ANALYSIS AND CLASSIFICATION OF NORMAL *VS.* PATHOLOGICAL INFANT CRIES

Hiya Chaudhari and Hemant A. Patil
{202101047, hemant_patil}@daiict.ac.in

Speech Research Lab @ Dhirubhai Ambani University (DAU formerly DA-IICT), Gandhinagar, Gujarat, India

***Abstract*—Infant cry analysis serves as a non-invasive diagnostic tool for detecting pathological conditions. This study analyzes infant cry acoustics using formant distribution analysis in 3-D across Mel, Bark, Cochlear and Gammatone scales, visualizing fundamental frequency (F_0), and the first two formants (F_1), and (F_2) for normal vs. pathological cry classification. The dataset comprises cries from Baby Chillanto, Baby Chillanto 2.0, and DA-IICT corpora, covering six pathologies. Harmonics-to-Noise Ratio (HNR), jitter, and shimmer are examined to assess phonatory irregularities, while violin plots highlight statistical dispersion in cry variability. Formant Space Area (FSA) via convex hull analysis quantifies spectral dispersion, offering a novel biomarker for pathology differentiation. For classification, both handcrafted spectral features (Mel Frequency Cepstral Coefficients, Gammatone Frequency Cepstral Coefficient, Cochlear Filter Cepstral Coefficients and Bark Frequency Cepstral Coefficients) and deep learning-based embeddings (HuBERT, wav2vec 2.0, XLS-R) are evaluated using CNN and Bi-LSTM models. BFCC outperforms MFCC, CFCC and GFCC, while deep learning embeddings enhance classification accuracy. Beyond classification, radar plots, convex hull mapping, and 3-D formant visualizations provide deeper insights into cry-based pathology detection.**

***Index Terms*—Formants,HNR, shimmer,Spectral Features, DL Models.**

I. INTRODUCTION

Infant cries serve as an early indicator of neurological and physiological conditions, providing crucial insights into a newborn's health status. Unlike adult speech, infant cries are reflexive vocalizations influenced by both neurological development and physiological constraints, making their acoustic properties valuable for diagnosing underlying pathologies. Traditional infant cry assessments are subjective, but advances in acoustic analysis and deep learning enable objective, automated pathology detection.

This study presents a comprehensive acoustic analysis of infant cries, examining formant distribution across four perceptual scales—Mel, Bark, Cochlear, and Gammatone—to assess how pathological conditions alter cry characteristics. Unlike previous studies relying on MFCCs [1] or conventional frequency analysis, this work explores multi-scale formant variations for a deeper perceptual understanding. Fundamental frequency (F_0), formant frequencies F_1,F_2, and voice quality parameters such as Harmonics-to-noise ratio (HNR), jitter, and shimmer are analyzed to quantify phonatory irregularities linked to medical conditions.

The dataset used in this study is compiled from Baby Chillanto, Baby Chillanto 2.0, and the DA-IICT Infant Cry dataset, and a combination of the two, covering six pathological condition.

Compared to existing studies on infant cry [2], this work explores handcrafted features and deep learning features as well for cry classification achieving higher accuracy and better pathology differentiation. Additionally, the use of convex hull-based formant space analysis provides a novel diagnostic marker, improving interpretability. The inclusion of Bark-scale features further enhances separability, offering a more perceptually relevant representation than traditional Mel or Gammatone scales.

II. TECHNICAL APPROACH

A. Voice Quality Analysis

Pathological cries often exhibit instability in pitch i.e. [3] and amplitude, which can be quantified using jitter, shimmer [4], and HNR [5] .

1) Jitter (Pitch Variation): Jitter measures the cycle-to-cycle frequency variation and is computed as:

$$J_{\text{local}} = \frac{\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{(N-1)\cdot \bar{T}}, \tag{1}$$

where T_i is the pitch (or fundamental) period of the i^{th} glottal cycle, and $\bar{T}$ is the mean period. High jitter values indicate greater irregularity in F_0, associated with pathological cries.

2) Shimmer (Amplitude Variation): Shimmer quantifies cycle-to-cycle amplitude fluctuations:

$$S_{\text{local}} = \frac{\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{(N-1)\cdot \bar{A}}, \tag{2}$$

where A_i is the amplitude of the i^{th} cycle, and $\bar{A}$ is the mean amplitude. Increased *shimmer* values indicate instability in vocal intensity.

3) Harmonics-to-Noise Ratio (HNR): It represents the ratio of harmonic (periodic) energy to noise energy:

$$HNR = 10\log_{10}\left(\frac{P_{\text{harmonic}}}{P_{\text{noise}}}\right), \tag{3}$$

where P_{harmonic} is the power of harmonic components, i.e. integral multiple of fundamental frequency (F_0) and P_{noise} is the power of noise components. Lower HNR values indicate greater breathiness or noisiness, common in asphyxiated or neurologically impaired infants.

III. Experimental Setup

In this Section, we describe the experimental setup and dataset used for the classification of normal *vs.* pathological infant cries.

A. Datasets Description

This study utilized multiple infant cry datasets to ensure robust classification performance. The datasets included Baby Chillanto [6] (D1), containing normal and pathological cries, and Baby Chillanto 2.0 [7], an extended version with additional samples as shown in Table I. Additionally, the DA-IICT Infant Cry Corpus (D2) [8], an in-house dataset was used. A combined dataset, D3, was formed by merging D1 and D2 comprising of 6 pathalogies.

B. Feature Extraction and Representation

1) Hand-crafted Features: Feature extraction was performed to derive meaningful representations of infant cries. The primary feature sets included Mel-Frequency Cepstral Coefficients (MFCCs) [9], which capture the spectral envelope of cry signals. Gammatone Frequency Cepstral Coefficients (GFCCs) [10] were utilized to incorporate auditory model-inspired features, enhancing classification performance. Bark scale-based features were extracted to represent frequency information in a perceptually relevant manner. Additionally, Cochlear Frequency Cepstral Coefficients (CFCCs) [11] were derived to simulate cochlear frequency mapping, offering an alternative biologically inspired representation of infant cry acoustics.

The Bark scale models human auditory perception and is given by:

$$B(f) = 13\arctan(0.00076f) + 3.5\arctan\left(\left(\frac{f}{7500}\right)^2\right), \tag{4}$$

where f is the frequency in Hz, and $B(f)$ is the frequency in the Bark scale.

2) DL Models: Feature extraction was also performed using self-supervised deep learning models to derive high-level representations of infant cries. wav2vec 2.0 [12], HuBERT [13], and XLS-R [14] models were employed to capture rich temporal and spectral structures beyond traditional handcrafted features. wav2vec 2.0 leverages contrastive learning to extract phonetic and acoustic variations directly from raw waveforms, while HuBERT utilizes masked speech modeling to learn discrete unit-based representations, improving robustness in pathology differentiation. XLS-R, a multilingual extension of wav2vec 2.0, was explored in three model sizes (300M, 1B, and 2B). These deep representations were compared with handcrafted features, demonstrating superior performance in distinguishing pathological cries [15]. All audio recordings were resampled to 16 kHz, with silence removed and amplitude normalized.

The dataset was split into 80% training, 10% validation, and 10% testing, ensuring balanced pathology distribution. Spectrograms were extracted for handcrafted features, while raw waveforms were used for deep learning models. Feature extraction included log-energy normalization and segmentation into fixed-duration frames for consistency.

TABLE I
Distribution of Pathologies in D1 and D2

Class	Baby Chillanto 2.0	DA-IICT
Asphyxia	340	-
Deaf	879	-
Asthma	-	215
Hypoxic-Ischemic Encephalopathy	-	182
Hypothyroidism	47	-
Hyperbilirubinemia	9	-

C. Classifiers

The classifiers used in this study include Convolutional Neural Network (CNN) [16] and a Bi-directional Long Short-Term Memory (Bi-LSTM) network. For CNN, an Adam optimizer was used with a learning rate of 0.003. The input shape was set to 20 × 893 × 1. The Bi-LSTM model was trained with a learning rate of 0.003, a batch size of 32, and a hidden size of 256. Dropout layers with a dropout rate of 25% were applied to prevent overfitting. In both models, the input feature dimension was fixed at 20 for consistency across all samples.

IV. Acoustic Analysis of Infant Cries

This section analyzes infant cry acoustics and perception, focusing on formant distribution, voice quality (jitter, shimmer, HNR), and perceptual scales (Mel, Bark, Cochlear, Gammatone). Variations in F_1 and F_2 are visualized using 3-D scatter and violin plots. Deep learning models (wav2Vec2, HuBERT, XLS-R) enhance classification accuracy, while formant-based analysis offers interpretable, clinically relevant insights for early diagnosis [6].

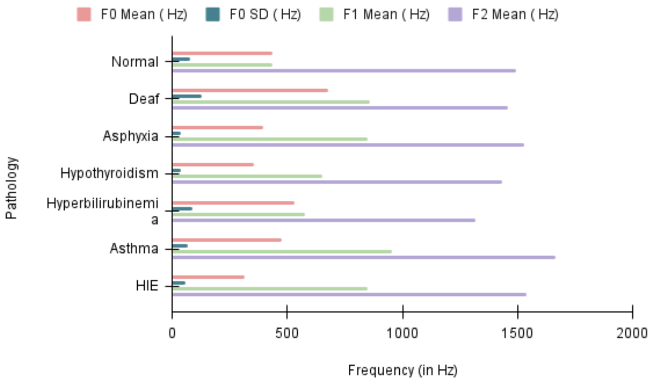

Fig. 1. Acoustic Feature Statistics of Infant Cries Across Different Conditions.

A. Analysis of F_0 and Voice Quality Measures

Fig.1 and Fig.2 represent the acoustic parameters F_0 mean, F_0 standard deviation, formants F_1 and F_2, jitter, shimmer, and

HNR measured infant cries across normal and pathological cries. Normal cries exhibit the highest HNR (9.16 dB) and lowest jitter (0.15%) and shimmer (0.18%), indicating stable phonation. In contrast, asphyxia, hypothyroidism, and hyperbilirubinemia show lower HNR and increased jitter/shimmer, suggesting aperiodic, weak cry signals. Deaf infants produce the highest mean F_0 (674.43 Hz), likely due to lack of auditory feedback. Normal cries show broad F_1 and F_2 distributions, whereas asthma and HIE exhibit more constrained patterns, indicating vocal tract shift in resonance. Asphyxia and deafness show wider F_1 and F_2 variability, reflecting unstable articulation. Hyperbilirubinemia and hypothyroidism display skewed formant distributions, suggesting irregular cry patterns.Pathological cries generally show more variation in formant frequencies, highlighting irregular vocalization patterns. Beyond voice quality markers such as jitter and shimmer, formant frequency shifts further characterize cry instability, making them crucial for pathology detection.

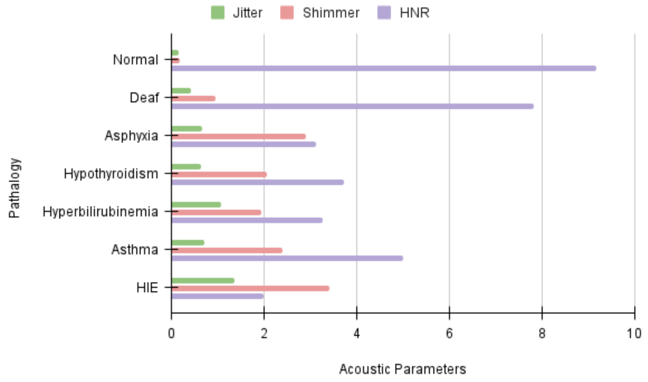

Fig. 2. Acoustic Feature Statistics of Infant Cries Across Conditions.

B. Formant Space and Trajectory Analysis

Fig.3 is a radar plot that provides a comparative view of F_0 , F_1 , and F_2 across multiple conditions [17].

Notably, asphyxia and hyperbilirubinemia exhibit wider formant areas, while asthma and HIE display more compact distributions. Fig. 4 further supports these findings, showing violin plots of F_1 and F_2 distributions, where asphyxia and deafness have broader variability, while asthma remains compact. These visualizations highlight formant structure differences across pathologies, providing valuable biomarkers for cry-based classification. Deaf infants exhibit higher F_0 values, suggesting pitch regulation deficits due to impaired auditory feedback, while hyperbilirubinemia presents reduced formant variability, reflecting restricted cry articulation. These visualizations highlight key acoustic markers that differentiate normal and pathological cries, aiding in early diagnosis.

Fig. 5 (formant trajectory) compares the temporal variation of F_1 and F_2 between normal and asphyxiated infants [18]. Normal cries show relatively stable formant patterns, whereas asphyxiated cries display greater fluctuations, indicating vocal instability due to respiratory distress. Fig. 6 visualizes the distribution of F_0 , F_1 , and F_2 across normal *vs.* pathological cries. Distinct clusters indicate that certain pathologies, such as asphyxia and hyperbilirubinemia, exhibit notable shifts in formant frequencies, reflecting altered vocal tract resonance. Bark scale . Among the perceptual scales used (Mel, Bark, and Gammatone), the Bark scale provided the cleanest distinction between normal and pathological cries, enhancing separability in acoustic analysis.

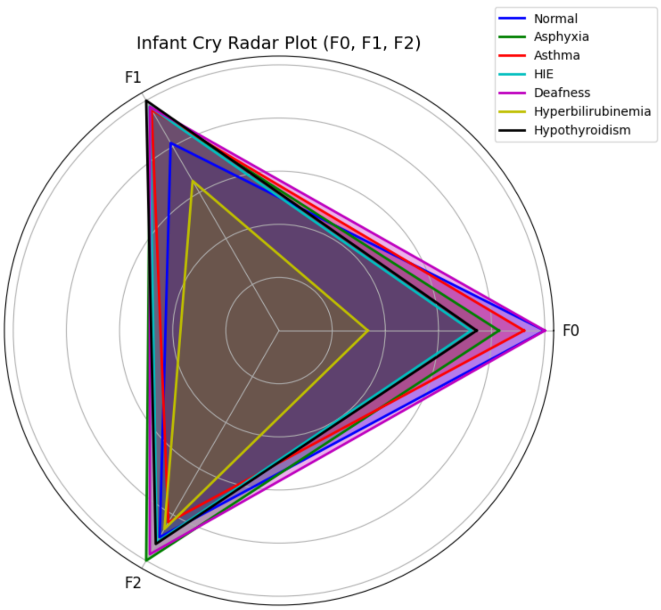

Fig. 3. Radar plot of infant cry formant features (F_0, F_1, F_2) across different infant cries.

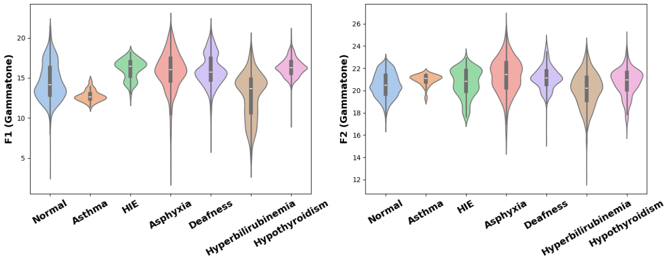

Fig. 4. Violin plot for F_1 (left) and F_2 (right) .

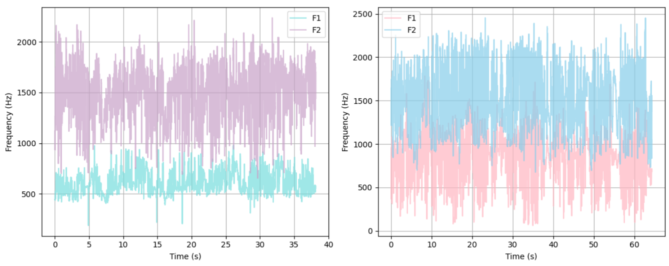

Fig. 5. Formant Trajectories of Normal *vs.* Asphyxia Infant Cries.

C. Convex Hull-Based Formant Space Area (FSA)

To quantify articulatory constraints across different cry conditions, we compute the FSA. Given a set of formant frequency coordinates $\mathbf{F} = \{(F_{1_i}, F_{2_i})\}$, the convex hull enclosing these points is computed as:

$$A_{\text{FSA}} = \frac{1}{2} \sum_{i=1}^{N} (x_i y_{i+1} - x_{i+1} y_i) \tag{5}$$

where (x_i, y_i) are consecutive points in the convex hull and N is the number of formant pairs forming the boundary Fig. 7 presents the formant space distribution of infant cries across conditions, with F_1 plotted against F_2. Convex hull-based Formant Space Area (FSA) quantifies articulatory constraints [19]. Larger FSA, observed in deafness and hyperbilirubinemia, reflects increased vowel dispersion due to altered neuromuscular control or lack of auditory feedback. In contrast, smaller FSA in HIE and asphyxia suggests restricted articulatory movement from neurological impairments. Asthma and hypothyroidism show moderate FSA, linked to respiratory and metabolic effects. Hunger cries exhibit variability due to distress, while normal cries display a balanced FSA, indicative of typical phonatory development.

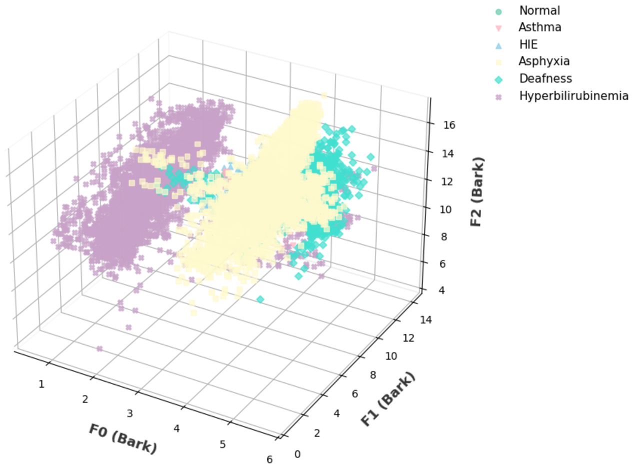

Fig. 6. Formant Distribution of Infant Cries in 3-D Bark Frequency Scale for 6 Pathologies.

V. Experimental Results

Beyond feature analysis, classification is crucial for automated pathology detection [20], [21]. Hence, we evaluate the effectiveness of both handcrafted and deep learning-based features in distinguishing normal *vs.* pathological cries.

A. Comparison Among Various Feature Sets

Table II compares classification accuracy across D1, D2, and D3 for different feature extraction techniques, highlighting the strengths of both deep learning and handcrafted features. XLS-R 300M achieves the highest accuracy due to its multilingual training, larger dataset exposure, and strong temporal modeling, effectively capturing nonlinear distortions and phonatory irregularities. However, deep learning models often lack interpretability, making handcrafted features such as BFCC and GFCC valuable for analyzing spectral patterns in a more explainable way.

TABLE II
Performance comparison of deep learning models and handcrafted feature sets for infant cry classification using Bi-LSTM and CNN classifiers

Classifiers	**Models**	**D1**	**D2**	**D3 (D1 + D2)**
Bi-LSTM	wav2vec 2.0	87.62	92.46	93.87
	XLS-R 1B	94.38	96.43	97.33
	XLS-R 2B	96.45	96.56	96.28
	XLS-R 300M	**99.68**	**98.95**	**99.94**
	HuBERT	81.23	84.89	94.62
	MFCC	96.66	88.98	92.15
	GFCC	96.58	91.65	93
	BFCC	**97.23**	**93.11**	**97.32**
	CFCC	96.79	90.43	94.01
CNN	wav2vec 2.0	87.62	92.46	93.87
	XLS-R 1B	94.38	96.43	97.33
	XLS-R 2B	96.45	96.56	96.28
	XLS-R 300M	**99.68**	**98.95**	**99.94**
	HuBERT	81.23	84.89	94.62
	MFCC	95.72	88.31	91.24
	GFCC	96	90.48	93.82
	BFCC	**97.05**	**92.45**	**95.33**
	CFCC	97	90.91	95

Among handcrafted features, BFCC outperforms others due to its Bark-scale resolution, while CFCC improves upon MFCC but lacks BFCC's perceptual optimization. The combined dataset (D3) further enhances accuracy, reducing dataset-specific biases and improving model generalization. Some conditions, such as asphyxia and HIE, remain harder to classify due to overlapping acoustic patterns, whereas deafness and hyperbilirubinemia show clearer spectral shifts, making them easier to distinguish. Overall, XLS-R 300M is the most effective feature extractor, with BFCC and GFCC as strong handcrafted alternatives. We employed two distinct DNN classifiers, namely, CNN and Bi-LSTM , in order to address any potential classifier model bias during performance evaluation. The results shown Bi-LSTM outperforms CNN.

VI. Summary of Key Findings

This study analyzes infant cries using acoustic features such as F_0, formants, jitter, shimmer, and HNR to identify pathology-specific vocal patterns. Pathological cries exhibit greater instability, with increased jitter and shimmer, whereas normal cries maintain periodic phonation.Formant space analysis shows that the Bark scale offers the clearest separation of pathologies. XLS-R 300M achieves the highest classification accuracy, while BFCC and GFCC also perform well. Combining datasets further boosts performance, highlighting the potential of acoustic features for AI-driven neonatal screening and early, non-invasive diagnosis.

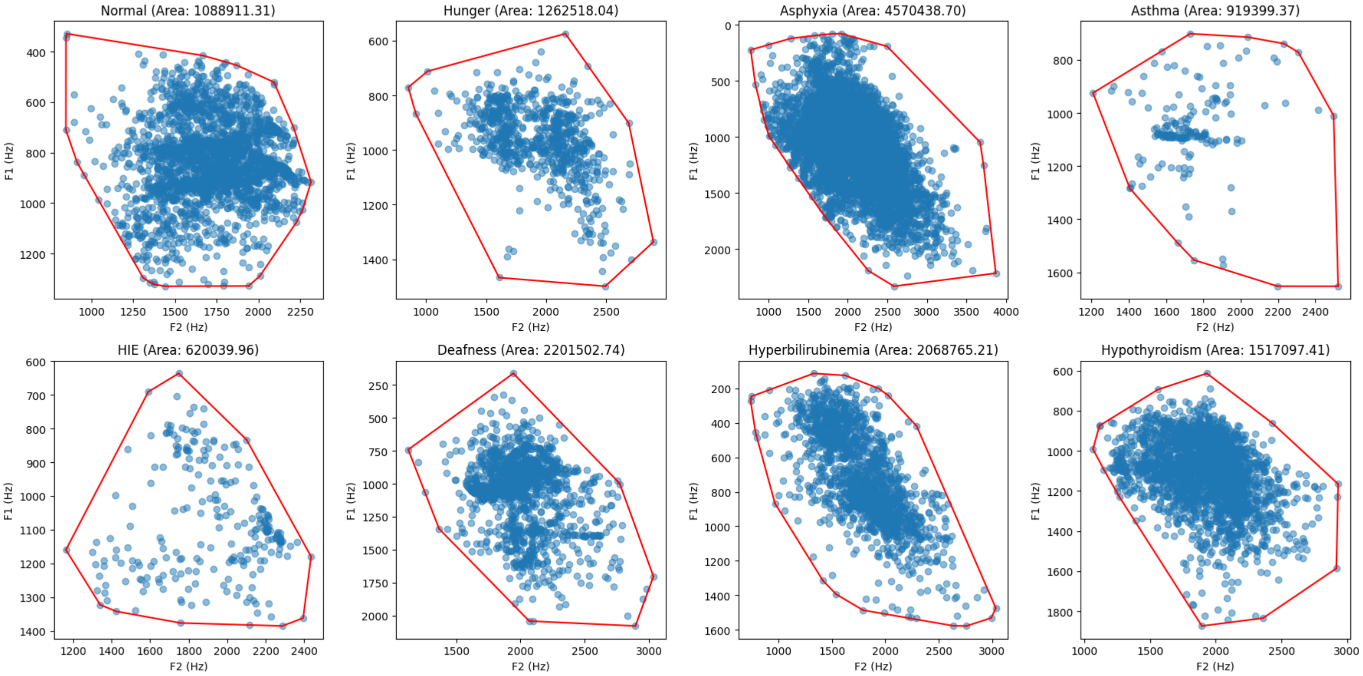

Fig. 7. Formant space area across normal *vs.* pathological cries.

References

[1] N. Meephiw and P. Leesutthipornchai, "Mfcc feature selection for infant cry classification," in *2022 26th International Computer Science and Engineering Conference (ICSEC)*, 2022, pp. 123–127.

[2] 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.

[3] 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.

[4] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition." in *Interspeech*, 2007, pp. 778–781.

[5] F. Severin, B. Bozkurt, and T. Dutoit, "Hnr extraction in voiced speech, oriented towards voice quality analysis," in *2005 13th European Signal Processing Conference*, 2005, pp. 1–4.

[6] O. F. Reyes-Galaviz, A. Verduzco, E. Arch-Tirado, and C. A. Reyes-García, "Analysis of an infant cry recognizer for the early identification of pathologies," in *International School on Neural Networks, Initiated by IIASS and EMFCSC*. Springer, 2004, pp. 404–409.

[7] 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.

[8] A. Chittora and H. A. Patil, "Data collection of infant cries for research and analysis," *Journal of Voice*, vol. 31, no. 2, pp. 252–e15, 2017.

[9] N. Meephiw and P. Leesutthipornchai, "Mfcc feature selection for infant cry classification," in *2022 26th International Computer Science and Engineering Conference (ICSEC)*. IEEE, 2022, pp. 123–127.

[10] Y. Zayed, A. Hasasneh, and C. Tadj, "Infant cry signal diagnostic system using deep learning and fused features," *Diagnostics*, vol. 13, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/12/2107

[11] S. Rathod, P. Gupta, A. Kachhi, and H. A. Patil, "Cochlear filter-based cepstral features for dysarthric severity-level classification," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 1095–1099.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," 2021.

[15] H. Chaudhari, A. J. Shah, and H. A. Patil, "Cross lingual speech representation for infant cry classification," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macao, China*, 2024, pp. 1–5.

[16] Y. Zhang, Y. Ma, and Y. Liu, "Convolution-bidirectional temporal convolutional network for protein secondary structure prediction," *IEEE Access*, vol. 10, pp. 117 469–117 476, 2022.

[17] K. Jafari, A. Tierens, A. Rajab, R. Musani, A. Schuh, and A. Porwit, "Visualization of cell composition and maturation in the bone marrow using 10-color flow cytometry and radar plots," *Cytometry Part B: Clinical Cytometry*, vol. 94, no. 2, pp. 219–229, 2018.

[18] I. Y. Özbek, M. Hasegawa-Johnson, and M. Demirekler, "Formant trajectories for acoustic-to-articulatory inversion." in *Interspeech*, 2009, pp. 2807–2810.

[19] H. Kim, M. Hasegawa-Johnson, and A. Perlman, "Vowel contrast and speech intelligibility in dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 187–194, 2011.

[20] R. Cohen and Y. Lavner, "Infant cry analysis and detection," in *2012 IEEE 27th convention of electrical and electronics engineers in Israel*. IEEE, 2012, pp. 1–5.

[21] S. Jeyaraman, H. Muthusamy, W. Khairunizam, S. Jeyaraman, T. Nadarajaw, S. Yaacob, and S. Nisha, "A review: survey on automatic infant cry analysis and classification," *Health and Technology*, vol. 8, pp. 391–404, 2018.