Identification and Classification of Baby Cry Sound Patterns for Infant Monitoring

[1]Amit Ambaji Sawant, [2]Veeresh Boragi, [3]Kalmesh Badiger, [4]Shrinivas Kadiwal
[5]M. C. Aralimarad

Basaveshwar Engineering College
(Autonomous) Bagalkot Karnataka.

***Abstract*- Baby cry detection involves tracking down and analysing a baby's cries to keep an eye on them and notify carers. Without the detection and classification of baby cry sound patterns for infant monitoring, carers wouldn't have a reliable technique to immediately assess and respond to the needs and well-being of infants, endangering their care and safety. This work targets the recognition and categorization of patterns in infant cry sounds for communication with the parents. This work targets the recognition and categorization of patterns in infant cry sounds for communication with the parents. It recently ranks as one of the most interesting medical study subjects. It makes the tasks of working parents or guardians easier and guarantees that the child receives the best care possible. It helps maintain track of a baby's activities, monitors the infant in the cradle, and minimizes the strain for nurses and doctors in the neonatal critical care unit. Using a cry detection technique that includes data collecting, pre-processing, Mel-Frequency Cepstral Coefficient (MFCC), feature selection, and classification, the main objective is to identify crying sounds based on crying patterns. It is challenging to categorize sounds and use Support Vector Machines (SVM) to produce accurate results. And tracking and observing children's activities is made easier by this material.**

***Keywords*: cry sound detection, cry sound classification, SVM, and MFCC.**

1. INTRODUCTION

Every year, 140 million babies are born in the world. It can be incredibly challenging to take care of new born infants, especially if you're a first-time parent. To handle the problems in real life, it is not enough to simply heed the counsel of other parents and that found in the literature. The main justification is that it could be difficult to understand a baby's cries. Infants utilize weeping to communicate with the outer world, unlike adults who express their feelings through voice, actions, etc. According to studies, infants scream in several ways to communicate their demands and emotions. Infants frequently scream when they are hungry, have less gas, sleeping, have burps, or are in pain. A language known as Dustan Baby Language (DBL), which contains expressions like "I am hungry," "I am sleepy," and others, is used by the baby to communicate. Babies are unable to communicate verbally or emotionally like adults. The infants' common language is separated into five groups called "baby language." [8] A baby's cry can disclose a lot about his or her identity, as well as their mental and physical state. These are the five words that babies use to express their wants, according to Priscilla Dustan. Those five words are "Neh" (hunger), "Eh" (need to go to the toilet), "Owh/Oah" (fatigue), "Eair/Eargghh" (cramps), and "Heh" (physical pain; feel hot or wet). A sobbing baby's fundamental frequency ranges from 250 Hz to 600 Hz. Dunstan Baby Language (DBL) classifies crying infants into three main phases. Pre-processing is the first stage, which normalizes all sound data; feature extraction is the second stage; and categorization is the third stage [8]. One of the frequently used feature extraction methods for audio processing is the Mel Frequency Cepstral Coefficient (MFCC), which is based on sound frequency. Many methods have been used for classification, but in this study, Support Vector Machine (SVM) classification is investigated to find the best situation for using this technique combination from feature extraction to classification. The paper is structured as follows: The literature study is covered in Section 2, baby cries are classified in Section 3, the suggested approach is implemented in Section 4, the results are shown in Section 5, and the conclusion and future scope are discussed in Section 6.

2. LITERATURE SURVEY

A review of current studies on hearing baby cries is given in this section.

Electrical Engineering Faculty at Telkom University in Bandung, Indonesia, Sita Purnama Dewi. [1] have shown that MFCC has some advantages in feature extraction that are used for the analysis of baby crying classification. These advantages include: It can identify the character of the sound to determine the pattern of sound; The output vector has a small data size but does not remove the noise characteristics in the extraction; and MFCC functions similarly to how a human listener functions in terms of providing their perceptions. Chunyan Ji, Yi Pan, Yutong Gao, and Thosini Bamunu Mudiyanselage. [2] have described the important scientific work in newborn cry analysis and classification and have offered information and tools that are useful to both researchers and medical professionals who work on this subject. It is shown that the lack of database resources hinders the expansion of the infant cry study. Large datasets with a variety of samples are necessary to enable deep neural networks.

The current trend in feature extraction is to create a mixed feature set that benefits from many domains to enhance discrimination. Relevant research results show that neural network-based architectures are gaining popularity. It performs more robustly and effectively than more traditional machine learning methods. Newborn screams are a window into an infant's emotions, according to studies by Ashwini K1, P. M. Durai Raj Vincent1, Kathiravan Srinivasan1, and Chuan-Yu Chang [3]. In this study, deep learning and machine learning are used to increase the effectiveness of the infant cry classification model even with small datasets.

Convolutional feature extraction-based machine learning classifiers perform well even with little datasets, although the SVM technique's hyperparameter change is computationally expensive.

3. BABY CRY CLASSIFICATION

i. Data Acquisition

Throughout the data collection stage, the baby's cries are recorded and classified. The bulk of databases are kept in homes or hospitals with labels provided by physicians, nurses, or parents. Digital recorders are placed close to infants and either instantly turned on to record each cry separately or left on for a significant amount of time to record the sounds around the infants. Infant sound is supposed to be a short-term stationary signal that is more immobile than other sounds because infants lack full control over their vocal tracts. Due to resource limitations and the delicate nature of the process employed to collect data on infant cries, the size of the infant cry database as a whole is severely confined. It contains signals for sleep, lower gas, hunger, discomfort, and pain, as well as five other signals. Each cry lasts exactly one second and a half. The Dunstan Baby Language database was made by Priscilla Dunstan, who also founded the Dunstan Baby Language theory. This database has also been used as a source of information in many literary works. There are numerous versions of the Dunstan Baby Language database because contributors extracted the audio recordings in diverse ways. The five "Dunstan words," each of which was taken to mean "Neh" for "hunger," "Eh" for "need to burp," "Oah" for "tired," "Eairh" for "low belly pain," and "Heh" for "physical discomfort," is an expression of newborn communication.[6]

ii. Pre-processing

Audio segmentation and denoising are the main responsibilities in the pre-processing stage. Infant scream signals get muddy due to the intricacy of the recording setting. Footsteps, adult voices, the sound of an air conditioner, an alarm, etc. may also be heard in a newborn care unit in addition to baby cry signals. Pre-processing the data is an essential step in correctly identifying or categorizing cry signals. Denoising, which removes background noise like speech, fans, footsteps, etc., is the initial stage of signal cleaning. We utilized the Audacity program to accomplish this.[6]

iii. Feature Extraction: Mel Frequency Cepstral Coefficient (MFCC)

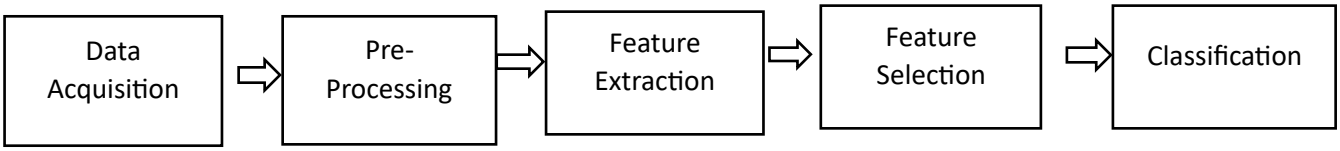

Fig: Baby Cry Detection

Feature extraction is the process of obtaining discriminative properties from audio signals for potential application in machine learning techniques. This is one of the most crucial elements in a machine-learning procedure. The fundamental work of baby cry analysis and processing is carried out by temporal or frequency domain feature extraction tasks. Calculating time domain features, such as zero crossing rate, amplitude, energy-based features, etc., is simple. Time domain features are not strong enough to capture the changes inside infant cry signals, and the features are susceptible to background disturbances. Frequency domain features, on the other hand, have a substantial ability to describe the qualities inside infant cry signals. The regularly used MFCCs, LPCCs, and LFCCs have proven to perform better than employing time domain features. On the other hand, it is shown that the baby scream signal is rhythmic and goes through cyclical changes as a result of breath and natural pauses.

MFCC

The specifics of feature extraction are covered in this section. A further Fourier transformation can be done to the almost stationary speech segments of a given voice sample with a duration of 20–40 msec. The speech sample phones that discriminate between two words can be identified using the spectral properties taken from the frames. MFCC is employed as a feature extractor. The Mel scale is a perceptual scale of frequencies which is dependent on the sensitivity of the human ear. The discrete audio signal is Si(n), where i is the frame number, and is the number of samples. The periodogram of audio signal [8] is computed by [1]

$$P_i(k) = \frac{1}{N}\left(\sum_{n=0}^{N-1} si\ (n)h(n)e^{-\frac{j2\pi kn}{N}}\right)^2 \qquad 0 \le k \le N\text{-}1 \qquad (1)$$

h(n) is a N sample long analysis window, N is the size of the DFT. An upper frequency f[m] and a lower frequency f[0] is chosen in Hz. Fig.1 shows Mel-Scale Filter Bank. [1]

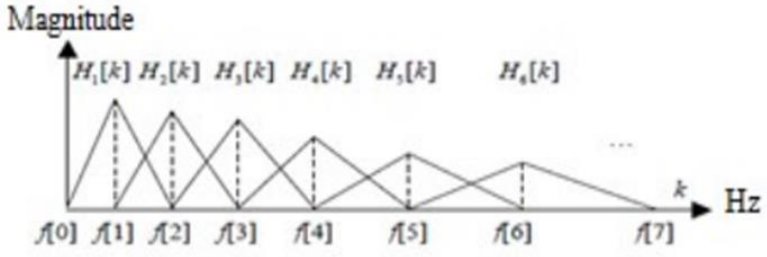

Fig: Mel-Scale Filter Bank

In Figure, X-axis represents frequency in kHz and Y-axis is the amplitude. The upper and lower frequency boundaries are M points uniformly placed in the Mel-scale and are given by

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_0) + m\frac{B(f_M)-B(f_0)}{M+1}\right) \quad (2)$$

Fs is the sampling frequency in Hz. B(f) is the transform from the Hertz scale to the Mel scale and B-1(b) is the inverse of the Mel Scale to the Hertz scale given by

$$B(f) = 1125\ln(1+f/700) \quad (3)$$

$$B^{-1}(b) = 700\left(e^{\left(\frac{b}{1125}\right)} - 1\right) \quad (4)$$

Differences and changes in lower frequencies are more easily perceived by the human ear than changes in higher frequencies. Therefore, equally spaced points in the Hertz scale will exhibit more density at lower frequencies than at higher frequencies in the Mel scale. A triangle filter Hm[k] is applied for each point M mel-bins to determine the amount of energy contained in each triangular in mel-bin. Finding the discrete cosine transform (DCT) for each of the M-filters in c[n] and computing the log energy for speech samples S[m] are the last two steps in determining the MFCC.

$$S[m] = \ln\left(\sum_{k=0}^{N-1} Np_i(k)H_m[k]\right) \quad 0 < m < M \quad (5)$$

$$C[n] = \sum_{m=0}^{M-1} S[m]\cos\left(\frac{\Pi n\left(m+\frac{1}{2}\right)}{M}\right) \quad 0 < n < M \quad (6)$$

Only the first 13 coefficients from each frame will aid in better decoding. The reason for this is that the human ear is more sensitive to lower frequencies than higher frequencies. The decoding result is never improved by the first 13 coefficients, which correspond to lower frequencies, and the remaining coefficients, which relate to higher frequencies. Therefore, for the recognition challenge, only the first 13 coefficients are taken into account. Cepstral Mean Variance Normalization (CMVN) is the feature transformer applied after the extraction of MFCCs. CMVN minimizes the effect of differences in variable environments like ambient noise, recording equipment, and transmission channels. CMVN (xt(i)) is calculated by [1]

$$(\hat{x_t}(i)) = \frac{x_t(i)-\mu_t(i)}{\sigma_t(i)} \quad (7)$$

Where xt (i) is the ith component of the original feature vector at time t and the mean μt(i) and standard deviation σt(i) are calculated over some sliding finite window of length N is given by

$$\mu_t(i) = \frac{1}{N}\sum_{n=t-N/2}^{t+\frac{N}{2}-1} x_n(i) \quad (8)$$

$$\sigma_t^2(i) = \frac{1}{N}\sum_{n=t-N/2}^{t+\frac{N}{2}-1}(x_n(i)-\mu_t(i))^2 \quad (9)$$

The first and second-order deltas Δ+ ΔΔ of the MFCCs can be calculated to add dynamic information to the MFCCs. For an acoustic feature vector x, the first-order deltas are defined as[1]

$$\Delta x_t = \frac{\sum_{i=1}^{n} w_i(X_{t+i}-X_{t-i})}{2\sum_{i=1}^{n_i} w_i^2} \quad (10)$$

where n is the window width and wi is the regression coefficients. The second-order delta parameters are derived in the same fashion as,

$$\Delta^2 x_t = \frac{\sum_{i=1}^{n} w_i(X_{t+i}-X_{t-i})}{2\sum_{i=1}^{n_i} w_i^2} \quad (11)$$

The combined feature vector becomes as,

$$X_t = [\,X_t \quad \Delta x_t \quad \Delta^2 x_t\,] \quad (12)$$

iv. Feature Selection

Feature selection is the process of selecting a subset of features from the initial features that were extracted from the audio signals using feature extraction techniques. The objective is to reduce the dimensionality of the characteristics while maintaining classification precision. Future smart systems for detecting and categorizing the cries of newborns will be viable and affordable to construct since they will require fewer features and less processing power. There may be some redundant data in the original characteristics as well, making it challenging to discriminate between the various cry signal types. By selecting the appropriate features for the task at hand, classification accuracy may be boosted. The many feature selection methods that have been used to analyse newborn screams are examined in this section. The F-ratio method was used to choose the top 20 MFCC features. Significant significance coefficients are related to higher F-ratio scores.[6]

v. Cry Classification

The most crucial step in the machine learning process is selecting the right classifier after the data has been cleaned, segmented, and features have been extracted, chosen, and normalized. We examine a Support Vector Machine (SVM) in this section.

- Support Vector Machine

The most popular probabilistic classifier for categorizing baby screams is the Support Vector Machine (SVM). The various SVM types include multiclass SVMs, linear SVMs, and binary SVMs using RBF kernels. Some of the features provided to the SVM include temporal, prosodic, and cepstral qualities. SVMs are designed to operate effectively with sparse instances and high-dimensional data, according to Onu et al.'s (2017) comparison of SVM to other nonlinear classifiers, such as neural networks, on the classification of asphyxia. The incremental SVM learning approach was used by Chang et al. in 2015 to categorize baby screams using FFT characteristics. During each training phase, this model continuously adds fresh data to the dataset, increasing accuracy by more than 18% compared to the original SVM model.[7]

4. IMPLEMENTATION DETAILS

Five different cry sound recordings are labeled in this study in connection to the Dustan baby language. The datasets are therefore marked with the terms "hunger," "uncomfortable," "lower gas," "burp," and "sleep." For every labeled cry sound, there are 50 identical datasets. The dataset's noise is removed using the Audacity application. After that, the datasets are trained and tested using machine learning techniques. The next step involves identifying a crying sound using a manually entered recording of a sound. Therefore, we employed a total of 250 datasets to predict the outcome of this s. The live recording of a baby's screams, which is collected as input from a microphone and has a threshold amplitude value of 0.8 for every five seconds, will then be analyzed using the SVM algorithm. The reason for the infant's crying will then be displayed on the screen or sent through email to the responsible parent or guardian. Depending on the outcomes, this will make it much easier for parents or other carers to keep an eye on their children. Try using the Support Vector Machine (SVM).

5. RESULTS

This study describes the creation of Baby Cry Sound Pattern Detection and Classification for Infant Monitoring. The classifier used is SVM, while the feature extractor is MFCC. This test makes use of a five-second real-time recording of a baby sobbing. With an accuracy of 92.12%, the trained model can recognize the sound of a baby sobbing. While accuracy fell below 60% when we used alternative classifiers like logistic regression and random forest. SVM is a better option for accurate infant cry sound recognition with fewer datasets, as shown by the above result.

6. CONCLUSION AND FUTURE SCOPE

This study implements the identification and classification of baby cry sound patterns for infant monitoring using MFCC as a feature extractor and SVM as a classifier. The accuracy of the baby cry detection system is 92.12%.

Future detection and classification of new born cry sound patterns hold significant promise for improving baby safety and well-being. The potential to develop trustworthy systems that can automatically recognize and categorize different types of baby cry sounds is interesting thanks to technological developments in audio signal processing and machine learning techniques. Wearables and baby monitors are only two examples of infant monitoring devices that can incorporate this technology, giving parents and other carers access to real-time alerts and data on their child's needs and well-being. Additionally, this technology may be utilized to detect specific medical conditions or abnormalities in infant vocalizations, going beyond simple monitoring and assisting medical professionals in early diagnosis and intervention. Overall, the identification and classification of baby cry sound patterns are improving, providing parents and medical professionals with helpful tools for the care and wellness of infants.

REFERENCES:

1. Anand H.Unnibhavi, D.S.Jangamshetti, Shridhar K. " Triphone Model Based Novel Kannada Continuous Speech Recognition System using Kaldi Tool" July 2020 International Journal of Innovative Technology and Exploring Engineering 9(9):452-458 DOI:10.35940/invitee.I7210.079920.
2. "Infant Crying Classification by Using Genetic Algorithm and Artificial Neural Network" in 2020 by Azadeh Bashiri, Roghaye Hosseinkhani.
3. "IoT Based Smart Cradle System with an Android App for Baby Monitoring", 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA).
4. "Sudden Unexpected Infant Death and Sudden Infant Death Syndrome", Centers for Disease Control and Prevention. Centers for Disease Control and Prevention 17 Apr. 2017, May 2017
5. Asthana S, Varma N, Mittal VK. An investigation into a classification of infant cries using modified signal processing methods. 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015. USA: IEEE, 2015
6. A review of infant cry analysis and classification Chunyan Ji, Thosini Bamunu Mudiyanselage, Yutong Gao & Yi Pan EURASIP Journal on Audio, Speech, and Music Processing volume 2021, Article number: 8 (2021)
7. K A, Vincent PMDR, Srinivasan K, Ashwini K, Chang CY. Deep Learning Assisted Neonatal Cry Classification via Support Vector Machine Models. Front Public Health. 2021 Jun 10;9:670352. doi: 10.3389/fpubh.2021.670352. PMID: 34178926; PMCID: PMC8222524.
8. C. A. Bratan et al., "Dunstan Baby Language Classification with CNN," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 2021, pp. 167-171, doi: 10.1109/SpeD53181.2021.9587374.