Deep Learning Based Effective Baby Crying Recognition Method under Indoor Background Sound Environments

Karinki Manikanta
Computational Engineering and Networking
Amrita Vishwa Vidhyapeetham
Coimbatore,India
Karinkimanikanta@gmail.com

K.P. Soman
Computational *Engineering and* Networking
Amrita Vishwa Vidyapeetham
Coimbatore, India
kp_soman@amrita.edu

M. Sabarimalai Manikandan
School Of Electrical Sciences
Indian Institute Of Technology
Bhubaneswar, India
msm@iitbbs.ac.in

Abstract— **Effective automatic baby cry sound detection plays a significant role in many applications of smart baby condition monitoring. The paper presents deep learning based effective baby cry sound detection (BCSD) method under different kinds of background sounds in indoor environments. We investigated the functioning of the three BCSD approaches by making use of mel-frequency cepstral coefficients (MFCC) and machine learning classifiers like one-dimensional convolutional neural networks (1D-CNN), feed-forward neural networks (FFNNs), and multi-class support vectors machines (MC-SVM). We created the baby crying sounds for both training and testing of three models. In this study, we looked i n t o the results of three methods under different frame lengths including 100 milliseconds, 250 milliseconds and 500 milliseconds. Evaluation results showed that the 1D-CNN with frame length of 500 MS provides promising results as compared to that of the frame lengths 100 milliseconds and 250 milliseconds. For frame length of 500 milliseconds, the 1D-CNN-based, FFNN-based and SVM-based BCSD method had F1-score of 98.86%, 98.46% and 97.97%, respectively for detecting cry sounds. The 1D-CNN based BCSD method had class-wise F1-score of above 98%. Results showed that three BCSD methods have promising results for the same test sound database including air-conditioner, fan, speech and music sounds.**

Keywords—*Cry sound recognition, audio classification, feed forward neural networks, 1D Convolutional neural networks and support vector machines.*

I. INTRODUCTION

Cry is the essential sign of life in everybody's life which is seen shortly after a baby's live birth.[1]- [5]. The cry is a multimodal and complex action with a lot of in formation. [1]. A detailed acoustic analysis was also carried out to measure and compare signals of the acoustic characteristics of child cry and have demonstrated the diagnostic possible of cry signals for pathological conditions [1]. Furthermore, some of the researchers have studied on the use of cry signals for identifying the systemic problems such as cleft lip and invisible defects. In addition with aforementioned cry pathological analysis, baby cry sound recognition enables timely notification and allows parents to remotely monitor when their baby is crying. Thus, an automated recognition of crying sound patterns has become a most popular research topic in developing smart baby monitoring systems and analysing various patterns of crying sounds in different contexts such as hunger, sleepiness, pain and so on. Nowadays, increasing penetration of Smart phones and their sensing, computing and communication facilities have led to their use for monitoring different kinds of sound patterns daily routine and remote monitoring assistance [3] [14]. Prior works are often restricted to recognition of patterns of crying sounds for early diagnosis and treatment of new-borns. Some of the cry sound detection methods were developed for cloud computing platform where an user's device sends the audio to a centralized cloud for per- forming recognition task. The user-cloud computing demands higher bandwidth utilization costs and power consumption due to the continuous transmission of the recorded audio from a place of monitoring. Few of the past projects focused on automatic detection of baby cries under different background sound environments or controlled settings.

In this paper, we present deep learning based effective baby cry sound detection (BCSD) method under different kinds of background sounds in indoor home environments. The BCSD methods are based on the mel-frequency cepstral coefficients and machine learning classifiers such as one-dimensional convolutional neural networks, feed-forward neural networks, and multi-class support vectors machines. The rest of the paper is structured as follows. Section II summarizes the cry sound detection and analysis methods. Section III presents three baby cry sound recognition methods. Section IV presents the evaluation results of three methods. Eventually, conclusions are drawn in Section IV.

A. *Related Works*

In this section, we briefly summarize the cry sound recognition and analysis methods. In [1], Y. Kheddache and C. Tadj (2019) proposed a method to identify the diseases in new-borns using advanced acoustic features of cry signals. The method classifies the signals into the healthy and pathological cry classes using the MFCC and probabilistic neural network (PPN) classifier. The method had preteen cry detection rate of 88.17% and full-term cry detection rate of 82%. In [2], C. Y. Chang, et al. (2017) presented infant cry classification method using the MFCC, as pitch, harmonic-to-noise ratio, spectral centroid, flux, roll off, and flatness and histogram of oriented gradients(HOG) was investigated for selection of best feature vector. The audio

978-1-7281-2619-7/19/$31.00 ©2019 IEEE

frame of 50 milliseconds and 400 milliseconds for detecting silence and extracting features from non-silence frame, respectively. The audio frame of 25 milliseconds was chosen for MFCC feature extraction. The system detects streams and cries from digital devices su ch as television and music players in urban environments th at categorize indoor (office and home), outdoor, people inter action, huge human meetings, machinery and audio frame and tested for signal-to-noise ratio(SNRs) ranging from -15 dB to 40 dB. The method had a DR of 93.16% and FAR of 4.76% for a SNR values of 20 dB and a SNR of 10 dB. In [4], H. F. Alaie et al. (2015) presented a non-invasive health care system by performing the acoustic analysis of cry signals to quantitatively isolate and measure those cries and distinguish between healthy and ill newborn babies. The system consists of mono-channel, pre-emphasizes, segmentation of recorded cry signals, static and dynamic MFCC feature extraction, Gaussian mixture model-universal background model (GMM-UBM) and a score-level fusion. Results showed that the boosting mixture learning (BML) based procedure has lower error rates than the Bayesian method or the maximum a posteriori probability (MAP) adaptation method. In [5], M. M. Jam, et al., (2013) presented automated wavelet-based cry recognition system for noticed children with hearing loss from normal infants based on Mel filter-bank discrete wavelet coefficients (MFDWCs), principle components analysis (PCA), neural network (NN) classifier. Results showed that MFCCs and LPCs based methods had correction rates of 91.5 % and 86.5% of respectively. In [6], R. Torres, et al. (2017) investigate hand crafted baby cry (HCBC) features for baby cry sound recognition and compared with the MFCCs and a state-of-the-art CNN classifier. Results showed that HCBC features is better than the CNN in terms of less computation and memory space. In [7], S. Tejaswini, et al. (2016) presented a method to classify healthy, hunger, pain, and discomfort cries. The method consists of Preprocessing, silence removal, wavelet transform, MFCC and SVM linear kernel classifier. Results showed that the identification accuracy was 93.09%, 90.27% and 71.29% respectively for food-discomfort, pain-hunger and discomfort-pain. In [8], Y. Lavner, et al. (2016) two machine-learning algorithms for automatic baby cry recognition in audio recordings are investigated: the logistic regression (LR) classifier with MFCC, pitch and formants; and the CNN with log Mel-filter banks. The CNN classifier shows preferable results as differentiate to the LR classifier under various types of domestic sounds, such as people talking and door opening. In [9], R. Sahak, et al. (2012) investigated asphyxiated infant cry recognition based on the MFCC, orthogonal least square (OLS), and SVM classifier. Results showed that the OLS-SVM method yields comparable accuracy of 92.5% as compared to the DS-SVM method. In [11], J. O. Garcia, et al. (2003) presented an automatic screening system with the objective of classifying two types of cries: normal and pathological cries of deaf babies based on the feature MFCC and feed-forward neural network (FNN) classifier. The method had the accuracy of 97.43%. In [12], M. Petroni, et al. (1995) presented infant cry vocalization category (angry, pain, fear) with the help of artificial neural networks such as feed-forward neural networks (FNN), recurrent neural networks (RNN), cascade correlation neural network and time-delay neural networks (TDNNs).

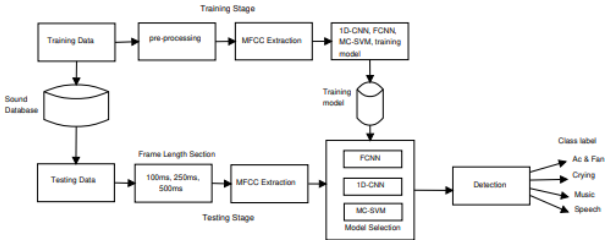

Fig.1. Cry sound recognition methods block diagram with the help of machine learning classifiers, MFCC features

II. CRY SOUND RECOGNITION METHOD

In this study, we present three cry sound recognition methods based on the MFCC features and 3 machine classifiers such as, FFNN, 1D-CNN and multi-class SVM. The performance of the three cry sound detection methods are evaluated under different background sounds encountered in indoor home environments. The block diagram of ML based cry sound recognition method is shown in the above mentioned diagram, cconsisting of four steps: pre-processing, extraction of features, cry models, identification. Each cry sound detection steps is explained in the further subsection.

.

A. Preprocessing

Practically, cry sound signals which are recorded are commonly corrupted with low-frequency noise components such as micro- phone artifacts, recording instrument biasing and power-line interference which are generated by the sensors movements and electromagnetic (EM) since the audio sensor is exposed tothe environments. Therefore, the audio recorded signal is sent between a High Pass Filter (HPF) with 60 Hz threshold frequency. Next, the signal is split up into frames with three frame lengths such as 100 milliseconds, 250 milliseconds and 500 milliseconds which are considered for performance evaluation. The intensity of the cry sound varies due to the stochastic nature of cry sound production and the area of the sound observing zone which is time fluctuating from acoustic sensors the source of sound. In practice the sound origin location can be unknown. Although the function is not sensitive to the amplitude rate, the normalization of amplitude is performed on the zero-mean audio formed signal to limit changes in the sensitivity of microphone.

B. MFCC Feature Extraction

In this study, we use the MFCC features for *recognizing the cry sounds signals. Each of the frame,* we extract MFCC features for the cry, fan and air-conditioner, music, and speech signals. The feature origin is described below as described in [13]:

(i) Pre-emphasis filter It is utilized to intensify huge frequencies which equilibrium the spectrum, as in general high frequencies have

low amplitudes related to lower frequencies and betters the SNR ratio. Pre-emphasis filter is executed as

$$y[n]=x[n]-\alpha x[n-1] \quad (1)$$

Where α is fixed as 0.97

(ii) Aftermath the pre-emphasis, every audio frame is executed with Hamming window(window function), the Hamming window function is defined as

$$h(n) = 0.54 - 0.46\, Cos(\frac{2\pi n}{N-1}) \quad (2)$$

where N($0 \le n \le N-1$) is audio frame length. The window function $h[n]$ is multiplied with the filtered signal $y[n]$.

a. **Fourier Spectrum**: Window formatted audio frames z[n] are taken by using FFT, which gives energy distributions over frequencies thus giving the magnitude of the spectrum.

b. **Mel-Frequency Spectrum**: The magnitude of spectrum is procreated by using 26 band-pass filters which have places at regular intervals on the Mel-scale. This concerns to linear frequency f expressed as:

$$(f) = 1125\, In\ (1 + \frac{f}{700}) \quad (3)$$

Mel-frequency is the log of f (linear frequency) which manifests similar effects of audio in the people's view.

c. **MFCC**: Filter bank coefficients are hugely corresponding to DCT. DCT is applied for de-correlating the filter bank's coefficient and these coefficients are calculated using

$$C_n = \sum_{k=1}^{k} \log S_k\ [n\ (k - \frac{1}{2}\frac{\pi}{K}) \quad (4)$$

n = 1, 2..... , K=26 , no. of triangular band-pass filters, S_k is the Kth triangular band pass filter energy output. The lower 13 of 26 coefficients for each frame is used in this method. The parameters that remove the pre-emphasis of the MFCC as 0.97, filter number 26, bottom band edge as 0 Hz, upper band edge as 8000 Hz. For each audio frame, the cepstral coefficients of 13.

(iii) *Description of Machine learning Classifiers*

The investigation, we measure is the performance of 3 machine learning classifiers such as FFNN, 1D-CNN and MC- SVM under numerous audio frames of size 100 milliseconds, 250 milliseconds, 500 milliseconds for automatically recognizing four classes. In further subsections, we briefly describe the machine learning classifiers and specifications used for classification of four classes of sound.

1) *Multi-Class SVM:* For trained information (xi, yi), where i = 1, 2, ...N, the optimization of SVM is equated as follows:

$$\arg min\ \{\frac{1}{2}|W|^2 + C\ \sum_{i=1}^{N} \xi i\} \quad (5)$$

Subject to the constraints

$$y_i(w.\ x\ - b) \ge\ 1 - \xi_i, \qquad \xi_i\ \ge 0 \quad (6)$$

Where ξ_i is slack with non-negative variable which is evaluated using a lagrazian formula of the same, decision function and make the multipliers i:

$$f(x) = sgn\ \{\sum_{i=0}^{N} y_i \alpha_i x.\ x_i + b\} \quad (7)$$

N represents training specimens number and feature vector (x). K (Non-linear kernel function) (xi, xj) is utilized to substitute the dot products xi and x, influenced by showing the data into a linearly separable higher dimensional space. The decision function is defined as:

$$(x) = sgn\ \{\sum^{N}{}_{i=0} yi\ \alpha_i K(x, x_i) + b\} \quad (8)$$

In the study, the Gaussian radical kernel used and defined as $K_R = ex(-\gamma|x - x_i|^2)$. Table-I mentions the specification of the cry sound recognition scheme based on MC SVM.

2) FFNN Classifier: It contains three layers.

They are input, output and hidden layer. The input for the corresponding layer is obtained from its preceding layer. The obtained input is evaluated and translated which is given as input to the immediate layer. Each layer has weight of its own. Fig. 2 represents the block diagram of the feed forwarded neural network forward network. Non-linear transformation can be defined as

$$h\ = f(W^i h^{i-1} + b^i),\ \ for\ 0\ \le i\ \le L \quad (9)$$

h0 equates to the input x. Wi, bi are weight matrices and bias vectors for the ith layer, and the desired output of final layer is hL. F (non-linearity) is usually a tanh/sigmoid.

Nevertheless, the ReLU activation function can be utilised for quick intersection of models. The ReLU is defined as f (x) = max(0, x), which is simpler than the activation function tanh or sigmoid because it needs a linear operator on a piece-wise basis than sigmoid or tanh stable point. The FFNN based cry sound detection identifications are listed in Table II.

3) *1D-CNN Classifier for Cry Sound Recognition:* The 1D-CNN based cry sound recognition block diagram is illustrated in Fig. 3. Table III summarizes 1D-CNN architecture specifications. In this work, a study was conducted to choose optimal no. of layers and parameters were altered to obtain better performance. In this study, the proposed 1D-CNN architecture contains of four convolutions, one drop out, two maxpooling, two fully connected layers and one flatten. The filter shifts for convolution and maxpooling are defined as one and two respectively. The maxpooling process minimizes the feature maps size and holds crucial and substantial audio feature functions. Two-Dimensional output is converted by flatten layer into One-Dimensional output.

TABLE I
SPECIFICATION OF MULTI-CLASS SVM CLASSIFIER.

Parameters	Frame Length(FL)		
	100ms	250ms	500ms
Audio Format	.wav	.wav	.wav
Channel	mono	mono	mono
Bit depth	16-bit	16-bit	16-bit
Class	4	4	4
Training Duration	384min	384min	384min
Testing Duration	96min	96min	96min
sampling rate	16000 Hz	16000 Hz	16000 Hz
Feature	MFCC	MFCC	MFCC
N0.of samples	1600	4000	8000
NFFT	2048	4096	8192
kernel	RBF	RBF	RBF
penalty parameter C	1	1	1
Gamma	0.0769	0.0769	0.0769
Degree	3	3	3
cache size	200	200	200
Tolerance	0.001	0.001	0.001
No .of Neurons	1600	1600	1600
Batch Size	128	128	128
Learning Rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	Infinity	Infinity	Infinity
Training time	42 Hour	5 Hour	1 Hour
Trained Model Size	28.9 MB	10.8 MB	5.0 MB

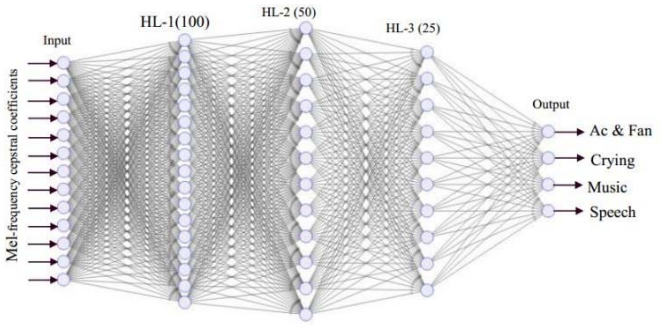

Fig 2. Block diagram of fully connected FFNN for cry sound recognition

To restrict the model overfitting, before each dense layer, a dropout layer is used. In the final stage of process, the neurons are interlinked using FC layer in the last layers. Stating the cry sound detection based on one-dimensional CNN is listed in Table III.

TABLE III
SPECIFICATION OF 1D-CNN BASED CRY SOUND RECOGNITION SCHEME FOR DIFFERENT FRAME LENGTH

layers(Type)	Frame Length					
	100ms		250ms		500ms	
	Shape	Parameters	Shape	Parameters	Shape	Parameters
1D-Conv	(12, 60)	180	(12, 60)	180	(12, 60)	180
1D-Conv	(11, 50)	5050	(11, 50)	5050	(11, 50)	5050
Max-pooling	(5, 500)	0	(5, 500)	0	(5, 500)	0
1D-Conv	(4, 100)	10100	(4, 100)	10100	(4, 100)	10100
1D-Conv	(3, 100)	20100	(3, 100)	20100	(3, 100)	20100
Flatten	300	0	300	0	300	0
Dropout	300	0	300	0	300	0
Dense	100	30100	100	30100	100	30100
Dropout	100	0	100	0	100	0
Dense	4	404	4	404	4	404

TABLE II
SPECIFICATION OF FFNN FOR CRY SOUND RECOGNITION

Parameters	Frame Length(FL)		
	100ms	250ms	500ms
Audio Format	wav	wav	wav
Channel	mono	mono	mono
Bit depth	16-bit	16-bit	16-bit
Class	4	4	4
Training Duration	384min	384min	384min
Testing Duration	96min	96min	96min
Sampling Rate	16000 Hz	16000 Hz	16000 Hz
Feature	MFCC	MFCC	MFCC
Number of samples	1600	4000	8000
NFFT	2048	4096	8192
Hidden layers	3	3	3
Activation Function	Relu, softmax	Relu, softmax	Relu, softmax
Fold	5	5	5
Nnumber of Neurons	1600	1600	1600
Batch Size	128	128	128
Learning Rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	30	30	30
Number of Parameters	7829	7829	7829
Training time	11.61 Sec	3.79 Sec	1.93 Sec
Trained Model Size	127.1 kB	127.1 kB	127.1 kB

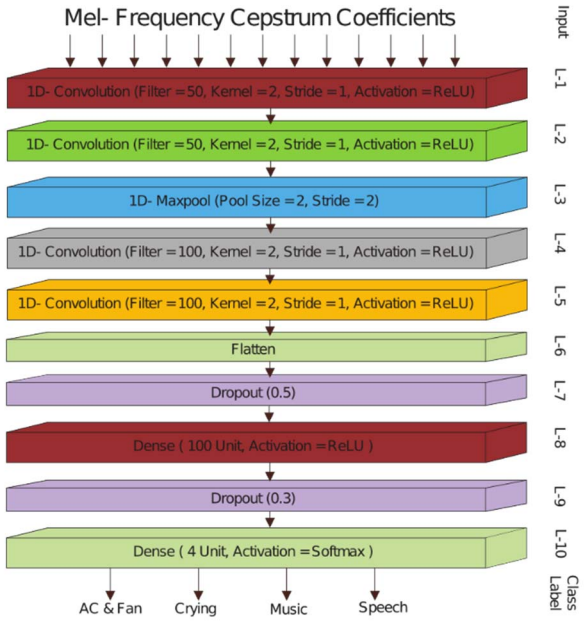

Fig 3. One-dimensional CNN architecture for cry sound recognition.

III. Results and Discussion

In this research, four sound classes such as cry, fan and air-conditioner, speech, and music are considered which can be encountered in indoor environments. We calculated the performance of three machine learning classifier based cry sound recognition methods using MFCC feature.

A. Validation Sound Database

There are no much freely-available cry sound databases in the literature. In the past published works, the cry sound databases were created for performance evaluation but not available for public access. Since the major objective of this paper was to evaluate the ML based baby crying sound detection methods performance under different indoor background sounds, we created the sound database including cry,

TABLE IV
SPECIFICATIONS OF 1D-CNN FOR CRY SOUND RECOGNITION

Parameters	Frame Length(FL)		
	100ms	250ms	500ms
Audio Format	.wav	.wav	.wav
Channel	mono	mono	mono
Bit depth	16 bit	16 bit	16 bit
Class	4	4	4
Training Duration	384min	384min	384min
Testing Duration	96 min	96 min	96 min
Sampling Rate	16000 Hz	16000 Hz	16000 Hz
Feature	MFCC	MFCC	MFCC
Number of samples	1600	4000	8000
NFFT	2048	4096	8192
Convolution layer	4	4	4
Dropout	2	2	2
Max-pooling layer	1	1	1
Activation Function	Relu, Softmax	Relu, Softmax,	Relu, Softmax
Fold	5	5	5
Number of Neurons	1600	1600	1600
Batch Size	128	128	128
Learning Rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	30	30	30
Number of Parameters	65904	65904	65904
Training time	52 min	23 min	10 min
Trained Model Size	843.0 kB	843.0 kB	843.0 kB

TABLE V
DESCRIPTION OF VALIDATION DATABASE

Title	Class	Contribution		
		Duration(hrs)	Self	Internet
Ac & Fan	Split ac, Central ac, Ceiling fan, Pedestal fan, Wall fan, Table fan.	2	100%	-
Crying	Crying	2	-	crying web sites
Music	Classical, country, Disco, Hiphop, Jazz, Metal, pop, Raggae, Rock, Tv news, tv programs, Movies.	2	90%	10%(GIZAN)
Speech	Male, Female, children.	2	100%	-

fan and air- conditioner, speech, and music that are described in Table V. We used EVISTR digital voice recorder and H1n Handy recorder for recording audio signals under various home con- dictions. The signals are converted to digital signals at sampling rate of 44.1 kHz and resolution of 16-bit. In addition to audio signals, we also collected audio from the public multimedia websites (freesound.org, soundsnap.com, pond5.com, GTZAN library, and YouTube). The audio total duration is around 8 hours. For training and test purposes, the audio recorded signal is resampled to 16 kHz.

B. Performance Metrics:

In this process of study, we assessed the methods performance using guiding principle metrics such as F1-score(F1), recall rate (RR) and precision rate (PR), that are defined as

$$\text{Precision Rate } (PR) = \frac{TP}{TP+FP} \times 100 \qquad (10)$$

$$\text{Recall Rate (RR)} = \frac{TP}{TP+FN} \times 10 \qquad (11)$$

$$\text{F1} - \text{Score} = 2 \times \frac{PR \times RR}{PR \pm RR} \qquad (12)$$

here TP denotes the true positive, FP denotes the false positives, and FN denotes the false negatives.

C. Performance Comparison

The comparison results of this study are summarized in Tables VI and VII for three methods with different lengths. We investigated the performance of three cry sound recognition methods under different frame lengths including 100 ms, 250 ms and 500 ms. Evaluation results showed that class- wise accuracy for the 1D-CNN is better than that of the other classifiers for all frame lengths. For frame length of 500 ms, the 1D-CNN-based, FFNN-based and SVM-based BCSD method had F1-score of 98.86%, 98.46% and 97.97%, respectively for detecting cry sounds.

IV. CONCLUSION

In our paper, we presented and discussed the results of three cry sound recognition methods using the MFCC and ML classifiers such as multi-class support vectors machines, fully connected feed-forward neural networks, and one-dimensional convolutional neural networks for indoor cry sound monitoring applications. Evaluation results showed that the 1D-CNN with frame length of 500 ms provides promising results as compared to that of the frame lengths 100 ms and 250 ms. The 1D-CNN based based method had class-wise F1-score of above 98%. Our preliminary results demonstrated that the CNN based cry sound recognition has a great potential applications in cry sound pattern analysis. In future directions, we attempt to implement our methods on real-time hardware for edge computing applications.

TABLE VII
COMPARISON OF PERFORMANCE OF DIFFERENT CLASSIFIERS WITH DIFFERENT FRAME LENGTHS.

		1D-CNN			FFNN			SVM		
FL	Class	PR	RR	F1	PR	RR	F1	PR	RR	F1
100ms	AC & FAN	99.96	99.97	99.96	99.95	99.84	99.89	99.81	99.97	99.88
	CRY	97.82	96.98	97.39	97.03	97.46	97.24	97	96.36	96.67
	MUSIC	96.38	95.87	96.37	96.19	95.97	96.07	96.23	95.74	95.98
	SPEECH	95.26	97.36	96.29	96.25	96.15	96.19	95.26	96.22	95.69
	Avg.	97.35	97.54	97.5	97.35	97.35	97.34	97.07	97.07	97.05
250ms	AC & FAN	99.95	99.99	99.96	99.96	99.98	99.96	99.85	99.99	99.91
	CRY	98.85	97.71	98.27	97.91	97.98	97.94	97.69	97.09	97.99
	MUSIC	97.35	97.71	97.52	97.08	96.75	96.91	96.83	96.11	96.46
	SPEECH	96.8	97.9	97.34	96.89	97.14	97.01	95.66	96.83	96.24
	Avg.	98.23	98.32	98.27	97.96	97.96	97.95	97.5	97.5	97.65
500ms	AC & FAN	99.97	99.99	99.97	99.97	99.98	99.97	99.85	99.96	99.9
	CRY	99.01	98.72	98.86	98.58	98.35	98.46	98.34	97.61	97.97
	MUSIC	98.18	98.19	98.18	98.21	97.02	97.61	97.33	96.57	96.94
	SPEECH	97.97	98.23	98.09	96.8	98.22	97.5	95.93	97.35	96.66
	Avg.	98.78	98.78	98.77	98.39	98.39	98.38	97.86	97.87	97.86

TABLE VI
CONFUSION MATRIX FOR THREE ML BASED BCSD METHODS FOR FRAME LENGTH(FL) OF 100 MS, 250 MS, AND 500 MS

	1D-CNN				FFNN				SVM			
Frame Length (FL) = 100ms												
	A&F	**C**	**M**	**S**	**A&F**	**C**	**M**	**S**	**A&F**	**C**	**M**	**S**
AC& FAN	**144098**	3	27	6	**143914**	10	172	38	**144092**	0	33	9
CRY (C)	1	**139244**	2114	2218	1	**139933**	2089	1554	16	**138353**	2480	2728
MUSIC (M)	44	1602	**137947**	4743	42	1976	**138531**	3797	184	1809	**138197**	4146
SPEECH (S)	12	1489	2298	**140125**	24	2297	3217	**138386**	74	2459	2897	**138494**
Frame Length (FL) = 250ms												
	A&F	**C**	**M**	**S**	**A&F**	**C**	**M**	**S**	**A&F**	**C**	**M**	**S**
AC&FAN	**57803**	0	4	0	**57800**	0	7	0	**57802**	0	5	0
CRY	5	**56117**	695	612	0	**56271**	731	427	8	**55760**	751	910
MUSIC	15	275	**56102**	1248	7	504	**55768**	1361	53	576	**55398**	1613
SPEECH	6	377	824	**56288**	11	697	934	**55853**	22	740	1057	**55676**
Frame Length (FL) = 500ms												
	A&F	**C**	**M**	**S**	**A&F**	**C**	**M**	**S**	**A&F**	**C**	**M**	**S**
AC&FAN	**28769**	0	1	1	**28766**	0	1	4	**28761**	0	6	4
CRY	1	**28564**	171	196	0	**28456**	198	278	8	**28243**	272	409
MUSIC	1	141	**28464**	380	3	220	**28123**	640	20	213	**27992**	761
SPEECH	6	142	354	**27980**	5	189	312	**27976**	13	254	487	**27728**

References

[1] Y. Kheddache, and C. Tadj, "Identification of diseases in newborns using advanced acoustic features of cry signals," *Biomedical Signal Processing and Control,* Vol. 50, pp. 35-44, Jan. 2019.

[2] C. Y. Chang et al., "DAG-SVM based infant cry classification system using sequential forward floating feature selection," *Multidim. Syst. Sign. Process.,* Vol. 28, No.3, pp. 961-976, 2017.

[3] A. Sharma, and S. Kaul, "Two-stage supervised learning-based method to detect screams and cries in urban environments," *IEEE/ACM Trans. on Audio, Speech, and Language Processing,* Vol. 24, pp.290-299, 2016.

[4] H. F. Alaie et al., "Cry-based infant pathology classification using GMMs," *Speech Comm.,* Vol. 77, pp. 28-52. Dec. 2015.

[5] M. M. Jam, and H. Sadjedi, "Wavelet-based automatic cry recognition system for detecting infants with hearing-loss from normal infants," *The J. Engineering,* Vol. 11, pp. 63-64, Sep. 2013

[6] R. Torres, D. Battaglino, and L. Lepauloux, "Baby cry sound detection: A comparison of hand crafted features and deep learning approach," in Proc. *Int. Conf. on Engineering Applications of Neural Networks,* pp. 168-179. Springer, Aug. 2017.

[7] S. Tejaswini, N.Sriraam, and G. C. M.Pradeep, "Recognition of infant cries using wavelet derived mel frequency feature with SVM classification," in Proc. *Int. Conf. IEEE Circuits, Controls, Comm. and Comput.*, pp. 1-4, Oct. 2016.

[8] Y.Lavner, R. Cohen, D. Ruinskiy and H. Ijzernab, "Baby cry detection in domestic environment using deep learning," in Proc. *Int. Conf. IEEE the Science of Electrical Engineering,* pp. 1-5, Nov. 2016.

[9] R. Sahak et al., "Detection of asphyxia from infant cry by linear kernel support vector machine enhanced with features from orthogonal least square," In Proc. *IEEE Int. Conf. on Computer Appl. and Industrial Electr.,* pp. 341-345, Dec. 2011.

[10] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, C. A. Reyes-Garca, "Evolutionary-neural system to classify infant cry units for pathology's identification in recently born babies," in Proc. *7th Int. Conf. IEEE Artificial Intelligence,* pp. 330-335, Oct. 2008.

[11] J. O. Garcia, and C. R. Garcia, "Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks," in Proc. *Int. Joint Conf. IEEE on Neural Networks,* Vol. 4, pp. 3140-3145, July .2003.

[12] M. Petroni et al., Stevens, "Classification of infant cry vocalizations using artificial neural networks," in Proc. Int. Conf. IEEE Acoustics, Speech, and Signal Processing, Vol. 5, pp. 3475-3478, May. 1995.

[13] S. Soni, S. Dey and M. S. Manikandan, "Automatic Audio Event Recognition Schemes for Context-Aware Audio Computing Devices," 2019 7th Int. Conf. Digital Inform. Process. and Comm. (ICDIPC), Turkey, 2019, pp. 23-28.

[14] Manikandan MS, Soman KP, " A novel method for detecting R-peaks in electrocardiogram (ECG) signal", Biomedical Signal Processing and Control. 2012 Mar 1;7(2):118-2