A Low-Cost Intelligent Hardware System for Real-Time Infant Cry Detection and Classification

PradeepPathirana, SagaraSumathipala

Abstract*—Cry of an infant serves as the main communication language to seek the attention of their caretakers. Acoustic characteristics of cries provide insights into the physiological and psychological states of the infants. To understand the reason behind the cries, caretakers pay attention to acoustic characteristics like tone, pitch, and loudness, etc. Infant cry classification has gained importance in both commercial and medical fields due to its applications such as baby monitoring and non-invasive diagnosis of health conditions of newborns in early stages. This paper discusses the implementation of a low-cost hardware device for real-time cry classification. Further, this presents the use of time and frequency domain features to detect cry words and identify the reasons. The proposed solution covers the use of classification techniques like artificial neural networks and k nearest neighbors to gain accuracy figures over 90% and 70% for cry detection and classification respectively while maintaining the resource utilization at a minimum level to keep hardware solution simple and low cost.

***Keywords:* Cry classification, Voice Activity Detection, Short-Time Energy, Short-Time Zero-Crossings, Mel-frequency cepstral coefficients, Artificial Neural Networks, K-Nearest Neighbors..**

I. INTRODUCTION

It is widely known that cry acts as the primary mode of communication of the infants to seek the attention of their parents [1]. Through the cries, infants express their basic emotions like hunger, tiredness, pain, and discomfort. Parents can often understand the cause of the cry based on its nature analyzing the tone, pitch, intensity, and duration, etc. Analysis of infant cry signals has become a major research area in medical fields as well to diagnose brain damages, hearing impairments [1], asphyxia, and problems related to the central nervous system [2]. With the introduction of the concept of Internet of Things (IoT), AI-based baby monitors have started to gain popularity with the inclusion of features like noise/cry detection, activity detection, and sleep detection, etc. Baby monitors have given parents greater flexibility since these solutions generally provide insights into the health and wellbeing of their babies [3].

Recently analyzing cry signals has received great importance in baby monitoring for non-invasive diagnosis of diseases. Most of hearing disorders can be cured if they were identified in first few months. Researches have shown that ratio between the dominant and fundamental frequencies of the cry differs with the hearing impairments [4]. Researches have also proven that frequency-domain features of the signal would be different to each other depending on the cause of the cry. To determine the cause, most of the researchers have successfully experimented with signal processing techniques like Fast Fourier Transform(FFT) [5], Wavelet Transform(WT), Mel Frequency Cepstral Coefficients analysis (MFCC)[6] and Empirical Mode Decomposition (EMD) [7]. Some researchers have applied deep learning techniques like convolution neural networks (CNN) trained with 2D short time spectrogram as the feature vector [8]. Although these researches involvecry audio recorded at 8kHz to 44.1kHz, it is experimentally shown that cry is distributed around 200-500Hz with an average at 320Hz for males and 400Hz for females [5]. Hence as per the Shanon-Nyquist theorem, audio recording at 8kHz sampling rate would be adequate for these applications [9].

Recently, intelligent hardware solutions for infant cry detection have been successfully experimented[3]. These solutions have complex hardware designs resulting in an expensive bill of materials. This paper discusses a hardware solution for automated infant cry detection in domestic environments in real-time at a low cost.

II THEORETICAL BACKGROUND

Success of audio processing mainly depends on the temporal and spectral characteristics of the signal. Temporal features of audio signals can be used to determine whether the signal contains a human voice. Further spectral characteristics are analyzed to confirm cry words and identify their causes. The audio signal was assumed to be a short-term stationary signal for both Temporal and Spectral feature extractions.

Short time energy

Short-time energy (STE) is defined as the average sample energy of the signal in a short time frame. Mathematically it can be defined as:

$$E(n) = \frac{1}{n}\sum_{i=0}^{N-1}[w(m)x(n-m)]^2$$

where w(m) represents coefficients of a windowing function (Hamming window) of length N to minimize the maximum side-lobs in the power spectral density (PSD) estimation. In general, Human voice contains higher STE

Revised Manuscript Received on 14 September, 2019.

PradeepPathirana, Department of Computational Mathematics, Faculty of Information TechnologyUniversity of Moratuwa, Sri Lanka.(E-mail: pdppathirana@gmail.com)

SagaraSumathipala,Department of Computational Mathematics, Faculty of Information Technology University of Moratuwa, Sri Lanka.(E-mail: sagaras@uom.lk)

compared to white noise and unvoiced signals [10]. Hence STE has been widely used in speech detectors as a parameter to identify the human voice.

Short time zero crossing

Short-time zero-crossing (STZC) is defined as the rate of change of sign over a short period of time. Mathematically STZC is defined as:

$$Z(n) = \frac{1}{2N}\sum_{m=0}^{N-1}\left|sign\big(x(n-m)\big) - sign(x(n-m-1))\right|$$

Where

$$sign\big(x(m)\big) = \begin{cases} 1 & x(m) \geq 0 \\ -1 & x(m) < 0 \end{cases}$$

In general, the human voice has a significantly lower STZC rate compared to noise since the human voice is scattered around the lower end of the spectrum [9]. Hence STZC plays a critical role given the lesser complexity of the computation.

To be consistent with the assumption of a stationary signal, typically all the parameters are estimated over a period of approximately 16ms (128 samples for 8kHz sampling rate). In contrast, research by Kevin et all(2010) [9] suggests estimating STZC over a large window to suppress the effect of impulsive voice artifacts such as coughs over a shorter period.

Fundamental frequency and dominant frequency

Fundamental frequency is defined as the lowest significant frequency in the PSD. Similarly, the frequency with the highest amplitude in PSD is defined as the dominant frequency. Harmonics of the fundamental frequency appear in the PSD as the critical frequencies out of which the most significant frequency becomes the dominant frequency. Since dominant frequency is a harmonic of fundamental frequency, ratio between dominant frequency and fundamental frequency is an integer. It has been shown that ratio between these two has a relation with the health of the baby [4].

Mel frequency cepstral coefficients

Mel Frequency Cepstral Coefficients (MFCC) analysis is considered as the most popular technique to analyze the audio signals since it is one of the best representations of the given signal with respect to the human auditory system [2]. Response of the human ear to different audible frequency bands vary across the human audible range significantly[2]. Human ear is more sensitive to low frequencies than high frequencies. Mel frequency scale was developed to mathematically model the above behavior of thehuman auditory system. MFCC is calculated as:

$$m(f) = DCT(\log(|fft(x(n)|))$$

Where m(f) denotes the MFCC vector while x(n) denotes the input signal after noise filters and windowing functions. Fast Fourier Transform (FFT) of the input signal is applied to determine the constituent frequencies. As the next step, Critical frequency bands are filtered by applying Mel scale-filter banks to the frequency response. This step discriminates different frequency bands to which human auditory system is sensitive differently. Logarithmic operation is performed to simulate the sensitivity scale of the human auditory system to different frequencies. As the final step, discrete cosine transformation (DCT) is applied to represent the result as a summation of discrete cosines and reduce the length of the vector.

Cry detection

Recently, few researches have been conducted towards automated noise/cry detection targeting baby monitors[3], [11]. These solutions include complex hardware designs to support the requirements. Figure 1 shows the basic steps involved in cry detection.

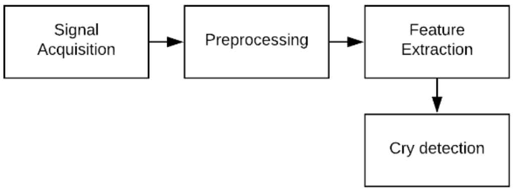

Figure 1 flow diagram of cry detection

Signal acquisition refers to the configuration of the hardware and continuous audio recording in the vicinity of the infant. Signal preprocessing involves the steps of noise filtering to remove sensor inherent noise and ambient noise. Feature extraction refers to the calculation of the features in time and frequency domains. These features will be used to train classifiers and then predict. Several researches have shown varying level of success in using different classifiers like k-nearest neighbors (KNN) [10], Artificial neural networks(ANN), support vector machine (SVM)[2] and neuro-fuzzy classifiers[12], etc.. Several researchers have experimented with probabilistic approaches like Hidden Markov Model (HHM), Parallel Hidden Markov Model (PHMM) and Probabilistic Neural Networks (PNN). Deep Learning with Convolution Neural Networks (CNN) and audio spectrogram as the 2D feature vector have shown promising results [1]. In addition, Radial basis neural networks(RBNN), probabilistic neural networks(PNN), etc. have been attempted in few occasions[13].

III METHODOLOGY

Hardware design

The scope of the hardware design of the proposed solution covers the steps from signal acquisition to cry classification and notifying the caregivers. Since audio processing requires a significant amount of processing power and memory, selecting a microprocessor is subjected to constraints of both performance and price. OSD3358 microprocessor (Beagle-Bone Board) was chosen to satisfy the above two requirements. Compared to the other microprocessors within the same range of performance and price, this has two additional microcontrollers(200MHz) known as Programmable Real-Time Units (PRUs) in same system on chip (SoC). Microprocessor can delegate selected tasks to PRUs to reduce the workload. MP34DT01 microphone is integrated to one of the PRUs to record the audio signal transmitted via Pulse Density Modulation (PDM) protocol. WL18MODGB Wi-Fi module is used tocommunicate with the server through Ethernet.

Agent-based firmwaredesign

Firmware which controls the external sensors and executes the algorithms was designed based on the Multi-Agent System concepts. Due to the limitations of resources, managing resources properly plays a key role in achieving the target successfully. Such a design provides performance by avoiding over-utilization/blocking of resources.

- Sensor agent resides in PRU and is responsible for audio recording via PDM.
- Preprocessor agent is responsible for noise filtering in audio signal.
- Classifier agent is responsible for voice activity detection, feature extraction, and cry classification.
- Auditor agent is responsible for auditing different processes and notifying.
- Network agent is responsible for sending messages and notifications to caregivers through server.
- Manager agent is responsible for monitoring life cycle of above agents

Signal acquisition and noise filtering

Audio interface was configured to sample the audio signal at 64kHz with sample size of 16bits. Due to the frequency selective attenuation in the sensor, audio signal was passed through a compensation filter to reverse the effect of frequency responseof the sensor. Output of the filter was down-sampled to 8kHz since according to the Shannon-Nyquist theorem, above sampling rate is adequate for the application. Further, this reduces the algorithm complexity in later stages. Signal is processed in real-time as a sequence of blocks of 4096 samples (512ms).

Voice activity detector (VAD)

For each block, VAD calculates STE and STZC to predict the presence of the human voice. The threshold for STE was experimentally determined to filter signals with the minimum required energy level. Similarly, the threshold for STZC was experimentally determined such that audio signals below the threshold would contain human voice. Blocks disqualified under the above conditions will not proceed further to feature extraction and classification. Such blocks will be regarded as noise blocks and will only be considered for classifier output.

Feature extraction

For further analysis of the blocks filtered by VAD, features are extracted as the next step by dividing blocks into non-overlapping frames of 512 samples (64ms). Ideally, the frame size should be in the range of 15-25ms to be consistent with the assumption of a stationary signal. But signal processing in real-time should maintain the equilibrium by processing the signal at a higher or equal rate compared to its production. Hence frame size was chosen as 64ms. MFCC feature vector was extracted for each frame. Further MFCC delta features were also evaluated against MFCC and did not show significant accuracy improvements.

Cry detection

To train the classifiers, features were extracted from the signals recorded from the same hardware to take the effect of the hardware into consideration. These features were used to train KNN and ANN classifiers independently. Both ANN and KNN classifiers were comparable in accuracy figures while the former was far better in terms of the processing time especially when the training data set is reasonably large. Hence ANN was chosen as the classifier to be included in firmware considering the resource limitations.

To classify an audio event, given event should be subdivided into frames of 64ms and processed by VAD and ANN classifier. Considering both the accuracy and performance, real-time audio stream is treated as a stream of non-overlapping windows of 5.12seconds (10 blocks of 4096 samples at 8kHz). This results in total of 80 frames of 64ms and based on majority votes cast by VAD and the classifier, an audio event of 5.12s will be labeled as quiet (discarded by VAD), noisy (passed through VAD but rejected by classifier) and cry (accepted by both).

Cry classification

Due to lower accuracy of ANN classifiers in cry classification and higher time complexity of KNN despite the higher accuracy, system breaks down cry detection and classification into two cascaded events. In this phase, feature vectors which were identified to be cry signals, are fed into KNN classifier to identify the cause of the cry if and only if the audio event of 5.12s was identified to be a cry signal. Hence compared to direct use of KNN for both cry detection and classification, time complexity and resource utilization have been reduced significantly due to the effective reduction in average number of feature vectors reaching the KNN classifier. KNN classifier was trained with a data set of infant cry signals annotated for "hungry", "belly pain" and "tiredness". To construct the required data sets, system recorded audio signals under each category allowing the system to filter features through VAD and cry detector before being included to data sets.

IV EVALUATION& RESULTS

Cry audio files used in this research were acquired from an online database along with the annotation provided by a group of experienced caregivers [14]. Since the feature extraction and classification are performed in the hardware design, above audio files were recorded again at 8kHz sampling rate using the hardware to capture both cry signal and sensor noise. Further, to increase the robustness of the system in noisy environment, recordings of a quiet room, white noise (lullaby), infant laughs, voices of adults, and multimedia files were also included to the training data set with the label of noise.

Frame wise cry detection accuracy was defined as

$$accuracy = \frac{m}{n}\ x\ 100\%$$

Where m and n denote the number of correct classifications and the number of total classifications respectively. Accuracy of classifying frames of cry and noise is 91.49% and 90.3% for ANN and KNN classifiers respectively.

Further, the proposed solution was evaluated against different scenarios that may frequently occur in practical use-case. To measure the accuracy under diverse conditions, recordings of white noise, songs/movies, audio clips of cries, laughs, adults' speeches were played allowing the system to analyze in real-time as a sequence of audio events. Table I summarizes the probabilities of the cry detector output under each scenario. With the limitations of hardware resources, processing time of each audio event is a critical measure of performance. To detect both cry and noise signals of 5.12seconds, ANN classifier took 100ms and "Quiet" audio events of the same length are detected under 1ms.

Table I Cry detection accuracy

Audio event	Quiet (%)	Noise (%)	Cry (%)
Quiet	99.54	0.46	0.00
Noise	3.62	93.21	3.17
Music	0.00	98.54	1.56
Laugh (baby)	0.00	99.38	0.62
Adult(male)	0.00	99.11	0.89
Adult(female)	0.00	99.01	0.99
Adult(stammer)	0.38	85.66	13.96
Cry	0.94	2.34	96.72

Similarly, the performance of KNN classifier was evaluated for the above-detected cry events. To analyze the performance, confusion matrix was formulated and further, precision, specificity(recall) and F1 measure was calculated. Table II and III summarize the frame-wise performance of the KNN classifier while Table IV and V summarize the event-wise performance under each category. Frame-wise performance of the classifier was measured from the audio frames produced by cry detector with the labels of "Cry" against the cry reason annotated in the data base. Similar to the evaluation of cry detection, performance of the classifier against cry events were estimated based on the results produced when respective cry signals were played in the vicinity of the device.

Table II Cry frame wise confusion matrix

Frame	Hungry	Belly Pain	Tiredness
Hunger	1887	315	220
Belly pain	528	2680	248
Tiredness	480	529	1551

Table III Cry frame wise performance evaluation

Frame	Hungry (%)	Belly Pain (%)	Tiredness (%)
Precision	65.18	76.05	76.20
Specificity	83.25	83.06	92.04
Recall	77.91	77.55	60.59
F1 Measure	70.98	76.79	67.66

Table IV Cry event wise confusion matrix

Event	Hungry	Belly Pain	Tiredness
Hunger	244	7	0
Belly pain	8	114	1
Tiredness	89	28	99

Table V Cry event wise performance evaluation

Frame	Hungry (%)	Belly Pain (%)	Tiredness (%)
Precision	71.55	76.51	99.0
Specificity	71.39	92.51	99.73
Recall	97.21	92.68	45.83
F1 Measure	82.43	83.82	62.69

Overall accuracy of the KNN classifier was estimated to be 72.51% and 77.46% for frame-wise and event-wise classifications respectively.

V CONCLUSION

This paper briefly describes the implementation of infant cry detection and classification in real-time using a simple low-cost hardware setup. Further, this presents a firmware design to optimize the resource utilization while maintaining the performance of the system to detect cry events and analyze the reason using voice activity detectors (VAD), cry detectors (ANN) and classifiers (KNN). ANN classifiers showed better performance for cry detection while KNN is better in terms of cry classifications/pattern recognition. Given nature of the algorithms, ANN is far better in terms of time complexity in comparison with KNN algorithm especially when the training data set is large. As the limitations, this shows that the adult voices with stammering and some white noise recordings with a significant energy in frequency ranges associated with the infant cries show the tendency of being identified as cries.

VI ACKNOWLEDGMENT

This research work was supported by Synergen Technology Labs LLC by extending its resources to continue the research. The authors thank its entire team for their contribution.

REFERENCES

1. M. J. Kim, Younggwan Kim, Seungki Hong, and H. 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), San Jose, CA, USA, 2013, pp. 1–4.
   R. Sahak, W. Mansor, L. Y. Khuan, A. Zabidi, and A. I. M. Yassin, "Detection of asphyxia from infant cry using support vector machine and multilayer perceptron integrated with Orthogonal Least Square," in Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, Hong Kong, 2012, pp. 906–909.

2. P. R. Myakala, R. Nalumachu, S. Sharma, and V. K. Mittal, “A low cost intelligent smart system for real time infant monitoring and cry detection,” in TENCON 2017 - 2017 IEEE Region 10 Conference, Penang, 2017, pp. 2795–2800.
3. G. Jr. Varallyay, Z. Benyo, A. Illenyi, Z. Farkas, and L. Kovacs, “Acoustic analysis of the infant cry: classical and new methods,” in The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Francisco, CA, USA, 2004, vol. 3, pp. 313–316.
4. S. Sharma and V. K. Mittal, “A qualitative assessment of different sound types of an infant cry,” in 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 2017, pp. 532–537.
5. A. A. Dixit and N. V. Dharwadkar, “A Survey on Detection of Reasons Behind Infant Cry Using Speech Processing,” in 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 190–194.
6. L. Abou-Abbas, L. Montazeri, C. Gargour, and C. Tadj, “On the use of EMD for automatic newborn cry segmentation,” in 2015 International Conference on Advances in Biomedical Engineering (ICABME), Beirut, Lebanon, 2015, pp. 262–265.
7. Y. Lavner, R. Cohen, D. Ruinskiy, and H. IJzerman, “Baby Cry Detection in Domestic Environment using Deep Learning,” p. 5, 2016.
8. K. Kuo, “Feature Extraction and Recognition of Infant Cries...,” p. 5.
9. L. Liu, Y. Li, and K. Kuo, “Infant cry signal detection, pattern extraction and recognition,” in 2018 International Conference on Information and Computer Technologies (ICICT), DeKalb, IL, 2018, pp. 159–163.
10. M. P. Joshi and D. C. Mehetre, “IoT Based Smart Cradle System with an Android App for Baby Monitoring,” in 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1–4.
11. J. Saraswathy, M. Hariharan, S. Yaacob, and W. Khairunizam, “Automatic classification of infant cry: A review,” in 2012 International Conference on Biomedical Engineering (ICoBE), Penang, Malaysia, 2012, pp. 543–548.
12. K. Srijiranon and N. Eiamkanitchat, “Application of neuro-fuzzy approaches to recognition and classification of infant cry,” in TENCON 2014 - 2014 IEEE Region 10 Conference, Bangkok, Thailand, 2014, pp. 1–6.
13. G. Veres, “donateacry-corpus”, [onlime] 2014, https://github.com/gveres/donateacry-corpus, (Accessed October 7,2019)

AUTHOR PROFILE

PradeepPathiranais following Master of Science in Artificial Intelligence at University of Moratuwa, Sri Lanka. He received Bachelor of Science in Engineering specialized in Electronics and Telecommunication Engineering from the same university. His research interest include wearable electronics and bio medical signal processing.

SagaraSumathipalareceived his Doctor of Engineering and Master of Engineering degrees from Nagaoka University of Technology, Japan. He is currently a Senior Lecturer in Department of Computational Mathematics, Faculty of Information Technology, University of Moratuwa, Sri Lanka. He is an active member of the Sri Lanka Association of Artificial Intelligence. His research interests include Natural Language Processing, Text Mining, and Intelligent Systems