining, and assisting the infant during each development stage; in early days of the infant, it plays the role of white noise emitter, a monitor for infant's basic needs and a night light.

The system that we propose aims to take into account the aforementioned functions by insisting on the reduction, in a first stage, of discomfort situations and, secondly, by having a medical aim with the possibility to monitor vital parameters and to diagnose certain disease. It allows also parents access their infant's status via a mobile application.

3. 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 [23].

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 work and simulate mother/caregiver-infant like

interactions via the different actuators and the response selection module (see Section 4). 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.

4. Overview of the proposed system

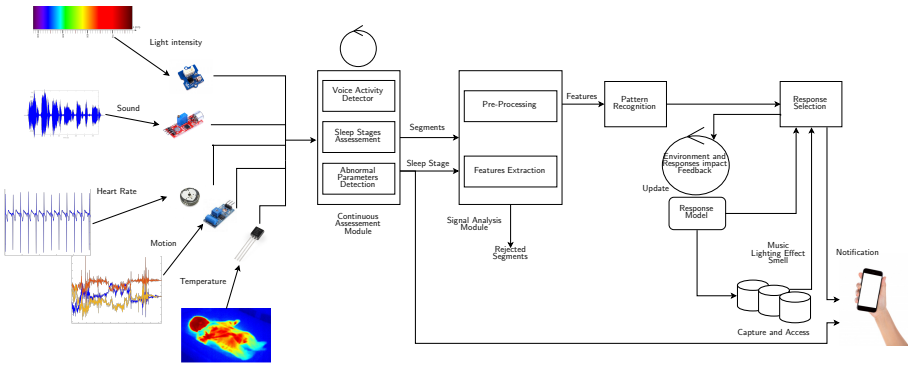


Figure 1: Architecture of the proposed smart baby monitor

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, detection of strong agitation that is caused mainly by overstimutlation in noisy environments. It also gives indications about vital parameters (body temperature, heart rate, oxygen saturation, body movements, respiration rhythm). 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.

4.1. Continuous Assessment Module

The personal assistant will include, in particular, a voice activity detector (VAD) [31] 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.

4.2. 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.

4.3. 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 [2]. 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 document, we refer to cry to indicate cry, pre-cry or phonetic sound).

4.4. 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.

4.5. 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 style as capture and access systems [11, 32], 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.

in the following, we are interested in the treatment of infant crying through the analysis of cry audio signals since it is these signs that concentrate all the information essential to detect accurately discomfort situations.

5. Audio Signal Preprocessing

Cry recordings are heterogeneous signals by nature, which are characterized by variable length, the presence of noise as well as different resolutions that are used by the analog-to-digital converters of the audio sensors. This is why it is important to start the pre-processing stage with signal *normalization* to obtain a uniform dataset of recordings in terms of resolution. A resampling of all audio recordings is performed; our original signals have a sampling rate of 44100 Hz, we choose to downsample the recordings at a sampling rate of 8 kHz followed by an anti-aliasing procedure.

In addition, because of the nature of audio signals and the dynamic behavior of sounds it is necessary to detect at what moment useful cry signal starts and ends using *end-point detection algorithm* [33]. To facilitate signal analysis the whole signal is split into frames which size is typically 32(ms). Additional signal preprocessing actions are also operated to prepare signal analysis including cry units detection [34] to isolate utterances in the signal as shown in Figure 2. Besides utterances (useful parts of the signal), the considered episode of cry encompasses moments of silence and also, breath inspiration sounds. These parts carry no valuable information about the cry and thus, have to be avoided in filtering step. Windowing eliminates spectral leakage and pre-emphasizing deals with spectral tilt. These steps will be detailed in the following.

5.1. Cry units detection

As we can see in Figure 2, a typical infant crying episode encompasses different parts; cry 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 crying and sometime by coughing. It is upon the same distinction that [10, 35] 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 [34], authors used, first, a threshold based on short-time energy so as to differentiate loud windows from the silent ones (see figure 4a

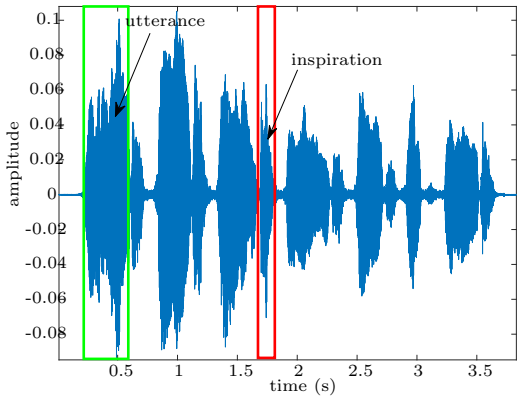


Figure 2: Waveform corresponding to an infant crying episode. We distinguish between a useful signal for cry detection (utterance box) and noise signal like inspiration box.

and 4b). The energy threshold is computed using a sliding window w of size m as follow:

$$\text{EnergyThreshold} = \frac{E_n(R)}{4} \tag{1}$$

where $E_n(R) = \frac{1}{N}\sum_m [x(m) \times w(n-m)]^2$ and R is a cry recording. Then, segments which have a duration less than 200ms [34, 19] are removed (see figure 4(a)). With this second step, inspiratory sounds are eliminated. Same procedure was conducted in [36] for the purpose of crying segmentation and in [33], zero-crossing rate is used in conjunction with short-time energy so as to detect end-points of speech utterances. Algorithm 1 describes a modified version of the usual cry units detection procedure by exploiting pitch data obtained using `Praat` [1].

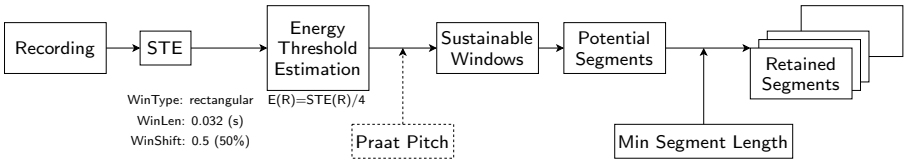


Figure 3: Block diagram of the short-time energy based segmentation procedure. Dashed part corresponds to the modified version. In this version, Pitch generated in `Praat` is used in conjunction with short-time energy in order to refine segmentation and decrease amount of sustainable windows that are dropped in the regular procedure.

[1] `Praat` [37] is a software that provides tools for speech analysis and synthesis. It is commonly used in phonetics and speech research.

Algorithm 1: Modified procedure for cry units detection using `Praat` pitch

Data: `Recording, winType, winLen, winShift, PraatPitch`
Result: Set of utterances/segments
Generate a windowing function of type `winType`;
Generate a set W of windows from `recording` according to `winLen` and `winShift`;
foreach $w \in W$ **do**
    Apply the windowing function on w;
    Compute energy of w;
    $E_n \leftarrow E_n + e_w$;
$E_{threshold} \leftarrow \frac{E_n}{4}$;
foreach $w \in W$ **do**
    if $e_w \geq E_{threshold}$ **then**
        Mark w as sustainable;
    else
Form clusters of sustainable windows;
Keep clusters c greater than 200ms in C;
foreach $c \in C$ **do**
    Expand c to the right while w has a `Praat` pitch;
    Expand c to the left while w has a `Praat` pitch;
return C;

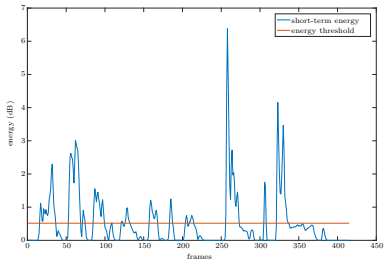
5.2. Framing

At this point of preprocessing, given a infant crying recording, we get a set of relevant segments corresponding to the utterances. Therefore, in order to achieve a fine-grained analysis on the audio signal, as the dynamic evolution of the statistical properties and signal non-stationarity must be taken into account — since the stationary assumption in case of speech is valid for 10 to 30 ms. Typically, the signal analysis has to be carried out on overlapping short-term frames with 32ms common length. We process theses frames with an overlapping of 50%. See Figure 5).

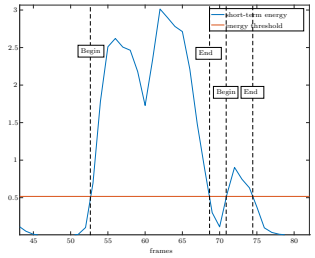
These frames are then ready to be processed in the time-domaine, however, concerning frequency-domain analysis, we have to deal with some issues related to this stage as well as the human speech production mechanism.

5.3. Dealing with spectral leakage

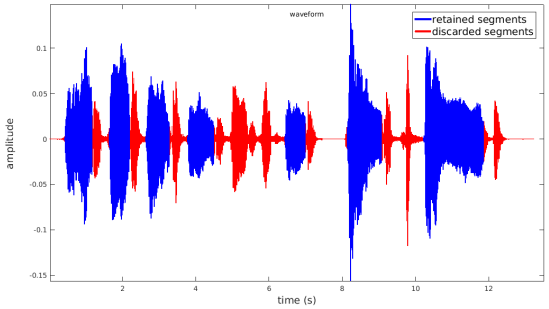
Filtering and transformation steps are needed to obtain high-quality voice signals that will allow an efficient and highly accurate features extraction and classification. The spectral leakage is caused when performing Fourier analysis on the signal treated after the framing step. In order to overcome this issue, a Hamming window is applied on each frame. This will result in flattening the signal at the edges so as to maintain a certain form of continuity between its beginning and its ending.



(a) Short-time energy of a crying episode.



(b) Zoom-in on the second utterance of the short-term energy waveform in 4a. Start and end points are highlighted.



(c) Valuable segments (blue), corresponding to the delimited parts in Figure 4b and lasting more than 200(ms), are kept for further processing stages.

Figure 4: Cry segments and end-point detection algorithm results

5.4. Source-filter model and spectral tilt

Human speech-production mechanism and modeling. Some features that are widely used in speech recognition such as MFCC, LPCC rely on the source-filter model presented in [38] 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 and so as to cop with it, high-frequency energy has to be boosted or pre-emphasized.

$$x'(n) \quad = \quad x(n) - \alpha \times x(n-1) \tag{2}$$

with $x(n)$ is a sampling point of the audio signal and $0.9 \leq \alpha \leq 1$.

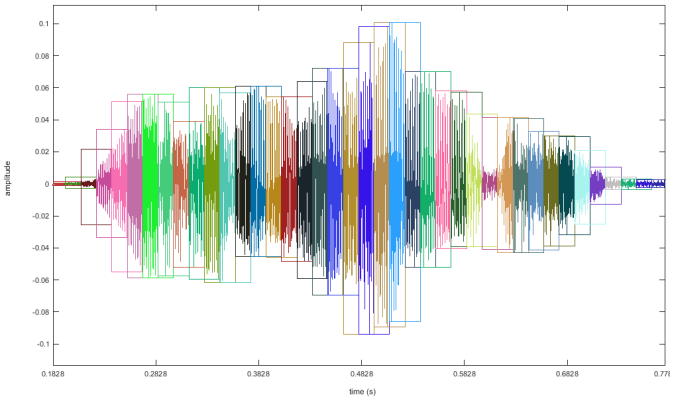


Figure 5: Part of a waveform after framing. Each rectangle represents a frame

Filter design. In order to design an optimal filter in the case of infant vocalizations, we have to find an optimal cut-off frequency. Figure 6 shows the different frequency responses of the finite impulse response (FIR) filter given above in Equation 2 as a function of the coefficient α. We can see that attenuation of low frequencies and improvement of the high frequencies can be achieved by setting the coefficient α between 0.9 and 1.

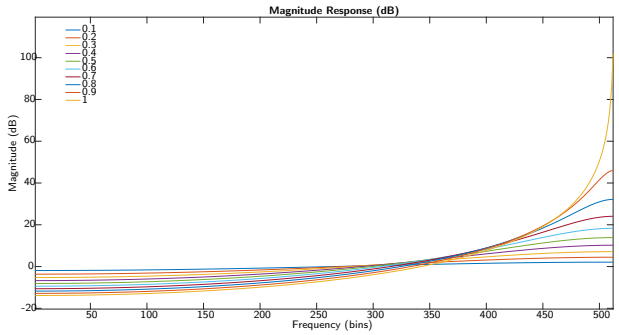


Figure 6: Magnitude response of the FIR filter given in Equation 2 as a function of the coefficient α.

At the end of these filtering and transformation steps, the resulting signals are ready to be processed in order to extract appropriate signal features.

6. Features extraction

In the following parts, a detailed description of the different features that are extracted after the segmentation, transformation and filtering steps. In this work four feature types are extracted, namely, temporal, spectral , prosodic and cepstral features to generate an exploitable dataset on which classification algorithms can be applied. All of these features provide more than 40 real attributes. We organise extracted features on temporal features, spectral features, prosodic features and cepstrum features. All of these features provide more than 40 real attributes to build machine learning dataset. Note that in the following, N refers to the number of samples per frame.

6.1. Temporal Features

The Root Mean Square of the cry signal energy is applied to approximate the loudness.

$$\text{RMS(frame)} \quad = \quad \sqrt{\frac{\sum_{i=1}^{N-1}(x(i))^2}{N}} \tag{3}$$

Zero crossing rate is a measure of the number of times the audio signal changes sign. It is computed for each frame as follows:

$$\text{ZCR(frame)} \quad = \quad \sum_{i=1}^{N-1} |sgn(x(i)) - sgn(x(i-1))| \tag{4}$$

where

$$sgn(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{otherwise} \end{cases}$$

Unvoiced sounds produced in the inspiratory phase of infant crying have higher zero-crossing rate while voiced sounds which correspond to the utterances have a lower rate one.

6.2. Spectral Features

Timbre is the property that distinguish sounds of same loudness and pitch *i.e.* fundamental frequency. Spectral centroid and spectral spread are among other descriptors used for characterizing timber. The following characteristics are intended to describe the shape of the computed spectrum, highlighting the spectral components of the signal and their distribution.

Fourier analysis, usually noted $\mathcal{F}$, of signals 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). In the following, $X(k)$ refers to the magnitude of the spectrum at the kth bin.

The Spectral Centroid is an indicator of the brightness of a given spectrum and represents the spectral center of gravity. It is defined as:

$$\text{SpectralCentroid(frame)} = \frac{\sum_{k=0}^{N-1}(k \times |X(k)|)}{\sum_{k=0}^{N-1}|X(k)|} \quad (5)$$

The Roll-off frequency is the point in the spectrum below which the 85% of the magnitude distribution is concentrated. This measure is useful in distinguishing voiced from unvoiced speech.

The cry signal bandwidth is measured by spectral spread and defined as the spread of the spectrum around its mean value.

$$\text{SpectralSpread(frame)} = \frac{\sum_{k=0}^{N-1}(k - \text{SC(frame)})^2|X(k)|}{\sum_{k=0}^{N-1}|X(k)|} \quad (6)$$

where SC(frame) is the spectral centroid computed with (5).

Similarly to music type classification used in [39] strength of peaks and valleys in spectral subbands are computed. The spectrum is divided into different octave-based subbands and spectral contrast is estimated from each sub-band. The spectral components are sorted in a descending order before peaks and valleys are computed. The spectral contrast of the bth band is defined as follows:

$$SpectralContrast_b(frame) = Peak_b - Valley_b \quad (7)$$

where $Peak_b$ is defined as:

$$Peak_b(frame) = \log\{\frac{1}{\alpha N_b}\sum_{i=0}^{\alpha N_b - 1} X'_b(i)\} \quad (8)$$

and $Valley_b$ as:

$$Valley_b(frame) = \log\{\frac{1}{\alpha N_b}\sum_{i=0}^{\alpha N_b - 1} X'_b(N_b - i)\} \quad (9)$$

and $X'_b(i)$ are the sorted magnitudes of the bth octave sub-band and N_b is the size of the bth octave sub-band. The value of the quantile α is set to 0.02.

6.3. Prosodic features

Variation in duration, pitch and intensity are the three acoustic cues that carry prosodic or suprasegmental information about infant cry and speech in general. They provide useful information about speaking style of a person and in the case of infant crying, these features along with their calculated statistics (mean, variance, skewness, etc.) describe the cry envelope or the shape that is formed by the successive frames.

Adult fundamental frequency (f_0) (or pitch) ranges between 85-180(Hz) however, 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 [40] In the recognition part, it is used to describe the shape, the contour or the melody of the crying across frames [36].

In this research study, the pitch is computed using an autocorrelation-based method [41] which is provided in `Praat`. This method shows good tracking performances outperforming other techniques according to [42]. While other techniques rely on the cepstrum, for example [43] , all of these have to cope with the high-pitched nature of crying which preclude a good tracking of fundamental frequency across time.

6.4. Cepstrum-domain analysis

Formants, LPCC and MFCC rely on the source-filter model developed in [38] . 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).

The Cepstrum is defined as the inverse Fourier transform (IFT) of the logarithm of the spectrum of an audio signal as follows:

$$PowerCepstrum \quad = \quad |\mathcal{F}^{-1}\{\log(|\mathcal{F}\{x(t)\}|^2\}|^2 \tag{10}$$

where $\mathcal{F}$ is the Fourier transform and $\mathcal{F}^{-1}$ is its inverse. Cepstral analysis is particularly reliable at isolating the contributions of the source and the filter in speech production mechanism yielding a set of coefficients among which the lowest ones encode the filter. The cepstrum is the basis for MFCC and LPCC calculation.

The formants or harmonics, are the vocal tract resonance frequencies. Their mean values vary, according to the study conducted in [44] from 1019(Hz) to 2944(Hz) and from 6632(Hz) to 8059(Hz) for the first (f_1) and second formant (f_2) respectively. This same study provides strong evidence for an active neurophysiological tuning mechanism among infants in early stages of their life, which is reflected by their ability to control f_1 and f_2 productions. Formants are computed through Linear Prediction Coefficients (LPC) [45] based on the pre-emphasized and windowed audio signal.

The Mel frequency cepstral coefficients (MFCC) are a widely used metric for describing timbral characteristics based on the Mel scale. The Mel scale is a perceptual psycho-acoustic scale derived from human hearing experiments and tends to mimic the human perception of sounds, so do the Bark and equivalent rectangular bandwidth (ERB) scales. In our experiments, we used a 40 bands filterbank and retained the first 13 coefficients as these will represent information merely about the vocal tract filter, discarding information about the glottal source.

At the end of this feature extraction step, each labelled infant cry signal is described by values of its computed features and it is represented as a set of learning examples. The next section gives more details about used data and obtained results.

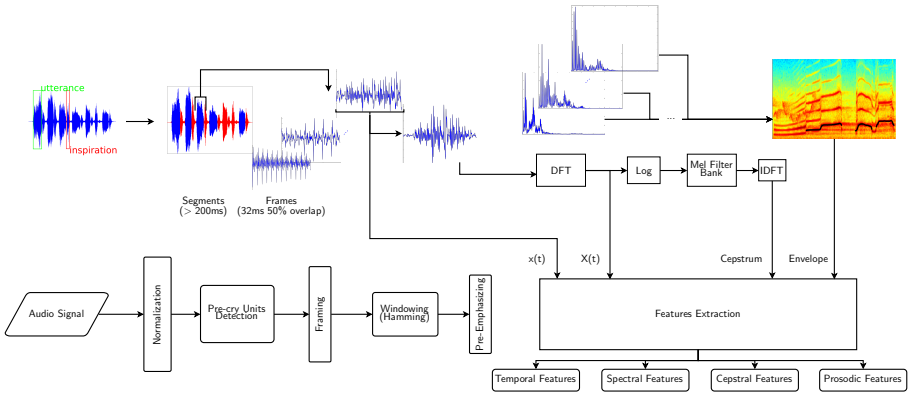


Figure 7: Block diagram of data preprocessing and features extraction

7. 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 with 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 3 [2]. For example, the *Owh* vocalization which corresponds to tiredness reflects infant yawning. It encodes in some manner the oval-shaped mouth associated to it.

Neh corresponds to the sensation of hunger. It is related to the sucking reflex which appears at birth and is common to all mammals. When baby's tongue touches the roof of their mouth, the sound 'n' can be heard distinctly at the beginning of the vocalizations.

Related to gas discomfort and pain, the *Eh* pre-cry vocalization is produced when baby needs to burp, *i.e.* release gaz bubbles up the esophagus. Physically, this sensation causes the baby to squirm.

Heh vocalization is produced when baby is too warm/cold or when it has a dirty diaper. This often causes baby to pant producing a distinctive breathy-sound starting with a 'h'.

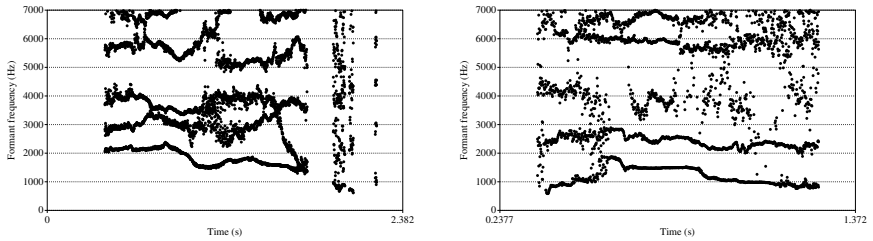
7.1. Qualitative study of a given infant cry signal

Here we give a brief description of the qualitative features for some vocalizations produced by infants in different situations, namely, tiredness and discomfort. This description is composed of the resonance frequency (formant) diagram obtained using `Praat`. It shows in some manner the complexity of the vocalizations and is considered as a "*good indicator for neuro-muscular maturation as*

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 3: Number of records per class and the corresponding encoding phonemes [2]

well as for the evaluation of pre-speech development" [44]. The variations of formant frequencies allows an estimation of articulatory activity during pre-speech vocalization.

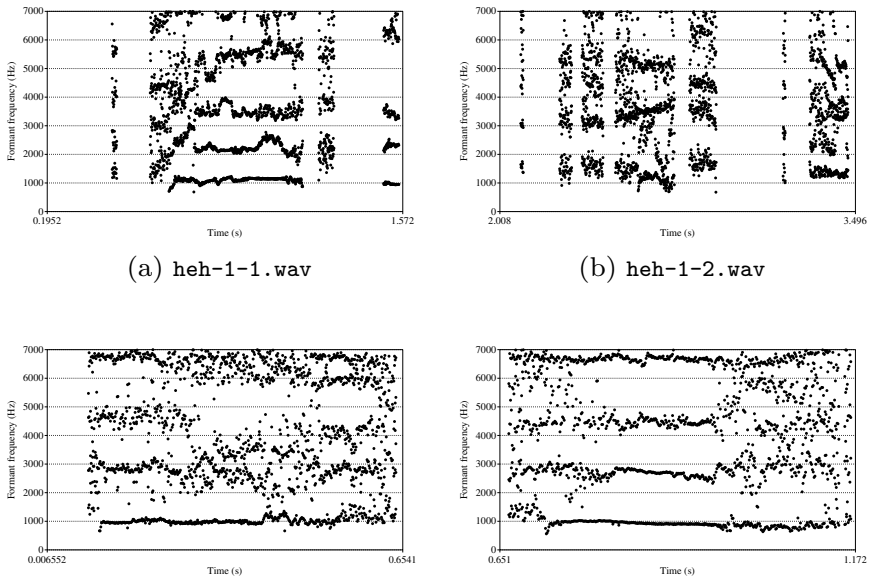


(a) `owh-10.wav` (b) `owh-1.wav`

(c) `owh-2.wav` (d) `owh-26.wav`

Figure 8: Qualitative study of *Owh* vocalizations. Resonance frequency diagrams of utterances produced by babies in situations of tiredness. All utternaces present a distinctive first formant (F_1) which contour is clearly defined. The remaining formants are also distinguishable but to a lesser extent. All utterances express variation patterns which correspond to the *raising-falling* pattern introduced, among other melodies, in [46]

`Praat` computes the LPC coefficients using the algorithm by Burg after resmapling, pre-emphasising and applying a Gaussian-like window on the original signal. The maximum number of formants that have been specified is five. The number of computed LPC coefficients is twice the number of formants.



(a) `heh-1-1.wav` (b) `heh-1-2.wav`

(c) `heh-2-1.wav` (d) `heh-2-2.wav`

Figure 9: Qualitative study of *Heh* vocalizations. These show contour of the first formants/harmonics of utterances produced by babies in situations of discomfort. All utternaces present a distinctive first formant (F_1) which contour is clearly defined. The remaining formants however are not clearly distinguishable, especially in the second part of Figure 9c where intersection is more pronounced. The formants, especially the first, have a flat tendency and don't show any variation pattern. This show a very low articulatory activity.

8. Experiments

From the previous features extraction steps, a machine learning dataset is built using real data of infant pre-cry vocalizations. The obtained dataset is characterized by a high degree of imbalance in addition to a neighborhood bias [47] 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 [48], 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 [47].

8.1. Cross-validation

The k-fold cross-validation is widely used for assessing performances of a prediction model. It performs, first, a random partitioning of the dataset into

k distinct folds. Then, at each iteration, a different subset consisting of $k-1$ folds is used to train the model while the remaining fold is used for the purpose of validation.

Usual partitioning and neighborhood bias. The random partitioning used in the case of segmented time-series introduces a neighborhood bias [47]. This bias consists in the high probability that adjacent and overlapping frames, that are typically obtained with a process similar to the one we have developed in Section 5 and that share a great deal of characteristics fall into training and validation folds in the same time. This leads to an overestimation of the validation results and goes often disregarded in the literature.

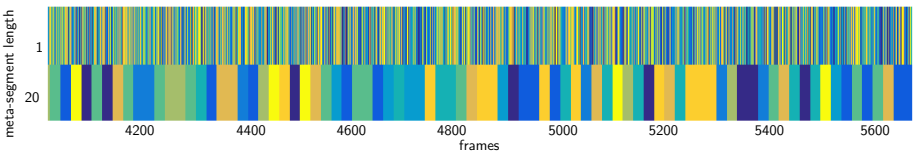


Figure 10: Partitioning of a portion of the final labeled dataset's frames over 10 folds using meta-segmented partitioning algorithm proposed in [47]. A segment length of one corresponds to the partitioning produced by the regular cross-validation procedure. Shown frames range from 4030 to 5670 are ordered by their time indexing across recordings. Each color corresponds to a different fold.

Meta-segmented partitioning. which is proposed in [47, Section 5] tries to circumvent this bias by, first, grouping adjacent frames into meta-segments of a given size. These meta-segments are then distributed on each fold. Figure 10 shows the partitioning of a subset of the generated frames over 10 folds with a meta-segment length of 1, which corresponds to the usual partitioning procedure, and a meta-segment length of 20.

After the application of the pre-processing step [2] as defined in Section 5, we obtain a set of frames which repartition amongst considered classes is presented in Table 3.

Table 4 summarizes some results about the resulting set of frames from the pre-processing stage which rely mainly 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 3, the dataset is highly unbalanced with a degree of imbalance of 3.4% corresponding to the amount of frames labeled as discomfort

[2] In this implementation, we used some functions provided by the `MIRtoolbox` [49] 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 4: Proportion of relevant frames

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.

8.2. 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 consider 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 [48], 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 [48]. Here, we list the different methods used in our experiments to compute the F-measure [3]. 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.

$$\mathrm{F}_{\mathtt{avg}} \quad = \quad \frac{1}{k} \cdot \sum_{i=1}^{k} \mathrm{F}^{(i)} \tag{11}$$

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 length of 100 (Figure 11). As we can see in Figure **??**, minority class is absolutely not represented in fold 7. This is worse for bigger segment lengths.

The second method computed as follow:

$$\mathrm{F}_{\mathtt{pr,re}} \quad = \quad 2 \cdot \frac{\mathrm{Pr}_{\mathtt{avg}} \cdot \mathrm{Re}_{\mathtt{avg}}}{\mathrm{Pr}_{\mathtt{avg}} + \mathrm{Re}_{\mathtt{avg}}} \tag{12}$$

[3] http://lipn.univ-paris13.fr/~hamidi/testsetperformances

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;

$$\mathrm{F}_{\mathtt{tp,fp}} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

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.

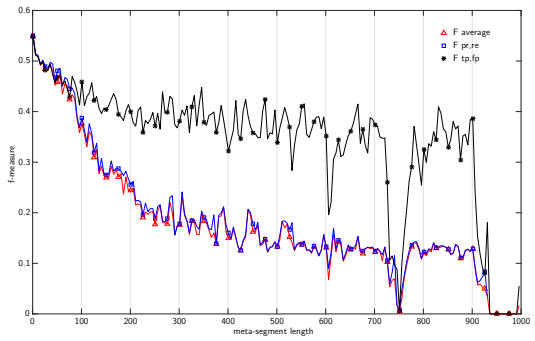


Figure 11: F-measures of a trained classifier (boosted trees) as a function of meta-segment length.

8.3. 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 5 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 [11], 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 [47]. 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.

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 5: Recognition performances of different classifiers for the discomfort/non-discomfort pre-cryings obtained with a regular cross-validation process. The F-measure use equation 13.

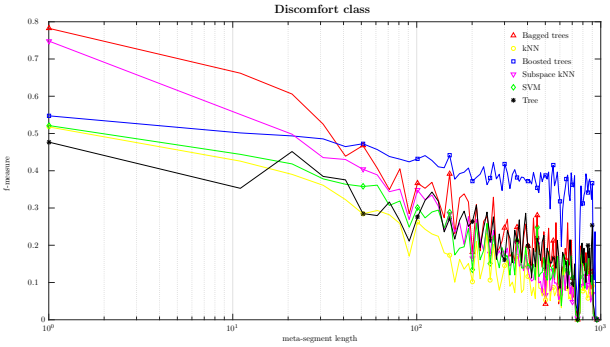


Figure 12: Prediction performances (F-measure computed with Equation 13) of several machine learning algorithms for discomfort class as a function of meta-segment length used during partitioning of the dataset. x-axis grows in a logarithmic scale.

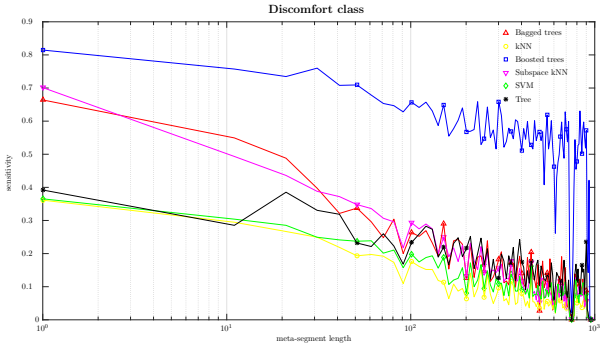
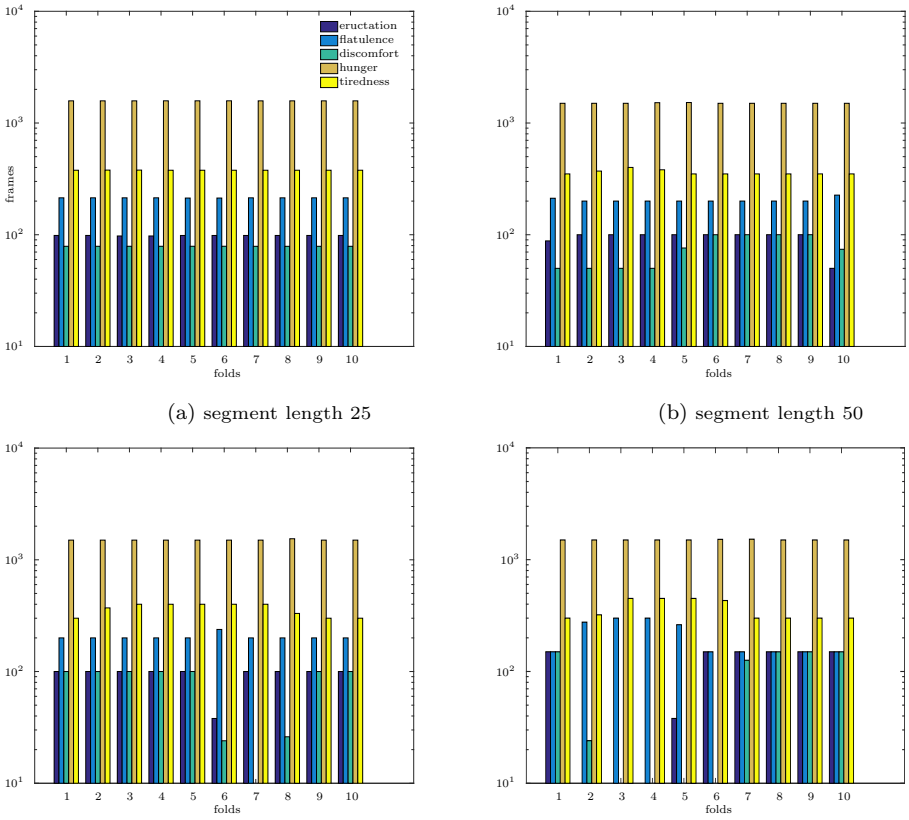


Figure 13: Sensitivity (or recall) as a function of meta-segment length. X-axis grows in a logarithmic scale.

Figure 12 shows the prediction performances, namely the F-measure computed using Equation 13, of different classifiers with an increasing segment length. The greater the size of the segments, the less the adjacent frames are found in different folds, which reduces the bias associated with an overestimation of the results.



(a) segment length 25

(b) segment length 50

(c) segment length 100

(d) segment length 150

Figure 14: 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

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 13).

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 16, 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 [2].

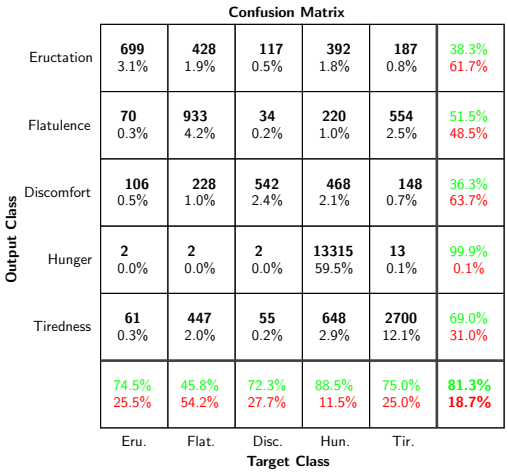


Figure 15: Confusion matrix for boosted trees classifier and a segment length of 25 frames.

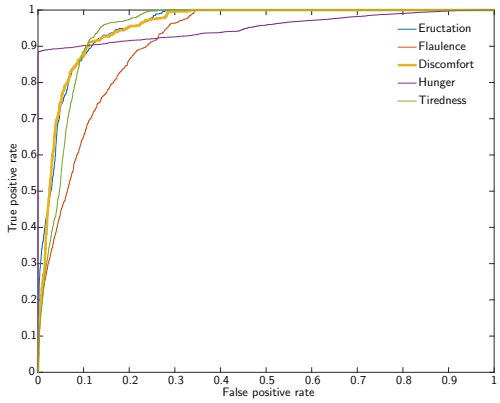


Figure 16: Receiver Operator Characteristics curve with all considered classes obtained with a segment length of 25. [best viewed in color]

9. Conclusion

This work 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.

10. Perspectives and future work

Several axes of improvement are possible for to this work. The first axe is characterized by the exploration of a set of new characteristics of the audio signal. Indeed, the most widly used acoustic feature extraction methods of current automatic speech recognition (ASR) systems are based on the assumption of stationarity [50].

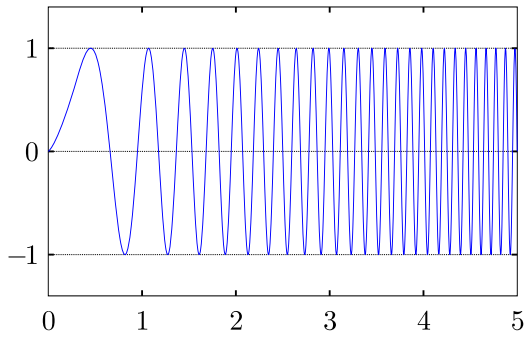


Figure 17: A linear chrip waveform: a sinusoidal wave that increases in frequency linearly over time within a short frame. (Wikipedia)

However, this assumption does not really hold even with short analysis windows as we can see in Figure 17. In [51], the authors propose to use Gammachirp filters in place of the Fourier transform which is usually used to calculate the spectral representation and hence all the characteristics that describe the spectrogram and which derive from it. More specifically, the short-time Fourier transform (STFT) which is given, in the discrete case, as follows:

$$X[n,k] \quad = \quad \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-jkm} \tag{14}$$

where w is a windowing function, can also be written as:

$$X(k) \quad = \quad \langle x(n), s_k(n) \rangle \tag{15}$$

which can also be formulated as a matrix multiplication:

$$\begin{bmatrix} X(0) \\ X(1) \\ \vdots \\ X(N-1) \end{bmatrix} = \begin{bmatrix} s_0^*(0) & s_0^*(1) & \dots & s_0^*(N-1) \\ s_1^*(0) & s_1^*(1) & \dots & s_1^*(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ s_{N-1}^*(0) & s_{N-1}^*(1) & \dots & s_{N-1}^*(N-1) \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix}$$

where and $s_k^*(n)$ corresponds to the STFT-equivalent filter of the window centered at frequency k and sample n. This is the formulation of the STFT in the form of a bank of filters. In this way, a natural re-formulation of the MFCCs would be to replace the above matrix and its $s_k^*(n)$'s by any other set of coefficents of a given filter bank.

Thus, the MFCCs could be easily adapted to the assumption of non-stationarity of the signal with, for example, the use of Gammachirp in place of the STFT-equivalent filter as proposed in [51]. These new techniques could allow us to obtain a more robust estimation of the vocal tract transfer function with the suppression of the regions between the harmonics [51].

Second axe of improvement, which is of high importance, is the experimentation of dynamic time warping (DTW). This technique calculates the similarity between time series, typically the audio signals. It is widely used in the field of speech recognition and has shown very good results in the context of classification of utterances of the same type but with variable lengths.

One of the possible applications of this technique in our case is to use it in conjunction with unsupervised learning techniques such as clustering. That being said, a more in-depth study of this technique is required, as the DTW is not a measure in its own. Indeed, it is said that a dissimilarity measure D induces a metric-space structure in the set V_L of parametric representations of words in a given vocabulary L, then the following properties must be satisfied, $\forall x, y, z \in V_L$:

- $D(x, y) \geq 0$; $D(x, y) = 0$ iff $x = y$
- $D(x, y) = D(y, x)$ (symmetry)
- $D(x, y) + D(y, z) \geq D(x, z)$ (triangle inequality)

However, the DTW does not satisfy the third condition above (triangle inequality) and thus does not induce a metric space but, rather, a semi-metric space [52]. This brings the question of how to interpret results obtained with measures that do not constitutes metrics in the context of machine learning techniques. For this reason, we consider using `Gaia`, a C++ library with python bindings which implements similarity measures and classifications on the results of audio analysis. It deals specifically with points in a semimetric space.

Another library which is considered, this time for filtering steps and features extraction, is `Essentia` which is written in C++ with Python bindings and distributed under the Affero GPLv3 license. This constitutes another axe of

development which consists in getting the existing code-base much closer to the final implementation on the electronic components.

Our choice veered to these libraries because they are written in C++, which provides considerable performance benefits [53], and the availability of Python bindings which facilitate the prototyping and accelerates experimentation of new audio analysis techniques.

Using deep neural network. We consider also to investigate the use of deep neural network in order to classify infant cries. This can be done in two ways: the first is to keep the proposed pipeline of audio filtering and features extraction and feed these vectors of characteristics to a certain type of artificial neural network, specifically a recurrent neural network (RNN) [54] that is able to deal with data characterized by temporal dependency such as audio signals; the second way of introducing neural network to this task is to build an end-to-end neural network pipeline [55, Section 1]. This way, less human expertise is required than the traditional approaches but requires much more data to get sustainable results.

Retroaction loop. Finally, one of the most important perspectives, is the development of the retroaction loop and environment feedback response. 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. Furthermore, a second part of this axe concerns the validation of the targeted response adaptation mechanism on a dataset that will be collected in real conditions.

11. References

[1] Ole Wasz-Hockert, Eero Valanne, Veli Vuorenkoski, Katarina Michelsson, and Antti Sovijarvi. Analysis of some types of vocalization in the newborn and in early infancy. In *Annales Paediatriae Fenniae*, volume 9, page 1, 1963.

[2] Priscilla Dunstan. *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, 2009.

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

[4] Aomar Osmani, Massinissa Hamidi, and Abdelghani Chibani. Platform for assessment and monitoring of infant comfort. In *2017 AAAI Fall Symposium Series on Artificial Intelligence and Human-Robot Interaction*, 2017.

[5] Grant Fairbanks. An acoustical study of the pitch of infant hunger wails. *Child Development*, pages 227–232, 1942.

[6] Orvis C Irwin and Thayer Curry. Vowel elements in the crying vocalization of infants under ten days of age. *Child Development*, pages 99–109, 1941.

[7] Ole Wasz-Höckert, Katarina Michelsson, and John Lind. Twenty-five years of Scandinavian cry research. In *Infant crying*, pages 83–104. Springer, 1985.

[8] Katarina Michelsson. Cry analysis in clinical neonatal diagnosis. *Precursors of early speech*, pages 67–77, 1986.

[9] Z Benyó, Z Farkas, A Illényi, G Katona, and G Várallyay Jr. Information transfer of sound signals. a case study: The infant cry. is it noise of an information? In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2004, pages 2774–2781. Institute of Noise Control Engineering, 2004.

[10] Taeko Tsukamoto and Yoh'ichi Tohkura. Perceptual units of the infant cry. *Early Child Development and Care*, 65(1):167–178, 1990.

[11] Chuan-Yu Chang, Chuan-Wang Chang, S Kathiravan, Chen Lin, and Szu-Ta Chen. DAG-SVM based infant cry classification system using sequential forward floating feature selection. *Multidimensional Systems and Signal Processing*, pages 1–16, 2016.

[12] M Petroni, AS Malowany, CC Johnston, and BJ Stevens. 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, pages 821–822. IEEE, 1995.

[13] MM Lewis. Infant speech: a study ofthe beginning oflanguage. *London: Kegan Paul, Trench, Trubner & Co. Ltd*, 1936.

[14] Orvis C Irwin. Infant speech: development of vowel sounds. *Journal of Speech & Hearing Disorders*, 1948.

[15] D Kimbrough Oller, Rebecca E Eilers, Dale H Bull, and Arlene Earley Carney. Prespeech vocalizations of a deaf infant: A comparison with normal metaphonological development. *Journal of Speech, Language, and Hearing Research*, 28(1):47–63, 1985.

[16] Björn Lindblom and Rolf Zetterström. Precursors of early speech. In *New York, Hampshire (Wenner-Gren International Symposium Series Vol. 44)*, 1986.

[17] Philip Lieberman, Katherine S Harris, Peter Wolff, and LH Russel. *Newborn infant cry and nonhuman primate vocalization.* ASHA, 1971.

[18] Mechthild Papoušek. Early ontogeny of vocal communication in parent–infant interactions. 1992.

[19] Hui-Chin Hsu, Alan Fogel, and Rebecca B Cooper. Infant vocal development during the first 6 months: Speech quality and melodic complexity. *Infant and Child Development*, 9(1):1–16, 2000.

[20] Orion Fausto Reyes Galaviz and Carlos Alberto Reyes Garcia. Infant cry classification to identify hypoacoustics and asphyxia with neural networks. In *Mexican International Conference on Artificial Intelligence*, pages 69–78. Springer, 2004.

[21] S Morris, I St James-Roberts, J Sleep, and P Gillham. Economic evaluation of strategies for managing crying and sleeping problems. *Archives of disease in childhood*, 84(1):15–19, 2001.

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

[23] Ian St James-Roberts. 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, 2007.

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

[25] Harriet Hiscock and Melissa Wake. Infant sleep problems and postnatal depression: a community-based study. *Pediatrics*, 107(6):1317–1322, 2001.

[26] T Weggemann, JK Brown, GE Fulford, and RA Minns. A study of normal baby movements. *Child: care, health and development*, 13(1):41–58, 1987.

[27] Danny Crichton. With mimo, mit alums are disrupting the baby nursery, onesie at a time. https://techcrunch.com/2015/01/27/with-mimo-mit-alums-are-disrupting-the-baby-nursery-onesie-at-a-time/, 2015.

[28] Felix Gillette. Baby's first virtual assistant. https://www.bloomberg.com/news/articles/2017-01-03/baby-s-first-virtual-assistant, 2017 (accessed April 28, 2017).

[29] Ali Messaoud and Chakib Tadj. A cry-based babies identification system. *Image and Signal Processing*, pages 192–199, 2010.

[30] Yousra Abdulaziz and Sharrifah Mumtazah Syed Ahmad. 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)*, pages 260–263. IEEE, 2010.

[31] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3, 1999.

[32] Efrain Rincon, Jessica Beltran, Monica Tentori, Jesus Favela, and Edgar Chavez. A Context-Aware Baby Monitor for the Automatic Selective Archiving of the Language of Infants. pages 60–67. IEEE, 2013. ISBN 978-0-7695-5087-9. URL `http://ieeexplore.ieee.org/document/6679821/`.

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

[34] María A Ruíz Díaz, Carlos A Reyes García, Luis C Altamirano Robles, Jorge E Xalteno Altamirano, and Antonio Verduzco Mendoza. Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis. *Biomedical Signal Processing and Control*, 7(1):43–49, 2012.

[35] HE Baeck and MN Souza. 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, pages 2174–2177. IEEE, 2001.

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

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

[38] G. Fant. Acoustic theory of speech production. 1960.

[39] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings of IEEE International Conference on Multimedia and Expo ICME*, volume 1, pages 113–116. IEEE, 2002.

[40] Rami Cohen and Yizhar Lavner. Infant cry analysis and detection. In *IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, pages 1–5. IEEE, 2012.

[41] Paul Boersma. 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*, volume 17, pages 97–110. Amsterdam, 1993.

[42] Onur Babacan, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich, and Thierry Dutoit. 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)*, pages 7815–7819. IEEE, 2013.

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

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

[45] Donald G Childers. *Modern spectrum analysis*. IEEE Computer Society Press, 1978.

[46] György Várallyay, András Illényi, and Zoltán Benyó. Melody analysis of the newborn infant cries. In *6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA 2009*. Firenze University Press, 2009.

[47] Nils Y Hammerla and Thomas Plötz. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1041–1051. ACM, 2015.

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

[49] Olivier Lartillot and Petri Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[50] Zoltán Tüske, Friedhelm R Drepper, and Ralf Schlüter. Non-stationary signal processing and its application in speech recognition. In *SAPA-SCALE Conference*, 2012.

[51] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Friedhelm R Drepper. Non-stationary feature extraction for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5204–5207. IEEE, 2011.

[52] Enrique Vidal Ruiz, Francisco Casacuberta Nolla, and Hector Rulot Segovia. Is the dtw distance really a metric? an algorithm reducing the number of dtw comparisons in isolated word recognition. *Speech Communication*, 4(4):333–344, 1985.

[53] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R Zapata, Xavier Serra, et al. Essentia: An audio analysis library for music information retrieval. In *ISMIR*, pages 493–498, 2013.

[54] Andrew L Maas, Ziang Xie, Dan Jurafsky, and Andrew Y Ng. Lexicon-free conversational speech recognition with neural networks. In *HLT-NAACL*, pages 345–354, 2015.

[55] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772, 2014.