New-born’s Cry Analysis using Machine Learning Algorithm

Parvi Agrawal [a], Manish Kumar [b], Vaishnavi Sriramoju [c], Kiran Deshpande [d], Nahid Shaikh [e]

[a] Information Technology, A.P. Shah Institute of Technology, Thane, India, parviagrawal@apsit.edu.in

[b] Information Technology, A.P. Shah Institute of Technology, Thane, India, manishkumar@apsit.edu.in

[c] Information Technology, A.P. Shah Institute of Technology, Thane, India, vaishnavisriramoju@apsit.edu.in

[d] Head of Department, Information Technology, A.P. Shah Institute of Technology, Thane, India, kbdeshpande@apsit.edu.in

[e] Faculty, Information Technology, A.P. Shah Institute of Technology, Thane, India, nashaikh@apsit.edu.in

Abstract

Generally, a child, especially a newborn's, needs the utmost care and awareness of parents or elders. But at present, parents are very occupied with their work and hardly get time to look after their baby. This may lead to physical and mental stress for infants as well as their parents. In this paper, grouping of newborn’s cry signals is done into five categories: hunger, discomfort, angry, tiredness and belly pain. The proposed system involves preprocessing the cry signal and finding the MFCC coefficients followed by feeding the recovered data into CNN for training and classification. Exploratory results show that the proposed method attains 74% perfection.

Keywords

Newborn’s cry analysis; MFCC; CNN

1. Introduction

For a newborn, the first communication with the outside world is in the form of crying of the infant. Crying for baby is crucial in ensuring survival, health and growth of the infant. The most important aspect of taking up this project is to help unskilled parents/trainee pediatricians/babysitters who would be aware of their baby’s needs. It is hard to define baby cry sound. So, our focus is to generate an automated and non-invasive platform to monitor baby's status to diagnose their emotional status. Many researches have been done regarding feature analysis of the infant cry but the most accurate for feature extraction is MFCC and also for categorizing baby cry sound use of various kind of ANN have been used. Exploratory study shows that Deep learning method based on CNN attains accuracy of more than

80\% [1]. The system will comprise of preprocessing including frame blocking and windowing followed by MFCC (Mel-Frequency Cepstrum Coefficients) feature extraction whose coefficients will be normalized [4] and sent to CNN (Convolutional Neural Network) for classification based on voice type and will classify the infant crying sounds into five classes: hunger, discomfort, angry, tiredness and belly pain.

2. Methodology

From Figure 2, the proposed system consists of the Input audio files which are in .wav format which are the infant cry sounds. These cry sounds are then sent for preprocessing which involves the frame blocking and windowing. The frame blocking is the process of breaking down the signal into frames. Signal analysis is done on short span periods of time which are called frames to get the accurate values. Typical values for frame blocking are dividing the frames into 709 samples and taking the overlapping length as 1s. This signal is then passed onto for windowing. Windowing is basically done in order to reduce the distortion of the signal. Finally, the signal is then multiplied with the hamming windowing function and this windowing signal serves as an input to calculate the Mel Frequency Cepstrum coefficients. Coefficients of MFCC are obtained by performing FFT (Fast fourier transform) which converts time domain to the frequency domain. FFT is performed in order to obtain the magnitude frequency of each frame. The coefficients of MFCC are then normalized and sent to Convolutional Neural Network for further classification. The CNN takes the MFCC as the inputs and does the classification of the cries. More the number of coefficients, better the system will be able to classify the data. CNN is a supervised learning algorithm that typically has one input layer, one hidden layer, and one output layer. The number of iterations in the training process is also influenced by the use of neurons in the hidden layer. As a result of the output neurons displaying the classes, the infant cry noises may be divided into five categories, as indicated in figure 1, namely hunger, discomfort, anger, weariness, and belly pain.

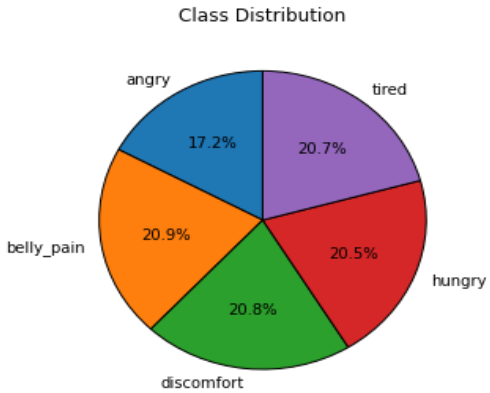

Figure 1: Class Distribution

Programming language used here is python and Google Collab as our Integrated Development Editor (IDE) are used for code and for matrix calculation and display, use key modules such as pandas, Numpy, Scipy, and Matplotlib. Librosa is a program that allows you to process audio and calculate features. Tensor Flow for classification model and Tqdm for calculating the features required for model analysis.

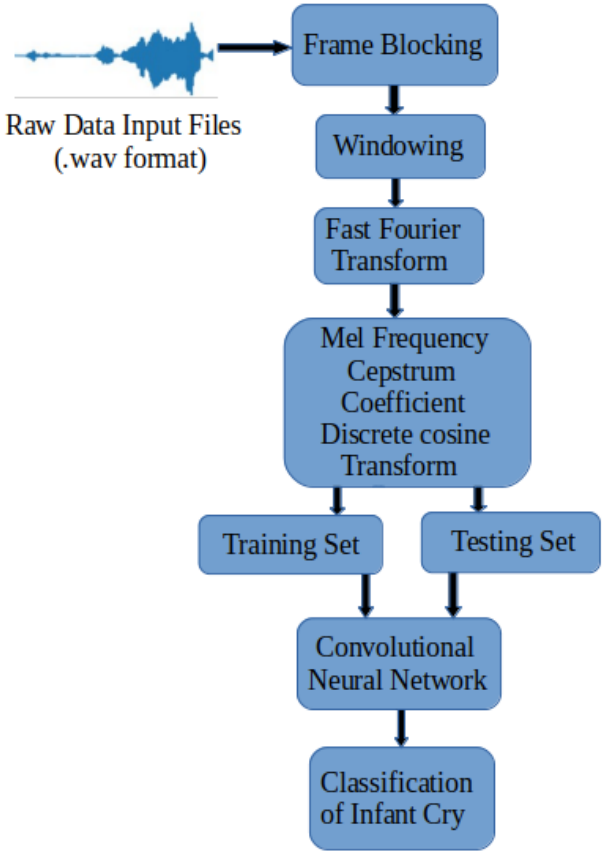

Figure 2: Proposed Method

3. Mathematical Formulas

3.1. Root-Mean-Square Energy

A weeping signal consists of weeping expression and terminated voice as well as non-voice, quietness or other antiquity. To determine the threshold, a root-mean-square-energy computation is utilised. The following is the RMSE equation:

$$E[p] = \sqrt{\frac{\sum_{s}^{Q} |z[p-s]|^2}{Q}}$$

where E[p] is the rmse function, z[p] is the signal equation, Q is the frame length, s means index of window, s starts with zero and p is the sample index.

3.2. Pre-emphasis

To amplify the high frequencies, the first step is to apply a pre-emphasis filter to the signal. A type of filter that is necessary to minimise noise and provide a spectrum that is not wrinkled. The input is a time domain signal that is expressed as follows:

$$T[p] = z[p] - cz[p-1]$$

Constant of emphasis filter is denoted by c, c should be greater than 0.9 and less than 1.

3.3. Frame blocking and Windowing

The total speech is fragmented into frames. Because speech is a time-varying signal, framing is essential because the values are not dynamic when viewed over a short period of time. To remove alias signal, this split frame will be multiplied using the FIR filter. Spectral leakage is reduced by the windowing technique. A proper Hamming window multiplies each frame.

$$T_1[p] = z_1[p]W[p]$$

Here 'p' should be greater than 0 and less than 1. The Hamming window is represented by W[p]:

$$W[p] = 0.54 - 0.46\cos\left[\frac{2\pi p}{Q-1}\right]$$

Here 'p' should be greater than zero and less than Q-1.

3.4. Fast Fourier Transform

Multiply the retaliation with collection of band pass filters which are triangular so that a smooth spectrum can be achieved. To convert time-series signal into frequency spectrum:

$$D_p = \sum_{l=0}^{Q-1} z_l \, e^{\frac{-2\pi elp}{Q}}$$

3.5. Mel-Frequency Wrapping

Using triangular overspreading windows and mel scale, map the spectrum's strengths onto the mel scale. Mel-scaled filter banks — triangle filters that imitate human perception:

$$F_{mel} = 2595 * log_{10}\left[1 + \frac{F_{HZ}}{700}\right]$$

3.6. Log spectrum computation

Log of the powers should be taken at each mel frequency. Filter banks are created by applying triangular filters to the power spectrum on a Mel-scale to obtain frequency bands.

$$D_v = log_{10}\left[\sum_{L=0}^{Q-1} |z[l]|U_v[l]\right]$$

Here 'v' can be from 1 to M [M refers triangle filters number] and Uv[l] shows the value of v-triangle filter which is having l as acoustic frequency value.

3.7. Cepstrum

To obtain the coefficients, which will then be normalised and given to CNN to be categorise. Compute the discrete cosine transform of the list of mel log powers as if it were a signal. Convert Mel-spectrum into the time domain using a Discrete Cosine Transform (DCT):

$$C_e = \sum_{e=0}^{L} D_e \cos\left[\frac{e(i-e)}{2\frac{\pi}{l}}\right]$$

4. Experiment and Results

4.1. Datasets

We have taken out infant data from Donate a cry corpus, Dunstan Baby Language and some are self-recorded from few sources. The proposed system consists of the Input audio files which are in .wav format which are the infant cry sounds. The following data is divided as 80% training set and 20% testing set. A total of five categories are being

considered: Hunger, Anger, Belly Pain, Discomfort and Tiredness. Total no. of datasets used are 709 and distribution is shown in Table 1.

	Angry	Hungry	Belly Pain	Discomfort	Tired
Train	64	396	42	27	37
Test	16	99	11	7	10
Total	**80**	**495**	**53**	**34**	**47**

Table 1: Dataset Distribution

4.2. Pre-processing

After accumulating newborn-crying signal, these cry sounds are then sent for pre-processing which consists of the frame blocking and windowing. This procedure employs a union of short-time methods including Energy Banks as well as Root Mean Square Energy (RMSE), with the backdrop of 10 s frame length and 1 s overlapping. Since the baby cry duration is of 5-10 seconds, perceptible threshold values equal to 0.005. Downsampling results are shown in figure 3.

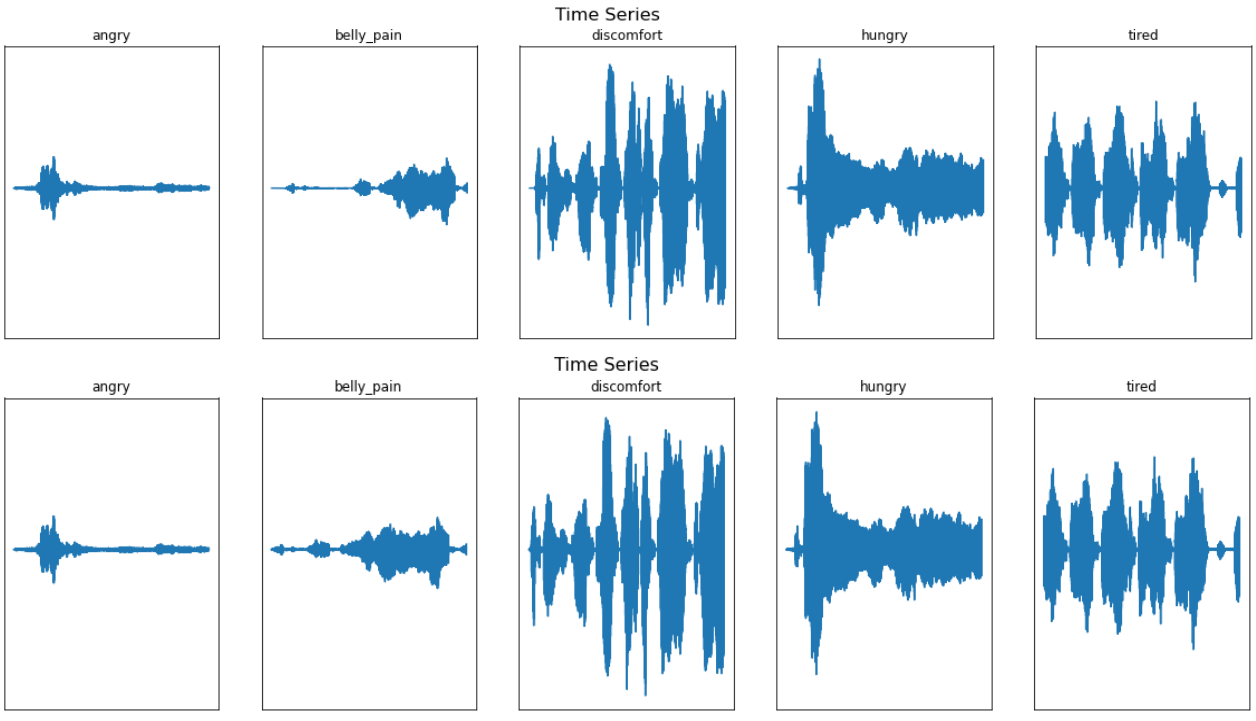

Figure 3: Data Before & After Downsampling

4.3. Mel – Frequency – Cepstrum Coefficient

Ultimately, the signal is then multiplied with the hamming windowing function and this windowing signal serves as an input to calculate the Mel Frequency Cepstrum coefficients. Remove the leftover signal time duration if it exceeds the starting point; otherwise, pad the signal with zero. The Librosa module's MFCC method is then used to compute variable by setting value of Mel Frequency Cepstrum Coefficient, value of Fast Fourier Transform, and no. of filters to 13, 1103, and 26 correspondingly. Aside from that, delta feature approaches are used to boost the amount of vital data in MFCC. Figure 4 depicts the MFCC results.

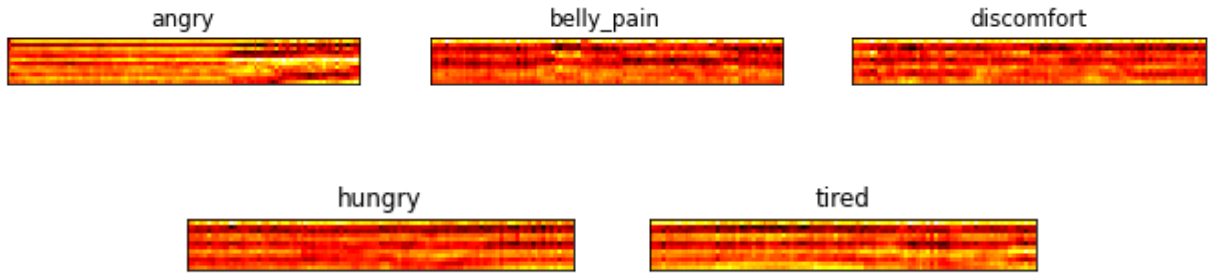

Figure 4: MFCC for Cry Categories

Obtained the coefficients of MFCC by executing FFT which is done in order to obtain the vast frequency of each frame and to convert time-series signal into frequency range is shown in figure 5.

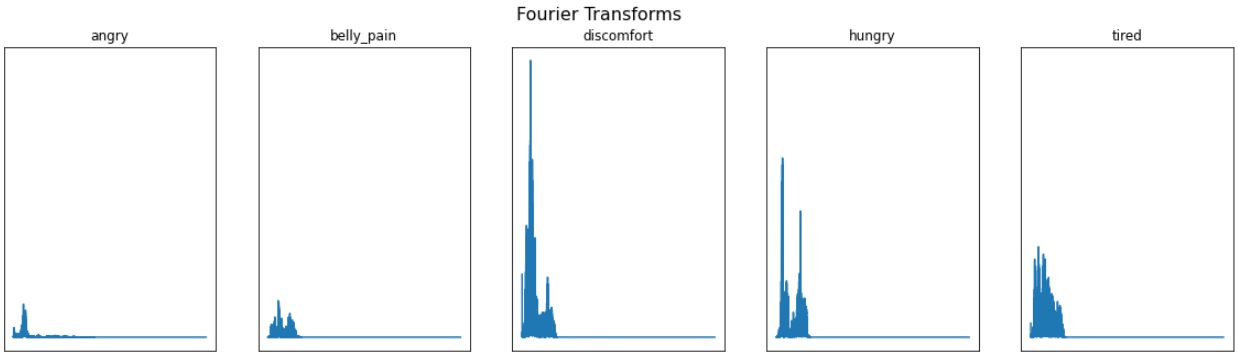

Figure 5: Fast Fourier Transform for Cry Categories

4.4. Convolutional Neural Network (CNN)

The coefficients of MFCC are then normalized and sent to Convolutional Neural Network (CNN) classifier for further categorization. From figure 6, The model consists of four deep two-dimensional convolutional layers (CONV2D).

16-128 filters showed boundary which is set in each layer and 3x3 size of kernel. The padding in the convolutional layer is unchanged. After the fourth convolutional layer, the max pooling procedure is used to help reduce computing demand by lowering dimensionality; however, now, keep up principal details. Dropout technique which helps in decreasing overfitting which is set to be 0.5, thus 50% of the nodes are dropped out randomly from the neural network. Other two variables for making system are reformer and loss function which are respectively Adam and categorical-cross entropy. After the model is constructed, training dataset of 13x9x1 input shape, 32 batch size and 10 epochs in figure. After ten rounds, the network obtains a test accuracy of 74%.

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 13, 9, 16)	160
conv2d_13 (Conv2D)	(None, 13, 9, 32)	4640
conv2d_14 (Conv2D)	(None, 13, 9, 64)	18496
conv2d_15 (Conv2D)	(None, 13, 9, 128)	73856
max_pooling2d_3 (MaxPooling2	(None, 6, 4, 128)	0
dropout_3 (Dropout)	(None, 6, 4, 128)	0
flatten_3 (Flatten)	(None, 3072)	0
dense_9 (Dense)	(None, 128)	393344
dense_10 (Dense)	(None, 64)	8256
dense_11 (Dense)	(None, 5)	325

Total params: 499,077
Trainable params: 499,077
Non-trainable params: 0

Figure 6: CNN Architecture

5. CONCLUSION

Exploratory outcome conveyed that the presented method contains high categorization correctness. According to the MFCC estimated result, the CNN with a frame duration of 10 ms, after having 10 iterations, the system gets 74% test correctness. Our primary outcome shown that cry sound based on CNN has higher possibility of getting required results. In the future scope, a stronger and a massive datatset of infant cries should be made available. The goal is for people to be able to grasp the meaning of infants' cries more easily and quickly by using the model. For future studies, a larger amount of audio data will be used for training and testing.

6. ACKNOWLEDGMENTS

The authors would like to extend gratitude towards respondents and experts for helping in process. Authors are extremely grateful to dunstanbaby website for support in making the model and in providing database and technical information concerning the “New-born’s Cry Analysis” research. Authors are truly grateful to efforts to improve technical writing skills. The blessings, assistance, and direction given from time to time will go a long way in the life path that the authors are about to embark on.

7. REFERENCES

[1] K. Teeravajanadet1, N. Siwilai2, K. Thanaselanggul3, N. Ponsiricharoenphan4, S. Tungjitkusolmun5, P. Phasukkit6,” An Infant Cry Recognition based on Convolutional Neural Network Method”, The 2019 Biomedical Engineering International Conference (BMEiCON-2019).

[2] Chunyan Ji, Thosini Bamunu Mudiyanselage, Yutong Gao & Yi Pan, “A review of infant cry analysis and classification”, EURASIP Journal on Audio, Speech, and Music Processing volume 2021, Article number: 8 (2021), Georgia State University, 25 Park Place, Atlanta, USA.

[3] Chuan-Yu Chang, Jia-Jing Li,” Application of Deep Learning for Recognizing Infant Cries”, 2016 International Conference on Consumer Electronics-Taiwan.

[4] Lichuan Liu, Senior Member, IEEE, Wei Li, Senior Member, IEEE, Xianwen Wu, Member, IEEE, and Benjamin X. Zhou,”Infant Cry Language Analysis and Recognition: An Experimental Approach “, IEEE 2019.

[5] Karinki Manikanta, K.P. Soman, M. Sabarimalai Manikandan,” Deep Learning Based Effective Baby Crying Recognition Method under Indoor Background Sound Environments”, IEEE 2019.