ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888707

Exploring the Differences between Deaf and Hearing Infant Cries

Enjamamul Hoq
University at Buffalo
Buffalo, USA
ehoq@buffalo.edu

Ifeoma Nwogu
University at Buffalo
Buffalo, USA
inwogu@buffalo.edu

***Abstract*—In this study, we propose an interpretable AI approach for analyzing infant cry signals, focusing on distinguishing between deaf and hearing infants. Using the Baby Chillanto dataset, we first conduct a human test to determine how well humans can detect the differences between hearing and deaf infants. We then explore the use of typical features used in audio signal analysis - chromagram, Log-Mel spectrogram and Mel frequency cepstral coefficients (MFCC), to determine whether a deep learning model can perform as well or better than humans. To explain the model's predictions, we apply SHapley Additive exPlanations(SHAP) which reveal the most significant portion of each feature set that contributes to the model's decisions. This interpretable approach provides valuable insights into how acoustic characteristics differ between deaf and hearing infants, potentially aiding early detection and understanding of hearing impairments through non-invasive cry signal analysis. While the combined human test results were near random, our deep learning approach demonstrated both high classification performance and enhanced explainability.**

***Index Terms*—Interpretable AI, infant cry, deaf, hearing, human study.**

I. Introduction

Early vocalizations, including infant cries, are considered precursors to speech and language development later in life. [1]. If the crying patterns of deaf infants can be clearly delineated from those of hearing infants, early interventions focusing on alternative communication methods (such as the teaching of sign language) could begin at a very young age, to mitigate any delays in education and support optimal development [2]. While it is possible that experienced parents, doctors, and nurses can occasionally identify the different crying patterns, from our empirical study, laypeople who have had some marginal training cannot readily tell the difference.

In this work, therefore, we propose an approach that explores various audio features and determines the most relevant one for delineating deaf from hearing infant cries, when passed through a deep learning classifier. We explain our findings in this approach and anticipate the use of a basic AI model such as this, for use in clinical settings, for early interventions.

A. Related Work

In 1940s, initial work on infant cry and understanding of child development started focusing on pitch level and voice breaks [3]. In 1960s, four different types of infant cries (birth, pain, hunger, pleasure) were identified based on the ability to recognize preverbal infant vocalizations [4]. Infants physical and emotional states were monitored by analyzing their cry sounds and ten distinct cry modes were identified based on time-frequency patterns using spectrograms [5]. Further analysis of this work extended from normal infant's cry signals to diagnose health conditions in infants [6]. This work correlated dysphonation and hyperphonation cry modes with the pathological cry signals through spectrographic analysis. Previous studies on infant cry analysis were manual tasks and mostly relied on physicians and professionals. The advent of machine learning has automated the process of classifying various infant cry signals by implementing various state-of-the-art machine learning algorithms. [7]. Support Vector Machine (SVM) algorithm is used to classify infant cry using Fundamental frequency($F0$), Mel Frequency Cepstral Coefficients (MFCC), and Constant-Q Chromagram (CQC) [8], [9]. Gaussian mixture model(GMM) [10], K-means clustering [11] are also recently used for infant cry classification. Feed Forward Neural Network (FFNN) based model is used for classifying pathological cries using MFCC features [12]. In 2019, Convolutional Neural Network(CNN) based transfer learning strategy was used on spectrograms of Baby Chillanto database to achieve promising results [13]. Recently, pre-trained models such as whisper [14] have also gained attention in classifying infant cry patterns. Leveraging latent representations from the whisper encoder module outperformed CNN and Bidirectional Long-Short Term Memory (Bi-LSTM) based network using MFCC features [15].

B. Contribution

In this paper, unlike other approaches working in the area of infant vocalizations that confound various types of cries - hunger, pain, asphyxia, etc., we focus only on deaf and hearing infants, taking a deep dive into explaining the nuanced differences between the crying sounds produced by the two classes of infants. Although we provide a robust classification approach, our main contributions are the results of the human study along with the interpretations derived from the outputs of the learning network.

II. The Study Approach

In this section, we describe our approach to understanding how well humans with a marginal level of training can delineate between the cries of deaf and hearing infants. We then visualize various acoustic features to determine the most informative feature for use in a clinical setting to provide real-time feedback.

A. Dataset

We utilized the *Baby Chillanto* dataset [16] which is the property of Mexico's NIAOE-CONACYT. It has healthy and pathological infant cry signals for cry classification. The dataset comprises hearing[1], deaf, hungry, pain, and asphyxia cry signals. However, for the purpose of this study, we specifically focused on the subset of the dataset that contained cry signals from deaf and hearing infants, to investigate the differences in acoustic features between the two groups. The Baby Chillanto dataset is particularly useful for this research because it provides audio recordings labeled based on the specific condition of the baby. It allows a detailed analysis of infant cry behaviors. The dataset includes recordings captured in controlled environments having a sample rate of 8KHz. A total of 1392 samples were used, with 885 samples from deaf infants and 507 samples from hearing infants.

B. The Human Test

To conduct the human test, we randomly selected 10 audio samples, five hearing and five deaf as the training data. The remaining data was treated as the pool of test data. We recruited 10 study participants, 3 male and 7 female, predominantly computer science graduate students. Each participant was initially "trained" to hear the differences between the two classes using the 10 pre-selected training samples. They were then played 10 cry signals randomly selected from the dataset. The same number of hearing and deaf cries were played to each participant, but the order was randomized. The participants recorded their selections and a summary of their results is presented in Fig. 1.

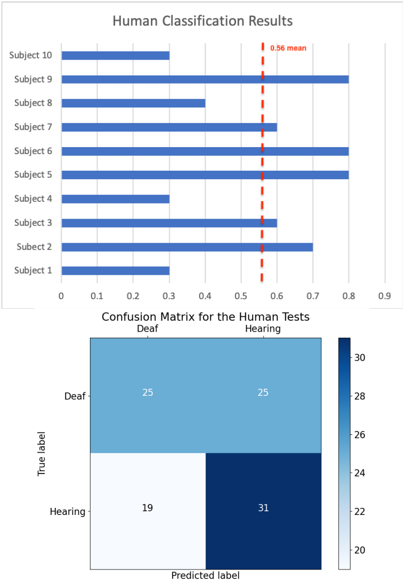


Fig. 1. Top: Per-subject accuracy from the human tests; bottom: overall human test confusion matrix for the 10 subjects over 100 tests.

Discussion on The Human Tests: The human judgments demonstrate that to the untrained or marginally trained ear, it is challenging to decipher the vocalizations, especially the crying sounds of deaf versus hearing infants. There were several individuals though that scored as high as 80 % accuracy suggesting that human ability may not necessarily be random, even though the F-1 score was only 0.532 for a binary classification task. It will be useful going forward to perform a more in-depth human study, to determine if gender, age, motherhood, etc might have played a role in the individual performance scores.

C. Acoustic Input Features

We examined three features that are commonly used in audio signal processing to capture essential characteristics of the sound: Log-Mel spectrograms, Mel-frequency cepstral coefficients (MFCCs), and Chroma.

The Log-Mel spectrogram is a time-frequency representation where the frequency axis is mapped to the Mel scale, which closely aligns with human auditory perception. In this work, the Log-Mel spectrograms were computed using a window size of 1024 samples and a hop length of 256 samples. A total of 128 Mel bands were used, and a logarithmic transformation was applied to the spectrogram to normalize the values. The MFCCs are another widely used feature in audio recognition, representing the short-term power spectrum of the sound. From each cry signal, we extracted 26 MFCC coefficients using the same window and hop lengths as used for the Log-Mel spectrograms. The MFCCs were normalized to a [0, 1] range for input into the model.

Lastly, we also explored the use of Chroma, another acoustic representation useful in sound analysis to refer to a representation of the twelve different pitch classes (such as C, C#, D, etc.) in the musical octave. Chroma features capture the energy or presence of each of the twelve pitch classes, regardless of the octave in which they occupy. We opted to include Chroma features in our analysis because the infants are non-verbal and oftentimes, their vocalizations could be perceived as somewhat musical in nature [17].

Each feature was extracted using Librosa [18] and to ensure uniform input size, acoustic features were resized to a 224×224 shape compatible with the neural network model. We present the T-SNE(t-distributed Stochastic Neighbor Embedding) [19] projections of the deaf and hearing cry samples in Fig. 2. We used T-SNE to project the high-dimensional acoustic feature representations into a 2D space using the scikit-learn library [20].

1) Discussion on Acoustic Features: Although we hypothesized that the musicality of infant vocalizations could be a driving force in delineating between deaf and hearing infant cries, from visualizing the feature projections, this was not the case. Log-Mel spectrograms were observed to provide greater

[1] The Baby Chillanto dataset originally labeled infant cries as "normal," but we updated it to "hearing" for inclusivity.

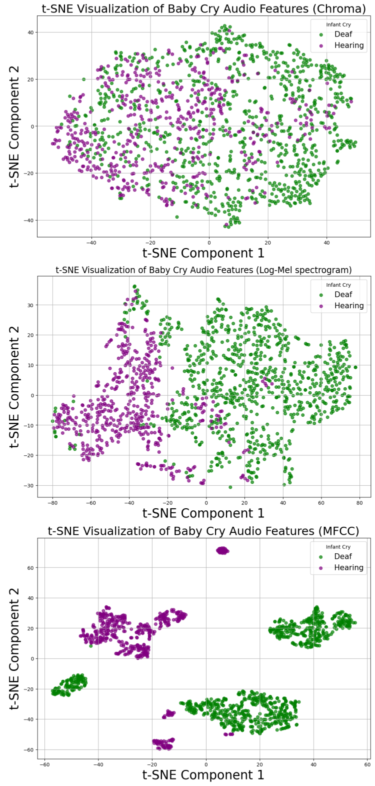


Fig. 2. Top: T-SNE plots of the input types extracted for deaf versus hearing infants; top: Chroma, middle: Log-Mel spectrogram and bottom: MFCC.

discriminability than Chroma features and MFCC provided the clearest separation as shown in Fig. 2. Visually, MFCC was observed to have the most separation between the two classes, but quantitatively, Log-Mel spectrograms performed best when fed as input to the deep learning classifier. It is possible that the MFCC features produces the best visual separation because it is a compact representation of the spectral envelope of the Log-Mel spectrogram. It extracts only the most important features from the Log-Mel spectrogram by computing a set of coefficients (typically 13–40 coefficients) that summarize the spectral information. But when the entire spectrogram is fed into a deep learning model, it makes its own selections depending on which aspects best minimize the loss function.

D. Training Details:

The classifier was trained using the Adam optimizer with a learning rate of 0.001, and a batch size of 16. The dataset was split into training, validation, and test sets in an 80:10:10 ratio. The model was trained for 100 epochs with early stopping to prevent overfitting. Training was carried out on an NVIDIA GeForce RTX 3090 24GB RAM GPU, For each experiment, the model architecture was tailored to accommodate the specific input features.

III. Results & Discussion

In this section, we implement a basic deep learning-based classification model with the intent of observing what aspects of the input data best explain the classification results. We discuss our findings at each stage of the study.

A. CNN-Based Classification

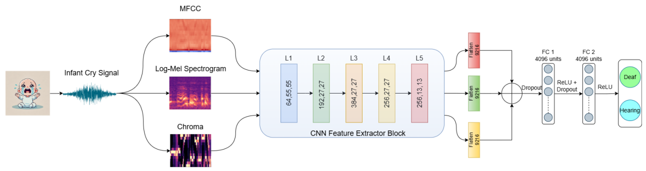
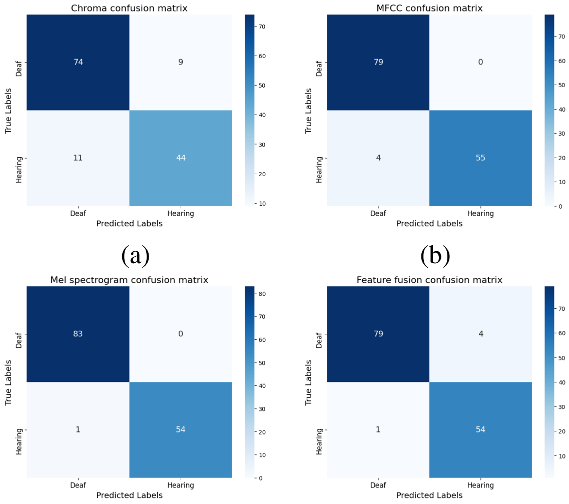


Fig. 3. Infant cry classification fusion model framework.

The binary classifier used consisted of three identical branches where each branch was responsible for extracting features from the Chroma, MFCC and Mel-spectrum inputs respectively. Each branch involved a CNN consisting of six layers: five convolutional layers with ReLU activation and max pooling, followed by two fully connected layers. The convolutional layers were primarily responsible for feature extraction, while the fully connected layers performed the classification. We initially trained each feature branch separately to obtain their individual classification results and subsequently fused all three branches by concatenation for comprehensive classification results with all inputs. Fig. 4 and Table I present our results from CNN-based classification task.

1) Discussion on The Classification Results: At first blush, from viewing the confusion matrices in Fig. 4, the CNN classifier performs extremely well on all input types provided, especially when compared to the human test result shown in Fig. 1 (bottom). Drilling down into the details as shown in Table I, the CNN-based classification far out-performs the marginally trained humans, where the CNN with the least separable auditory input type (Chroma) still had an almost 30% improvement over human judgments.



(a) (b)

(c) (d)

Fig. 4. Confusion matrices from the deaf versus hearing infant cry classification task using different input features. (a) Chroma; (b) MFCC; (c) Log-Mel spectrogram; and (d) all three input features fused.

Although the Log-Mel spectrogram performed best, when conducting the explainability study and testing on all three features, we found MFCC to be the most reliable, showing consistent results across samples. For this reason, we present our explainability results with MFCC inputs.

TABLE I
DEAF VERSUS HEARING INFANT CRY CLASSIFICATION RESULTS USING DIFFERENT INPUT TYPES. (A) CHROMA; (B) MFCC; (C) LOG-MEL SPECTROGRAM; AND (D) ALL THREE INPUTS (IN 3-BRANCH CNN).

Input Type	**Accuracy**	**Precision**	**Recall**	**F-1 Score**
Human	0.56	0.568	0.5	0.532
Chroma	0.855	0.850	0.845	0.847
MFCC	0.978	0.978	0.975	0.977
Log-Mel Spectrogram	0.992	0.994	0.990	0.992
All Inputs	0.963	0.959	0.966	0.962

B. SHAP (SHapley Additive exPlanations)

We use the SHAP values to better explain the behavior of the classification network. Shapley value is used in game theory to evaluate the contribution of each player in a coalition of players in a cooperative game [21]. The notion of using these values for interpreting neural network models was proposed by Anacona *et al*. [22]. Their use in additive explanations (SHAP) was presented by Lundberg *et al*. [23]. For a particular model prediction $f(x)$, the Shapley value for feature i is given by:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right] \quad (1)$$

where F is the set of all features, and S represents a subset of those features. SHAP calculates how the inclusion of feature i changes the model's prediction by comparing the output with and without that feature. In this work, SHAP is applied to both Log-Mel spectrogram and MFCC representations, allowing us to understand how each feature (such as frequency bands or cepstral coefficients) contributes to the model's decision of whether a cry is from a deaf or hearing baby. Examples of our findings are shown in Fig. 5.

1) Discussion on The SHAP Outcomes: From reviewing the SHAP results, we observed that the top horizontal portion representing high-frequency components in the MFCC inputs was consistently the most prominent feature responsible for the classification. When the infants were deaf (top images in the 2 pairs), the Shapley values in the top region of the MFCC were high for the deaf class and vice versa, for the hearing class. We observed this pattern across many samples reviewed. Interestingly, the Y-axis of an MFCC represents the frequency coefficients, where the lower values relate to the broad spectral shape of the sound, capturing the most important aspects of the spectral envelope. This is the region that is most perceptible to humans. The higher coefficients capture more subtle variations and higher frequency components. The observation that the classifier focuses on this higher frequency region might help explain why humans have a harder time delineating the signals.

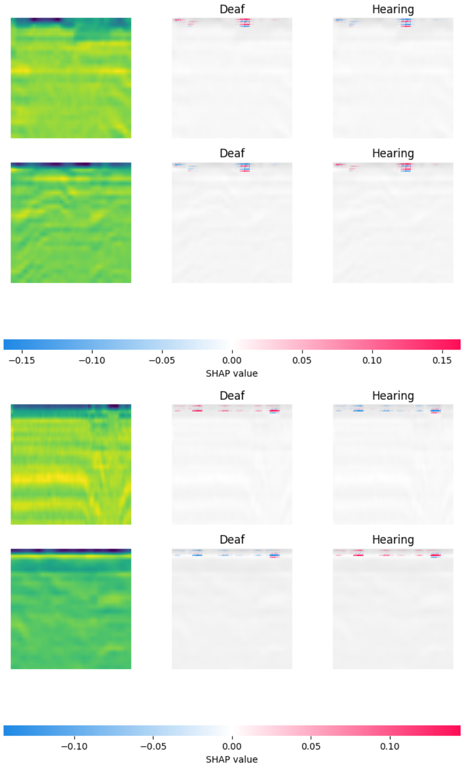


Fig. 5. Examples on SHAP output for two pairs of MFCC inputs. Each pair has a deaf (top) and hearing (bottom) infant signal.

This finding is consistent with the results reported in a 1985 study by Oller et al. [24], where spectrographic displays were used to analyze the differences between the prespeech vocalizations of deaf and hearing infants, and they did indeed find clear differences.

IV. CONCLUSION

We have presented an interpretable AI approach for analyzing infant cry signals, focusing on distinguishing between deaf and hearing infants. We perform tests to compare human judgments with machine classifications and discover that while humans in general have a challenging time delineating the crying acoustics, a CNN-based model provides almost perfect scores. We also observe that the model focuses on the high-frequency regions of the MFCC inputs, during classification. This interpretable approach can provide insights into how acoustic characteristics differ between deaf and hearing infant cries, potentially aiding early detection and understanding of hearing impairments through non-invasive cry signal analysis.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. 2223507. The infant image was created with the assistance of DALL·E 2. We like to thank Dr. Carlos A. Reyes-Garcia, Dr. Emilio Arch-Tirado and his INR-Mexico group, and Dr. Edgar M. Garcia-Tamayo for their dedication of the collection of the Infant Cry data base.

References

[1] D. K. Oller, G. Ramsay, E. Bene, H. L. Long, and U. Griebel, “Protophones, the precursors to speech, dominate the human infant vocal landscape,” *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1836, p. 20200255, 2021.

[2] M. D. Clark, K. R. Cue, N. J. Delgado, A. N. Greene-Woods, and J.-L. A. Wolsey, “Early intervention protocols: Proposing a default bimodal bilingual approach for deaf children,” *Maternal and Child Health Journal*, vol. 24, pp. 1339–1344, 2020.

[3] G. Fairbanks, J. H. Wiley, and F. M. Lassman, “An acoustical study of vocal pitch in seven-and eight-year-old boys,” *Child Development*, pp. 63–69, 1949.

[4] O. Wasz-Höckert, T. Partanen, V. Vuorenkoski, K. Michelsson, and E. Valanne, “The identification of some specific meanings in infant vocalization,” *Experientia*, vol. 20, pp. 154–154, 1964.

[5] Q. Xie, R. K. Ward, and C. A. Laszlo, “Automatic assessment of infants’ levels-of-distress from the cry signals,” *IEEE transactions on speech and audio processing*, vol. 4, no. 4, p. 253, 1996.

[6] H. A. Patil, ““cry baby”: Using spectrographic analysis to assess neonatal health status from an infant’s cry,” *Advances in speech recognition: Mobile environments, call centers and clinics*, pp. 323–348, 2010.

[7] C. Ji, T. B. Mudiyanselage, Y. Gao, and Y. Pan, “A review of infant cry analysis and classification,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 8, 2021.

[8] G. Z. Felipe, R. L. Aguiar, Y. M. Costa, C. N. Silla, S. Brahnam, L. Nanni, and S. McMurtrey, “Identification of infants’ cry motivation using spectrograms,” in *2019 International conference on systems, signals and image processing (IWSSIP)*. IEEE, 2019, pp. 181–186.

[9] M. Huckvale, “Neural network architecture that combines temporal and summative features for infant cry classification in the interspeech 2018 computational paralinguistics challenge,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech Communication Association (ISCA), 2018, pp. 137–141.

[10] I.-A. Bănică, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu, “Automatic methods for infant cry classification,” in *2016 International conference on communications (COMM)*. IEEE, 2016, pp. 51–54.

[11] K. Sharma, C. Gupta, and S. Gupta, “Infant weeping calls decoder using statistical feature extraction and gaussian mixture models,” in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019, pp. 1–6.

[12] A. Zabidi, L. Y. Khuan, W. Mansor, I. M. Yassin, and R. Sahak, “Classification of infant cries with asphyxia using multilayer perceptron neural network,” in *2010 Second International Conference on Computer Engineering and Applications*, vol. 1. IEEE, 2010, pp. 204–208.

[13] L. Le, A. N. M. Kabir, C. Ji, S. Basodi, and Y. Pan, “Using transfer learning, svm, and ensemble classification to classify baby cries based on their spectrogram images,” in *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*. IEEE, 2019, pp. 106–110.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[15] M. Charola, A. Kachhi, and H. A. Patil, “Whisper encoder features for infant cry classification,” in *Proc. INTERSPEECH*, vol. 2023, 2023, pp. 1773–1777.

[16] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García, “Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies,” in *2008 Seventh Mexican international conference on artificial intelligence*. IEEE, 2008, pp. 330–335.

[17] K. Wermke, M. P. Robb, and P. J. Schluter, “Melody complexity of infants’ cry and non-cry vocalisations increases across the first six months,” *Scientific reports*, vol. 11, no. 1, p. 4137, 2021.

[18] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python.” in *SciPy*, 2015, pp. 18–24.

[19] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[21] A. E. Roth, “Introduction to the shapley value,” *The Shapley value*, vol. 1, 1988.

[22] M. Ancona, C. Oztireli, and M. Gross, “Explaining deep neural networks with a polynomial time algorithm for shapley value approximation,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 272–281.

[23] I. Covert, S. M. Lundberg, and S.-I. Lee, “Understanding global feature contributions with additive importance measures,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 212–17 223, 2020.

[24] D. K. Oller, R. E. Eilers, D. H. Bull, and A. E. Carney, “Prespeech vocalizations of a deaf infant: A comparison with normal metaphonological development,” *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 1, pp. 47–63, 1985.