Baby cry detection based on audio signals using deep neural networks

Xie, J.

Award date:
2019

Baby Cry Detection Based on Audio Signals Using Deep Neural Networks

Jiali Xie
Embedded Systems, Eindhoven University of Technology,
Email: j.xie.1@student.tue.nl

***Abstract*—Automatic baby cry detection from acoustic signals is essential for remote baby monitoring that can help parents better understand their baby's status to provide better soothing. In this thesis, we propose a method to detect baby cry in real-time at home. The major challenge of this study is the extremely imbalanced data set, which only contains 0.59% of cries from 152 hours' recordings from five babies. We used a two-step approach. The first step was to identify the segments without sound (i.e., background noise) under different recording circumstances using an adaptive thresholding method with a low computational complexity. The second step was to discriminate between cries and other sounds such as parents' talking, music, vacuum cleaner, etc. and background segments missed in the first step. A convolutional neural network (CNN) operating on log linear-filter bank (LLFB) energies of audio signals was implemented to detect cry segments. Leave-one-subject-out cross-validation was used. On average, an accuracy (across subjects) of 98.64 $\pm$ 0.66% in the background detection was achieved, where only 2 out of 3209 cry segments were missed. For cry detection, we achieved an average accuracy of 0.92 $\pm$ 0.04 and a Cohen's Kappa coefficient of 0.74 $\pm$ 0.05.**

***Index Terms*—Baby cry detection, audio signal processing, convolutional neural network, audio classification.**

I. Introduction

CRY is an essential way for babies to communicate. Many studies have shown that some information from a baby, such as health condition, emotion, hunger, and disability, can be interpreted by analyzing the acoustic characteristics of their cries [1]. Furthermore, cry is a biological alarm to express the needs and wants of the baby and notify the caregivers to take actions [2]. It is imperative for caregivers to respond to the cries of babies so that the needs of babies can be satisfied in time. However, it is very difficult for caregivers (e.g. parents) to pay attention to their babies all the time. In addition, first time parents confuse cries and other baby sounds such as moans, and they may not know the proper timing to take care of their baby. A tool to assist them to recognize cries of their baby is preferred. Therefore, a reliably baby cry detection algorithm allows for helping parents to provide improved care.

In the past decades, many studies with regard to baby cry have been conducted, among which the two main study perspectives are classification and/or interpretation of cries (hungry or not hungry, healthy or unhealthy, etc.) and cry recognition/detection using audio signals [3].

Some studies analyzed acoustic signals of baby cries trying to automatically interpret or identify cries associated with certain status. For example, a study of classifying cry of babies with or without hearing disorders by analyzing the fundamental frequency and the dominant frequency of the corresponding audio signals was done by Vrallyay [4] and a preliminary accuracy of 79% was reported. Hidayati et al. [5] developed a system to recognize baby cry with different emotions such as pain, sadness, and fear. In this system, pitch and formants were used to as features and a K-means algorithm was applied, where they achieved an accuracy of 90%. Ntalampiras [6] investigated several machine learning algorithms, such as Multi-layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Reservoir Network (RN), to classify pathological states of cry using audio-based features, such as temporal modulation features and Mel frequency cepstral coefficients (MFCC), etc. An RN classifier had the best performance in terms of recognition rate ($\sim$94.5%). In addition to that, Al-Azzawi [7] proposed an automatic system (accuracy = 96%) to recognize baby cry of several different physiological statuses and diseases. Features were extracted with a fuzzy transform and then fed into an MLP artificial neural network (ANN). Azlee et al. [8] used MLP ANN combined with a set of selected MFCC features to classify healthy babies or babies with hypothyroidism, achieving an accuracy of 88.94%.

Regarding cry detection. Vrallyay et al. [9] used a short-time energy function and a harmonic product spectrum method to retrieve information on energy contact and spectral content from baby audio signals. The information was further analyzed to find cry periods. Lavner et al. [10] introduced a k-nearest neighbors algorithm to detect baby cries at an achieved detection rate varying from 15% to 100% depending on the signal noise ratio changing from -5 to 40. Pitch related parameters, MFCC, and short-time energy parameters were extracted from the signal as the input of the algorithm. Kim et al. [11] proposed a method to distinguish between cry and non-cry sounds. In that work, features based on weighted segment-based two-dimensional linear frequency cepstral coefficients (LFCC) were computed and a Gaussian mixture model was used where an error rate of 4.42% was reported. In a recent study, to detect baby cries in a domestic environment, Lavner et al. [12] compared two machine learning algorithms (low-complexity logistic regression and CNN) and they concluded that the CNN classifier (detection rate of 82.55% with a fixed false positive rate of 1%) outperformed the other.

Although many studies have shown the feasibility of audio-based baby cry detection, there are still remaining problems. For example, the data used in previous studies were mostly

recorded in a control environment [4] [6] [7], and more importantly, manually selected with a very good balance between cries and other sounds [7] [8] [10], or with a short recording duration [11] [12]. However, there can be many challenges in practical use at home. For example, data can be strongly imbalanced where baby cries only account for around 1-2 hours per day in total, various noise (parent talking, music, car engine, etc.) can present in combination of cries, and the microphone placement can vary a lot, depending on the own preference and the layout of the baby's room. Hence, the algorithms described in those studies might not be suitable or optimal for continuous long-term (24/7) baby monitoring for home use in real-life applications. In this study, we aimed to tackle these problems by using a data set containing continuous recordings (up to 24 hours per day per baby) collected in a home scenario.

In this thesis, we present a method to automatically detect baby cries in real-time based on Convolutional Neural Networks (CNN) in a domestic environment. The data used in this project contains more than 95% background segments (without sound or with little noise). It is therefore very time consuming if all the data are processed with the cry detection algorithm, where CNN is expected to be used. To tackle this problem, thus, we propose a two-step approach. The first step is background detection using a computational effective algorithm, which aims at removing background segments, ideally without missing cry segments. In this step, the challenge is the variability between babies and between recordings caused by different placements and types of the microphones. Therefore, methods using a fixed amplitude threshold would not be appropriate. In this work, we designed an adaptive thresholding method to identify background segments. This method, compared with a machine learning algorithm, is expected to have a much lower computational complexity. The second step is cry detection. CNN is widely used in classification problems and has achieved many practical successes. Compared to traditional machine learning methods, which require a lot of human efforts and domain expertise to extract good features, CNN can potentially make full use of the input data and learn important features automatically [14]. Thus, CNN is chosen in this work. Log linear-filterbank (LLFB) energies are employed to represent audio signals as the input of CNN. Details of the two-step approach will be introduced later.

II. DATA SET

A. Data Recording and Subjects

For each baby, the audio data used for this study was collected using a Philips baby monitor at the baby's own bedroom in 2016 or 2012. There was no specific requirement for the placement of the device. Age of the baby subjects was between 3 and 18 months. All the babies were healthy during the date collection. Recordings from five babies were annotated, and these babies were tagged as B1, B2, B3, B4, and B5. The recording time of each baby is shown in Table I. The video data were converted into WAV audio files with a sampling frequency of 16000Hz.

TABLE I: Recording time of audio signal of each baby.

Baby	B1	B2	B3	B4	B5
Time(h)	23.89	47.89	45.03	23.95	11.0

B. Event-based Annotation

Audio signals were annotated as one of nine categories including moan, cry, other baby sound, other human sound, non-human sound, moan with other sound, cry with other sound, background, and other baby sound with other sound (Table II). Although in this work we focused on baby cry, annotating more categories can help analyze and understand what lead to incorrect detection.

TABLE II: Annotation categories and labels.

M	Moan
C	Cry
OB	Other baby sound (Heavy breathing, coughing, gurgling, laughing, hiccuping, ...)
OH	Other human sound (Parent, sibling, nanny, ...)
NH	Non-human sound (car passing, baby hitting arm against bed, dog barking, stairs creaking, babys sound which do not come from the throat like farting,)
M+	Moan with either OH or NH at the same time
C+	Cry with either OH or NH at the same time
BG	Background (No sound and able to score)
OB+	Other baby sound with either OH or NH at the same time

Annotation was done by a human annotator through hearing and visual inspection of audios, with assist of videos when necessary. Note that the data without baby visible in the video was discarded before annotation. The annotation for the raw data was done based on crying events instead of time segments with a fixed length because a cry should be considered as a continuous event. The start time and the end time of each event were annotated. The annotation rules are listed in the following.

1) If the sound wave is a straight line, we annotate it as BG.
2) If there is sound, but its duration is less than 0.5s, we annotate it as BG.
3) If the interval between two periods of sound is less than 0.5s, then they are considered as one period (event).
4) If there are many short periods of noise, and there is no baby sound in it, but the intervals are different($<$3s), we annotate it as a long period of noise (NH).
5) If there are many short baby's sounds ($<$1s), we annotate them as one event of baby sound. Note that this event can be cry, moan, and other baby sound.
6) If OH and NH exist together in a period, we annotate the period as the one that dominate (has a higher percentage of) the period.

C. Segmentation and Labelling

Since the event length can be inconsistent, it is challenging to use them as inputs of classifiers. Therefore, a segmentation step is required to equalize the length of the input. After segmentation, labels should be given to all the segments according to the original annotation. Fig. 1 shows boxplot of the time duration of events of all baby sounds. It shows that the

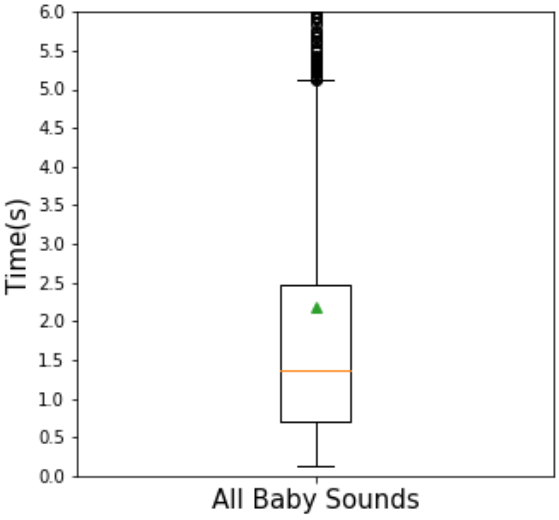


Fig. 1: Time duration of events of all baby sounds (outliers bigger than 6 are not shown).

median duration was approximately 1.3-1.4s so that we chose 1.5s as the segmentation length. Overlap of 0.5s was added to each segment to avoid losing signal information especially at the edge.

In addition, rules about relabeling these 1.5s segments are given in the following.

1) If BG occupies $\geq$1s, we label it as BG.
2) If OB or OB+ has a time of $\geq$0.5s, and it is longer than any other baby sounds (cry or moan) in the same period, then it is labeled as OB or OB+. This is because the median duration of most other baby sounds was around 0.5s.
3) Otherwise, we label it with the category that dominates the 1.5s segment.

TABLE III: Amount and percentage of classes.

Class	Segments Amount	Percentage	Percentage without BG
C	2435	0.72%	15.70%
M	1836	0.55%	11.84%
OB	1261	0.37%	8.13%
OH	1263	0.38%	8.14%
NH	6461	1.92%	41.66%
M+	699	0.21%	4.51%
C+	1373	0.41%	8.85%
BG	321265	95.39%	
OB+	208665	0.05%	1.17%
Sum	545440	100%	100%

Table III shows the percentages and amounts of segments of all the categories. It can be seen that the data set is extremely imbalanced, where most of the segments were BG (-95.39%) and only less than 5% contained sounds. Among the data without BG, 24.55% were baby cry segments (C and C+, 3209 segments). In this work, we merged C and C+ as baby cry (CRY), and all other sounds (M, OB, OH, NH, M+, and OB+) as non-baby cry (NCRY). Thus, three classes were considered including CRY, NCRY, and BG.

III. Method: Background Identification

As stated, in this study we proposed the use of an adaptive thresholding method to identify BG segments.

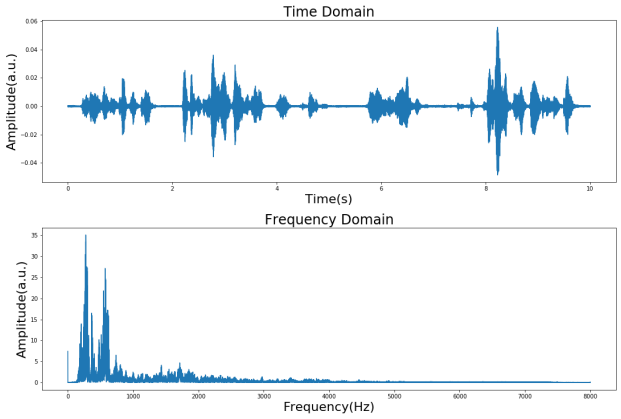
First, the baseline of background noise should be computed, which can differ from recording to recording due to difference in microphone placement and type. A sample entropy measure was used attributed to its capability of quantifying signal complexity or randomness [13], which is expected to separate noise and sound. It is important to select parameters of sample entropy, including the sample length m and tolerance r. Recommended in [15], m should be set to 2 and r should be 0.2 times the standard deviation of the signal. A greater value of sample entropy indicates less self-similarity, more complex or more noise. Given an audio recording, a threshold T was determined to make sure that we could correctly find BG segments to derive a baseline of background noise of the recording (BL). The baseline measure (derived from these BG segments) should then be used to normalize each recording. In other words, a "global" threshold for identifying BG over the recording should be adapted by the baseline measure. Based on T, 50 identified BG segments in the beginning of each recording were included and the baseline measure BL was the median of the mean absolute values of all the 50 segments.

For each 1.5s segment, signal was divided into 15 0.1s sub-segments $s_1, s_2, ..., s_i, ..., s_{15}$. We then calculated the volume of each sub-segment. For a given sub-segment s_i, its corresponding relative volume is $v_i = 20\log_{10}(\mu/BL)$, where μ is the absolute mean value of the signal in this sub-segment and BL was used to adapt the volume, Then we ranked the volume values of all the sub-segments in a descending order and computed the sum of the top 5 ranked volume values S_v (of 0.5s, equivalent to the boundary duration of annotating BG). A threshold T_v was eventually used to identify BG segments, where a 1.5s segment is classified as BG if $S_v < T_v$, otherwise as non-background (CRY or NCRY). All parameters used in this algorithm was experimentally optimized by maximizing the BG identification performance. It is also important to note that, since the ultimate goal is to detect baby cry (with a specific cry detection algorithm following the BG identification algorithm), it is acceptable to misclassify some BG segments as non-BG segments while we should not misclassify too many cry segments as BG segments.

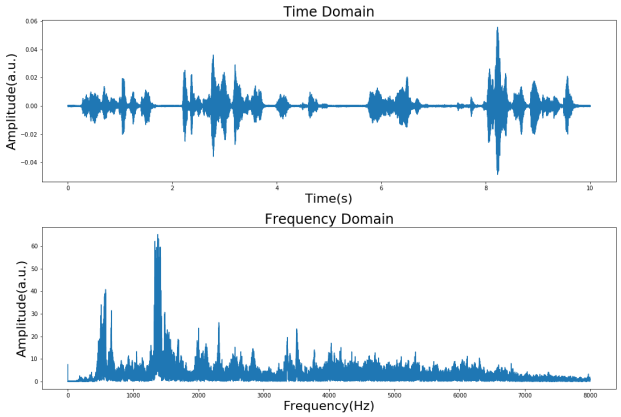
IV. Method: Cry Detection

A. Baby Cry Analysis

Baby cry signals contain typical spectral characteristics compared with other human sounds from adults. Fig. 2 shows the time domain and spectral amplitude of a 10-s signal from a mother's talking and a baby' cry, respectively. From the graphs, it can be seen that the pitch of the mother's talking signal was around 200Hz with clear harmonics solely in the relatively low-frequency range. However, the pitch of the baby cry was higher (around 500Hz), with clear harmonics in both low and high-frequency ranges.



(a) A mother's talking



(b) A baby's cry

Fig. 2: Time domain and spectral amplitude of a 10- s signal from a mothers talking and a babys cry,

B. Representation of Audio Segments

Mel-scale features are widely used in voice activity detection and speech recognition [16]. The Mel-scale is designed to mimic the non-linear human perception of sound. The formula to convert Hertz (f) to Mel (m) is defined as

$$m = 2595 \log_{10}(1 + \frac{f}{700}). \quad (1)$$

MFCC and log Mel-filterbank (LMFB) energies are two popular sets of features in the Mel-scale. Compared with LMFB energies, MFCC includes a discrete cosine transform (DCT) during its computation, in order to decorrelate the filter bank coefficients. In CNN, however, correlated data can be a benefit [12]. Additionally, some underlying non-linear information would be undesirably overlooked, as DCT is a linear transformation. Even so, LMFB energies still do not fit well for cry detection. Fig. 4 shows the Mel-scale filters associated with LMFB energies, from which it can be observed that the distance between filters is larger in higher frequencies. The energies after applying Mel-scale filters has a higher resolution in the lower-frequency range and lower resolution in the higher-frequency range. This property can lead to "blurry" representation in the high-frequency range where some important information might be missing. Log linear-filterbank (LLFB) energies can solve this problem. Fig. 4 illustrates the linear-scale filters, where the filters have the same distance in between. This means that LLFB energies give same "emphasis" to both low and high frequencies. Through the analysis of baby cry acoustics, we have learned that baby cry signals have components in the high-frequency range where LLFB energies should have a better representation than LMFB energies. Fig. 5 shows the LMFB energies and LLFB energies of a 10-s baby cry audio signal. We can see that, in the high-frequency range (e.g. >2000 Hz), the LLFB energies are clearer than the LMFB energies. Given these reasons, we applied LMFB energies to represent audio signals for each 1.5s segment and the corresponding cry detection results will be compared with those using LMFB energies.

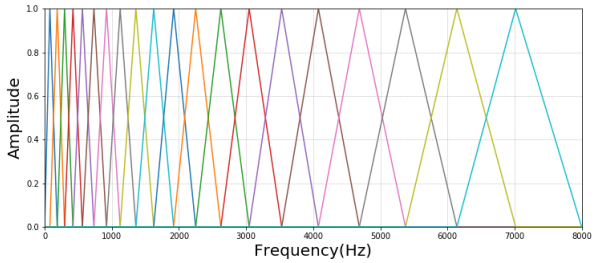


Fig. 3: Mel-scale filters (LMFB).

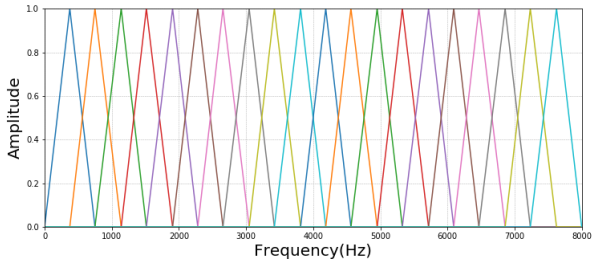
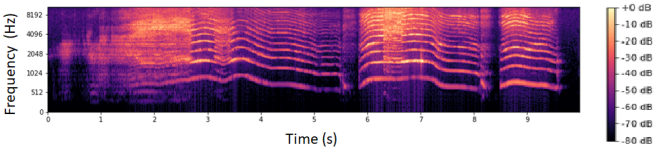
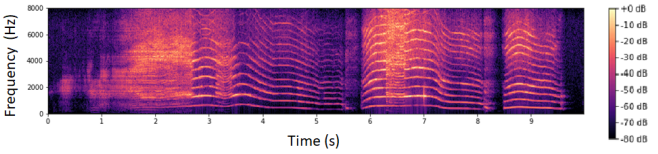


Fig. 4: Linear-scale filters (LLFB).



(a) LMFB energies



(b) LLFB energies

Fig. 5: LMFB energies and LLFB energies for 10s signal of cry.

Before feeding the LLFB energies to CNN, pre-processing is required to select the frequency range and re-scale the input. Frequency larger than 6000 Hz was left out where the cutting

boundary was experimentally chosen (which led to the best cry detection performance). After that, z-score normalization was used to re-scale LLFB energies for each segment to have zero mean and unit standard.

C. Convolutional Neural Networks

CNN was employed to classify CRY and NCRY or BG. The architecture of our networks is depicted in Fig. 6, consisting of two convolutional paths, one fully connected layer, and one output layer. Instead of using square filters, rectangle filters with size 2×10, 2×8, 10×2 and 8×2 were selected here to capture information from both frequency domain and time domain. The columns and rows of each input LLFB image correspond to time and frequency, respectively. Thus, "wide" filters (2×10, 2×8) can have more information from frequency domain and "tall" filters (10×2, 8×2) can have more information from the time domain.

The left convolutional path using "wide" filters has three convolutional layers, where each layer has one 2D convolutional layer with 'relu' as the activation function, one 2D max pooling layer, and one dropout layer with a dropout rate of 0.2. The right convolutional path with "tall" filters has a similar design except for the number of filters. The left path has more filters than the right one, the reason is that we think frequency features are more valuable than time features based on two facts. The first fact is that cry events with the length longer than 1.5s were randomly cut into different segments during the segmentation part, which make the time features become discontinuous. The second one is that the inconsistent length of cry events may differ the time features from each other.

The outputs of two convolutional parts were then flattened and concatenated together in order to be linked to the fully connected layer. The last part of the model is an output layer with a 'sigmoid' activation function.

To train this network, stochastic gradient descent (SGD) was used as the optimizer with a learning rate of 0.01. Binary cross entropy was chosen to be the loss function. 32 and 30 were used as the batch size and epoch number, respectively.

V. Evaluation Criteria

Leave-one-subject-out cross-validation was used as evaluation method, which means that four babies were treated as a training set while one baby was treated as a test set. It is more reasonable to use this method to generate the results because it is almost impossible for cry signals to exist in the training set when applying this algorithm in the real world.

A. Confusion Matrix

A confusion matrix is often used to present the performance of a supervised learning algorithm. Table IV describes the confusion matrix, where each row stands for the instances of a predicted class while each column stands for the instances of an actual class [17]. Some evaluation metrics that can be computed based on the confusion matrix are also shown in the table.

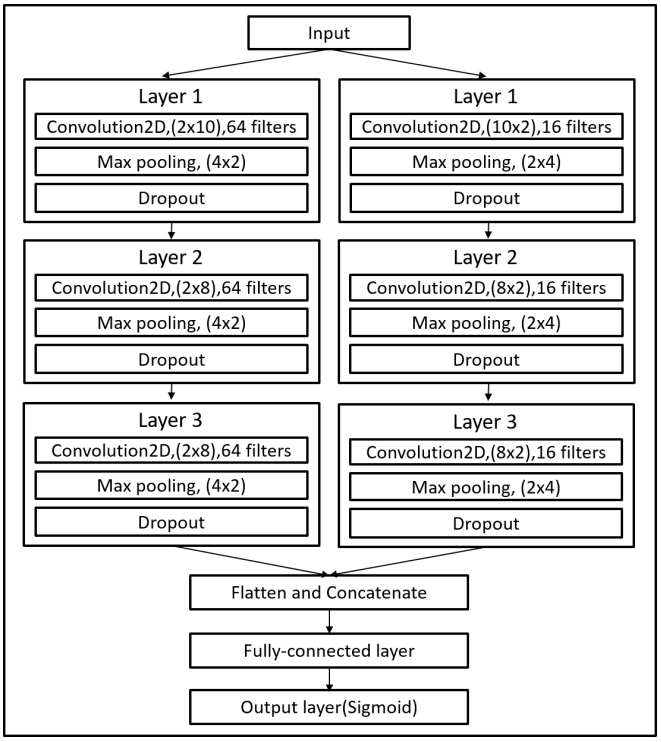


Fig. 6: The CNN architecture for cry detection.

B. Precision-Recall

A Precision-Recall (PR) curve is a helpful measurement of success of classification when the classes are strongly imbalanced [20]. Precision (TPR) is the rate of real true cases out of predicted true cases, and recall indicates how many true cases are returned. The precision-recall curve demonstrates the trade-off between precision and recall by changing the decision-making threshold. A bigger area under the precision-recall curve (AUCPR) means a better performance, which also means a high precision and a high recall.

C. Cohen's Kappa

The Cohen's Kappa coefficient is used to measure agreement between two raters (here between ground truth and predicted result) [18]. It is a robust measure for imbalanced data due to the fact that it is computed by taking the chance agreement into account. Kappa can be calculated as [19]

$$k = \frac{p_0 - p_e}{1 - p_e}, \tag{2}$$

where p_0 is the overall agreement (identical to accuracy) and p_e is the proportion of the raters that are expected to agree by chance. Landis and Koch defined an interpretation system for the value of k [19]:

- <0.00: poor agreement
- 0.01-0.20: slight agreement
- 0.21-0.40: fair agreement
- 0.41-0.60: moderate agreement
- 0.61-0.80: substantial agreement
- 0.81-1.00: almost perfect agreement

TABLE IV: Confusion matrix.

Total		True condition		
		Positive	Negative	
Predicted condition	Positive	True positive (TP)	False positive (FP)	Positive predictive value (precision, PPV) = $\frac{TP}{TP+FP}$
	Negative	False negative (FN)	True negative (TN)	Negative predictive value (NPV) = $\frac{TN}{FN+TN}$
		True positive rate (TPR, recall) = $\frac{TP}{TP+FN}$	True negative rate (TNR, specificity) = $\frac{TN}{FP+TN}$	Accuracy = $\frac{TP+TN}{Total}$

TABLE V: Confusion matrix of background detection.

		Ground Truth									
		B1		B2		B3		B4		B5	
		BG	Other	BG	Other	BG	Other	BG	Other	BG	Other
Detection	BG	44011	877	87279	220	101275	256	51795	331	34005	231
	Other	98	1352	373	2359	2073	4185	335	1967	21	3732

D. Onset Detection Rate and False Cry Alarm

Onset detection rate (ODR) is a metric to assess the success rate of detecting a cry event given a certain time delay. For each cry event, we examine the time delay (based on the start time of this cry event) between the annotation and the detection. ODR_0 is associated with no time delay, and ODR_1, ORD_2, and ODR_3 correspond to a time delay of 1s, 2s, and 3s, respectively, from the annotated cry event to the detected cry event. When the time delay is more than 3s, we consider that the cry event is missed by our cry detection algorithm (ODR_m).

False cry alarm is a metric proposed to quantify the misclassifications of the algorithm with a certain level of tolerance. A false cry alarm is a segment being incorrectly detected as cry but without any cry segments before and after 3s. In fact, the adjacent moan and other baby sound before and after cry segments can be easily misclassified as cry. These misclassifications are not counted as false cry alarms within the relax time range [-3s,+3s]. This is assumed to be likely acceptable for parents to respond to the cry alarms within this time range.

ODR and cry alarm can help us know how many cry events are seized in time while how many incorrectly events are reported.

VI. Result

TABLE VI: TPR, accuracy and Kappa of background detection of each baby.

	TPR	Accuracy	Kappa
B1	99.78%	97.90%	0.72
B2	99.57%	99.34%	0.88
B3	97.99%	97.83%	0.77
B4	99.36%	98.78%	0.84
B5	99.93%	99.34%	0.96
Average	99.33 $\pm$ 0.70%	98.64 $\pm$ 0.66%	0.83 $\pm$ 0.08

Table V shows the confusion matrix of background identification. The numbers of TP and TN indicate that most of background segments were correctly identified. TPR, accuracy, and Kappa are presented in Table VI, where it shows that 99.33 $\pm$ 0.70% of background segments were correctly identified in average. The average Kappa was 0.83 $\pm$ 0.08 (almost perfect agreement between the ground-truth and our algorithm). The corresponding performance associated with baby sounds and cry is presented in Table VII, in which TPR of cry stands for the proportion of cry segments remained after removing identified background segments, and the TPR of all baby sound means the proportion of baby sounds segments left. It can be seen that, on average, 99.92 $\pm$ 0.11% of the cry segments and 96.30 $\pm$ 2.93% of all the baby sound segments were remained for further cry detection when 99.33 $\pm$ 0.11% of the background segments were removed. Table VIII gives the percentages of all the classes when excluding the identified background segments, where 19.44% were cry (C, C+) segments and the proportion of background segments reduced from 95.39% to 17.58% compared with Table II.

TABLE VII: TPR of cry and all baby sounds for background detection (TPR of cry stands for the proportion of cry segments remained after removing identified background segments, and the TPR of all baby sound means the proportion of baby sound segments left).

	TPR of cry (C, C+)	TPR of all baby sounds (M, C, OB, M+, C+, OB+)
B1	100.00%	96.00%
B2	99.71%	98.31%
B3	100.00%	98.47%
B4	100.00%	90.72%
B5	99.91%	98.01%
Average	99.92 $\pm$ 0.11%	96.30 $\pm$ 2.93%

TABLE VIII: Amount and percentage of classes after background detection.

Class	Segments Amount	Percentage
M	2373	14.39%
C	1834	11.12%
OB	1116	6.77%
OH	1154	7.00%
NH	4892	29.66%
M+	698	4.23%
C+	1373	8.32%
BG	2900	17.58%
OB+	155	0.94%
Sum	16495	100.00%

TABLE IX: Confusion matrix of cry detection.

		Ground Truth									
		B1		B2		B3		B4		B5	
		Cry	Other	Cry	Other	Cry	Other	Cry	Other	Cry	Other
Detection	Cry	692	51	312	200	478	84	145	25	866	56
	Other	161	546	36	2184	251	5445	48	1967	218	2613

TABLE X: Kappa, AUCPR, and accuracy of cry detection by using LMFB (rectangle filters), LLFB (rectangle filters), LLFB (3x3 filters), and LLFB (5x5 filter) respectively.

		B1	B2	B3	B4	B5	Average
Kappa	LMFB (rectangle)	0.51	0.60	0.70	0.77	0.63	0.64 ± 0.09
	LLFB (rectangle)	**0.71**	0.68	**0.71**	**0.78**	**0.81**	**0.74 ± 0.05**
	LLFB (3x3)	0.53	0.70	0.68	0.65	0.80	0.67 ± 0.09
	LLFB (5x5)	0.51	**0.71**	0.70	0.74	0.77	0.69 ± 0.09
AUCPR	LMFB (rectangle)	0.95	0.77	0.85	**0.88**	0.89	0.87 ± 0.06
	LLFB (rectangle)	**0.97**	**0.83**	**0.86**	**0.88**	**0.95**	**0.90 ± 0.05**
	LLFB (3x3)	0.96	**0.83**	0.81	0.85	0.93	0.88 ± 0.06
	LLFB (5x5)	0.95	0.81	0.81	0.86	0.92	0.87 ± 0.06
Accuracy	LMFB (rectangle)	0.75	0.88	**0.95**	**0.97**	0.83	0.88 ± 0.08
	LLFB (rectangle)	**0.85**	0.91	**0.95**	**0.97**	**0.93**	**0.92 ± 0.04**
	LLFB (3x3)	0.76	**0.92**	0.94	0.96	0.92	0.90 ± 0.07
	LLFB (5x5)	0.74	**0.92**	0.94	**0.97**	0.90	0.89 ± 0.08

Note: the rectangle filter size is 2x10, 2x8, 10x2, or 8x2 (see Fig. 6).

TABLE XI: Performance of cry detection in terms of cry alarm (false cry alarm with moan, false cry alarm with baby sound, total false cry alarm, true cry alarm, total cry alarm and total processed segments).

	False cry alarm (with moan (M, M+) in the range)	False cry alarm (with other baby sound (OB, OB+) in the range)	False cry alarm without any baby sound in the range	Total false cry alarm	True cry alarm	Total cry alarm (true cry alarm + total false cry alarm)	Total processed segments
B1	13	1	0	14	729	743	85872
B2	86	30	24	140	372	512	172113
B3	34	1	2	37	372	512	161832
B4	13	0	2	15	155	170	86089
B5	11	2	0	13	909	921	39540

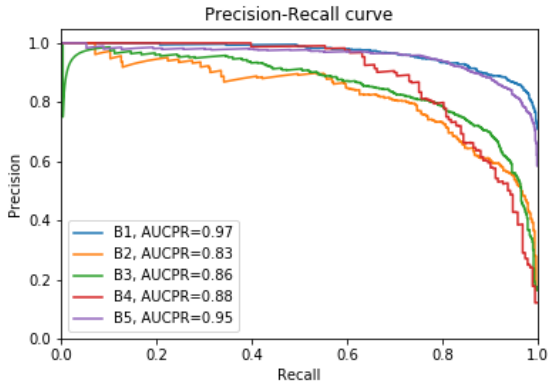


Fig. 7: Precision-Recall curve for all the babies.

The results of cry detection in Table IX, Table X and Fig. 7 were based on the data after background identification. The confusion matrix in Table IX shows that the classification results of all the babies correspond to a low FP except for baby B2 and most segments were correctly recognized. Kappa, AUCPR, and accuracy for cry detection of each baby are presented in Table X. A substantial agreement at an average Kappa of 0.74 ± 0.05 was achieved. Accordingly, the average AUCPR, and accuracy of 0.90 ± 0.05, and 0.92 ± 0.04 were obtained, respectively. Fig. 7 compares the PR curve between babies, and it shows that baby B1 and B5 had a bigger AUCPR than the others. Table X compares the cry detection performance between using LMFB energies with rectangle filters, LLFB energies with rectangle filters, LMFB energies with 3×3 filters and LMFB energies with 5×5 filters. In general, LLFB energies with rectangle filters had the best cry detection results.

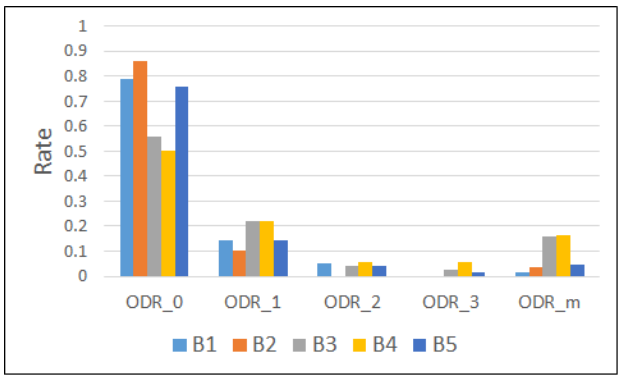


Fig. 8: ODR_0, ODR_1, ODR_2, ODR_3, and ODR_m for each baby.

The ODR_0, ODR_1, ODR_2, ODR_3, and ODR_m for each baby are shown in Fig. 8, from which it can be seen that most cry events were detected immediately when it happened (0s) or in a very short time (1s). Notably, baby B3 and B4 had a relatively high ODR_m at ∼15%. The cry alarms are

presented in Table XI. In general, our algorithm generated a small number of false cry alarms except for baby B2. Most false cry alarms were associated with moan (M, M+) or other baby sounds (OB, OB+).

VII. Discussion and future work

A. Discussion and Conclusion

From Table V, we see that many other segments were incorrectly classified as background for baby B1 while many background segments were incorrectly recognized as other sounds for baby B3. Upon a close look, we found that these misidentified segments for baby B1 contained longer and higher-amplitude noise (caused by e.g.ventilation) compared with those for the other babies. However, it is acceptable and even desirable to classify these segments as background noise. Through listening to the misclassified segments of baby B3, we found that many of their annotations were not correct due to the low scale of the audio signal. This is because we first visualized the audio signal so that those noise or sounds were overlooked. Therefore, it is important to zoom in to these segments during annotation.

From the results of cry detection, we can see that the use of LLFB led to a better performances than LMFB. This is because the high-frequency components are important in cry detection, where baby cries often correspond to a relatively higher frequency that other sounds. By using rectangle filters, we achieved a slightly better cry detection performance than square filters. Comparing the results between babies in Table IX and Table X, baby B2 has a relatively low precision. The reason found by listening to the audio recording is that it is difficult to distinguish cry and moan segments where the annotations were not confident. It is also shown in Table XI that most false alarms are caused by moan and other baby sounds for baby B2. Additionally, the higher miss rate with regard to ODR_m for baby B3 and B4 compared with the other babies was due to the difficulty in separating cry and moan and in identifying the cry events with a relatively short duration (e.g. $<$1s). Misclassifications in cry detection can also be caused by variability between babies with regard to their cries, which can depend on age, gender, etc. Therefore, a larger data set with more babies should be included to further improve (and generalize) the algorithm.

According to the results and analysis of background identification, we can conclude that the algorithm can adapt to different placements and types of microphones and may have better performance than human annotation based on the data from 5 babies. For the data in this work, the adaptive thresholding method can help identify $\geq$ 99% of the background segments. In general, we achieved a good performance with an average Cohen's Kappa of 0.74 across the five babies in cry detection, which is significantly better than that in a previous study [21].

B. Future Work

The adaptive thresholding algorithm for background identification described in this work was developed based on five babies only. In the future, more data with different placements and types of microphones should be used to further verify the effectiveness of this algorithm.

For cry detection, it is suggested to further improve the annotation of baby sound events, in particular in distinguishing between cry and moan which is one of the major barrier of annotation. In addition, studies should be focused on dealing with the false detections caused by variability between babies in terms of crying. A potential method to tackle this problem is to train a more generalized neural network model on a larger data set with more subjects (babies). Furthermore, adversarial neural networks can be applied to extract a set of features that is optimal in characterizing cries without being sensitive to the differences between babies.

References

[1] Saraswathy, J., M. Hariharan, Sazali Yaacob, and Wan Khairunizam. "Automatic classification of baby cry: A review." In Biomedical Engineering (ICoBE), 2012 International Conference on, pp. 543-548. IEEE, 2012.

[2] Singer, Lynn T., and Philip Sanford Zeskind. Biobehavioral assessment of the baby. Guilford Press, 2001.

[3] Cabon, Sandie, Fabienne Poree, Antoine Simon, Olivier Rosec, Patrick Pladys, and Guy Carrault. "Video and audio processing in paediatrics: a review." Physiological measurement (2019).

[4] Vrallyay Jr, Gyrgy. "Infant cry analyzer system for hearing disorder detection." spectrum 18, no. 19 (2004): 20-21.

[5] Hidayati, Rahmat, I. Ketut Eddy Purnama, and Mauridhi Hery Purnomo. "The extraction of acoustic features of infant cry for emotion detection based on pitch and formants." In International Conference on Instrumentation, Communication, Information Technology, and Biomedical Engineering 2009, pp. 1-5. IEEE, 2009.

[6] Ntalampiras, Stavros. "Audio pattern recognition of baby crying sound events." Journal of the Audio Engineering Society 63, no. 5 (2015): 358-369.

[7] Al-Azzawi, Nemir Ahmed. "Automatic Recognition System of Infant Cry based on F-Transform." International Journal of Computer Applications 102, no. 12 (2014).

[8] Zabidi, Azlee, Wahidah Mansor, Lee Yoot Khuan, Ihsan Mohd Yassin, and Rohilah Sahak. "Classification of infant cries with hypothyroidism using multilayer perceptron neural network." In 2009 IEEE International Conference on Signal and Image Processing Applications, pp. 246-251. IEEE, 2009.

[9] Vrallyay, Gyrgy, Andrs Illnyi, and Zoltn Beny. "Automatic infant cry detection." In MAVEBA, pp. 11-14. 2009.

[10] Cohen, Rami, and Yizhar Lavner. "Infant cry analysis and detection." In 2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel, pp. 1-5. IEEE, 2012.

[11] Kim, Myung Jong, Younggwan Kim, Seungki Hong, and Hoirin Kim. "ROBUST detection of infant crying in adverse environments using weighted segmental two-dimensional linear frequency cepstral coefficients." In 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1-4. IEEE, 2013.

[12] Lavner, Yizhar, Rami Cohen, Dima Ruinskiy, and Hans IJzerman. "Baby cry detection in domestic environment using deep learning." In 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), pp. 1-5. IEEE, 2016.

[13] Richman, Joshua S., and J. Randall Moorman. "Physiological time-series analysis using approximate entropy and sample entropy." American Journal of Physiology-Heart and Circulatory Physiology 278, no. 6 (2000): H2039-H2049.

[14] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436.

[15] Lake, Douglas E., Joshua S. Richman, M. Pamela Griffin, and J. Randall Moorman. "Sample entropy analysis of neonatal heart rate variability." American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 283, no. 3 (2002): R789-R797.

[16] Jung, Youngmoon, Younggwan Kim, Hyungjun Lim, and Hoirin Kim. "Linear-scale filterbank for deep neural network-based voice activity detection." In 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), pp. 1-5. IEEE, 2017.

[17] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

[18] Pontius Jr, Robert Gilmore, and Marco Millones. ”Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment.” International Journal of Remote Sensing 32, no. 15 (2011): 4407-4429.
[19] Landis, J. Richard, and Gary G. Koch. ”The measurement of observer agreement for categorical data.” biometrics (1977): 159-174.
[20] Boyd, Kendrick, Kevin H. Eng, and C. David Page. ”Area under the precision-recall curve: point estimates and confidence intervals.” In Joint European conference on machine learning and knowledge discovery in databases, pp. 451-466. Springer, Berlin, Heidelberg, 2013.
[21] Fu, Qinyi. *Infant Cry Detection Based on Audio Signal*, Master's thesis, Eindhoven University of Technology, Eindhoven, the Netherlands, 2017.