INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Whisper Encoder features for Infant Cry Classification

Monil Charola, Aastha Kachhi, Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology
(DA-IICT), Gandhinagar-382007, Gujarat, India

monil_charola@daiict.ac.in, aastha_k@daiict.ac.in, hemant_patil@daiict.ac.in

Abstract

Identifying the pathology in infant cry is an important and socially relevant research problem, as it can save the lives of many infants. This study proposes the use of transfer learning based approach using Whisper Encoder Module which is compared against state-of-the art MFCC feature set for classification of normal *vs.* pathological infant cry. Moreover, we also present multi-class pathological infant cry classification using CNN and Bi-LSTM networks. Our study finds that whisper encoder module coupled with DNN classifiers such as CNN and Bi-LSTM outperform MFCC features with absolute increment of 4% and 1% on CNN and Bi-LSTM respectively. Furthermore, whisper encoder features are analysed using statistical parameters and t-SNE plots. The experiments are performed using the *10*-fold cross-validation on *Baby Chillanto* dataset, *In-house DA-IICT* dataset, and also on the datasets formed by combining these two datasets.

Index Terms: Infant Cry Classification, Whisper Encoder features, Transfer Learning, CNN, Bi-LSTM.

1. Introduction

Infants primarily communicate by crying, which is frequently the first sign of their demands and general health. Within a few months of birth, millions of infants die from a variety of illnesses, starvation, and diseases that can be prevented by vaccination. Among the most frequent causes of baby mortality include asthma, asphyxia, and Sudden Infant Death Syndrome (SIDS) [1]. Many measurements, including images from head ultrasound (HUS), computed tomography (CT), and magnetic resonance imaging (MRI) scans, which show damaged areas of the brain, are used to clinically detect these disorders. However, in many underdeveloped nations, pathology detection is time consuming and expensive, which may have an effect on the infant's health because not all infants have the luxury of immediate access to healthcare and support from paediatricians.

For instance, asphyxia, one of the illnesses, can be identified by its visual manifestations, such as bluish and pale limbs. However, the infant's substantial brain damage would have already occurred by then [1, 2]. The degree of the hearing loss, length of rehabilitation type, and the age at which the pathology was detected all have an impact on the acoustical and perceptual characteristics of deaf infants. As a result, there is a growing need to create diagnostic aids that use infant cries to help paediatricians identify the first symptoms of such illnesses [3]. The study on classification of infant cries involves researchers from the physiological, neurological, paediatrics, engineering, developmental linguistics, and psychological disciplines. Examining the acoustic characteristics of infant cries, researchers can learn more about numerous aspects of newborn health and development, such as pain, hunger, and neurological abnormalities. As stated in [4], infants have a unique musical predisposition, where melody contour (i.e., F_0 and its dynamics) is most salient for them. This perception begins around the third trimester of pregnancy and hence, examining baby cries can reveal information about a child's emotional and social growth. Furthermore, researchers can learn more about infants' emotional expressions and caretaker's reactions by analysing their cries.

Early $1960s$ marks the exploratory work of infant cry analysis on normal cries using spectrograms. It was pioneered by *Xie et. al.* [5]. The attribute of cry signal, by and large defined as "cry phonemes" or "cry modes," were investigated in the time and frequency-domains. Cry modes explored were dysphonation, vibration, hyperphonation, and inhalation. This study was extended to the pathological cries in [6], where some of these cry modes were found to be correlated with pathology. Mel Frequency Cepstral Coefficients (MFCC), a state-of-the-art feature set, has recently been applied for cry classification tasks employing Gaussian Mixture Models (GMM) as classifiers [7], [8]. Recent studies on infant cry classification also involve detecting different types of pains, for example: belly pain, burping, discomfort, hungry, and tired using state-of-the-art features with traditional classifiers [9].

However, there are various challenges in analysing infant cry signals, using signal processing techniques, such as lack of sufficient number of pathological cry samples, extraction of excitation source, and vocal tract related features, and unbalanced data for classification. In the recent past, advance signal processing techniques and machine learning algorithms have been used for the classification and analysis of infant cries [10]. In order to address the above stated issues, transfer learning-based methods for infant cry classification tasks have been proposed in the literature [11, 12]. This paper presents a transfer learning strategy for multi-class and binary class infant cry classification using the pretrained **W**eb-scale **S**upervised **P**retraining for **S**peech **R**ecognition, also called Whisper [13]. The performance of whisper encoder based features are compared with baseline MFCC using Convolutional Neural Network (CNN), and Bidirectional Long-Short Term Memory (Bi-LSTM) Network as DNN classifiers. To the best of authors' knowledge, this is the first work of its kind reported on use of transfer learning-based approach using Whisper's Encoder Module for infant cry classification and analysis task.

2. Proposed Work

Whisper is an open source pre-trained sequence-to-sequence Transformer model similar to that described in [14]. It was designed for multilingual and multitask automated speech recog-

10.21437/Interspeech.2023-1916

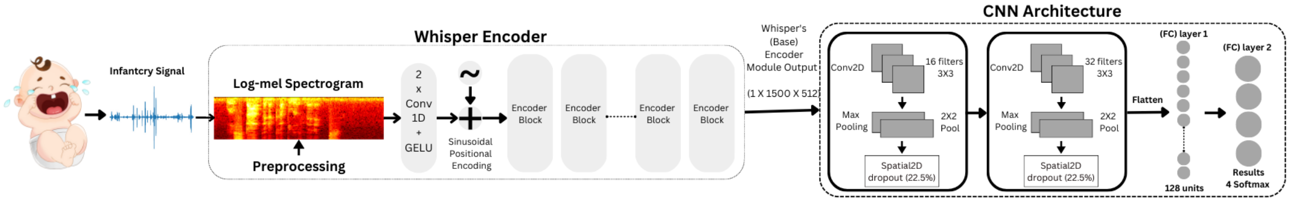

Figure 1: *Functional block diagram of proposed Whisper features in tandem with CNN*

nition (ASR) and was recently released in September 2022, at `https://github.com/openai/whisper`. Whisper is derived from the acronym **WSPSR**, which stands for **W**eb-scale **S**upervised **P**retraining for **S**peech **R**ecognition [13]. Whisper essentially emphasises the idea that training on a large and diverse supervised dataset and focusing on zero-shot transfer considerably enhances the system's endurance and performance.

Whisper model is trained by scaling weakly supervised audio paried with its' transcripts, scraped from the Internet, consisting of 680,000 hours of audio data from which 117,000 hours covers other languages, and 125,000 hours of the data is translation from other languages to English [13]. As a result, the Whisper model is trained on highly diversified audio data, encompassing a wide range of sounds from several different environments, recording setups, speakers, and languages. The huge volume and variety in audio samples certainly helps in training Whisper Models to generate high quality vector representations of audio signal whilst ensuring high robustness.

There are five different whisper models with increasing number of encoder-decoder blocks and number of trainable parameters, namely, tiny, base, small, medium, and large. The features generated by Whisper encoder module's are of fixed dimension which depends on the size of model. In this study, we have employed the Whisper's Base model with the dimension of output features at the end of encoder module fixed to $1 \times 1500 \times 512$. The size of the vectors obtained, grows in proportion to the size of the whisper model.

Transfer learning has been demonstrated to be successful in a variety of natural language processing and speech recognition tasks. This paper presents a transfer learning-based strategy for the classification of infant cries. We propose that the pre-trained Whisper model's sequence-to-sequence Transformer Encoder module effectively captures sequential/melodic structure embedded in the Mel Spectrogram (i.e, without DCT) of the cries which can be the discriminatory information of various classes. We picked this approach because the variety of the training dataset makes the model more robust and relevant to our situation. The suggested transfer learning approach makes use of the model's ability to extrapolate to previously unseen data. Furthermore, fine-tuning the pre-trained model for our particular task reduces training time whilst improving model performance.

2.1. Pipeline based on Transfer Learning approach

Transfer learning is a machine learning technique of training a model that leverages already learnt information from a related activity, to enhance learning for a new one. Transfer learning happens when an already trained model is retrained using a new dataset. While retraining the model, the information gained from the original dataset is preserved by freezing some trainable hyperparameters and neurons [15].

The pipeline employed for this study is shown in Figure.1. For the first step , the speech signal is preprocessed and prepared to be fed to the Whisper encoder module. For the preprocessing, the input infant cry signal is first resampled to 16*kHz* and then for maintaining uniformity the signal is padded to the time duration length of 30 seconds. Then, a 80-channel Log-Mel spectrogram is computed using window length of 25 *ms* and stride of 10 *ms*, whose coefficients are then normalised to values between $[-1, 1]$. These values are then passed through two convolution layers of kernel size 3, with GELU as activation function. Also, sinusoidal embeddings are used to assist the Whisper encoder in learning the relative places within the input as stated in [13]. The processed signal is then directed towards the Whisper encoder block, which produces a fixed dimensional vector as an output in its final hidden state. This output is then fed into a DNN classifier, which classifies the cry signal into respective classes.

For the training phase, Whisper's encoder module was kept frozen , while only the DNN classifier's weights were modified while back-propagating the errors. Moreover, in order to ensure that Whisper's encoder-based features are not biased against a particular DNN architecture, we conducted the experiments using Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory (Bi-LSTM) Network as classifiers [16, 17]. Furthermore, the number of classes can be modified by changing the number of units in the last dense layer of the DNN architecture shown as in Figure.1.

3. Experimental Setup

3.1. Datasets Used

In this work, we employ three datasets namely, Baby Chillanato dataset ***(D1)***, the property of Mexico's NIAOE-CONACYT [18, 19]. In-house DA-IICT dataset ***(D2)***, the property of [20,21], and combined $((D1) + (D2))$ dataset. The third dataset, referred as the combined dataset ***(D3)***, includes all cry utterances of Baby Chillanto dataset as well as the cry utterances of In-house DA-IICT dataset. The utterances of each dataset were resampled to a uniform sampling frequency of 16 *kHz* for fair experimentation. Table 1 shows the statistics of all datasets.

Table 1: *# Cry Utterances in all datasets Considered*

Class →	**Healthy**			**Pathology**			
Dataset ↓	**Normal**	**Hungry**	**Pain**	**Asphyxia**	**Deaf**	**Asthma**	**HIE**
D1	507	350	192	340	879	-	-
D2	793	-	-	-	-	215	182
D3	1842			1616			

3.2. Details of the Feature Set and Classifiers Used

In this study, the performance of Whisper Encoder features is compared with the state-of-the-art Mel Frequency Cepstral Coefficients *(MFCC)* features [22]. The baseline MFCC features were extracted from the audio files at a fixed sample rate of 16

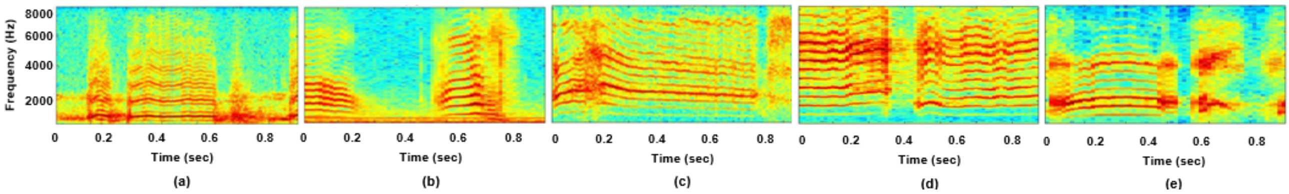

Figure 2: *Spectrographic Analysis for Various Pathologies in Infant's Cry. (a) Healthy cry, (b) Asphyxia, (c) Asthma, (d) Deaf, and (e) HIE cry. Best viewed in colour.*

kHz, with the *window length* of 512 samples and *window shift* of 256 samples. A total of 13 MFCC coefficients were extracted. Additionally, delta and double-delta features were extracted and appended with the original MFCC features, resulting in *39-D* cepstral features.

3.2.1. Convoluitonal Neural Network (CNN)

Convolutional Neural Network (CNN) work by mimicing how a human brain percieves an image and hence, it was employed as one of the classifier. A CNN model with 2 convolutional layers, each with kernal size of 3×3 was used [23]. Each convolutional layer is followed by max-pooling layer of size 2×2, along with Spatial 2D Dropout layers with dropout probabilities 0.225. At the end, two Fully-Connected (FC) layers were used with ReLU and Softmax activation functions respectively [24]. The Adaptive Moment Estimation *(Adam)* optimizer was used with learning rate of 0.001, with categorical-cross-entropy being employed as the loss function [23, 25]. The classifier was trained for 10 epochs for each fold.

3.2.2. Bidirectional Long Short Term Memory Network (Bi-LSTM)

Bidirectional Long Short-Term Memory (Bi-LSTM) belongs to the class of Recurrent Neural Network (RNN), which is broadly explored in sequencial modeling task, such as natural language processing and speech recognition. Bi-LSTM is an addendum to conventional LSTM architecture. Bi-LSTM performs forward and backward processing of the input sequence, enabling it to gather data from both previous and next time steps. This study uses 2 Bi-LSTM layers each consisting of 32 units with dropout probability of 0.1 at the end of each layer. Lastly, a dense layer of 4 units with Softmax activation function is used as the output layer.

3.3. Performance Evaluation Metrices

For this study, we have used variety of widely accepted metrics to evaluate our model's performance. Statistical metrics, such as the F1-score [26], Jaccard's index, which measures the similarity and dissimilarity of two classes [27], Mathew's Correlation Coefficient (MCC), which shows the degree of association between the expected and actual class [28], and Hamming loss, which is calculated on the basis of the number of samples that are incorrectly predicted [29], are used. The final scores for each metrics are calculated by averaging out the values across each fold.

4. Experimental Discussion

4.1. Spectrographic analysis

Figure 2 represents the spectrographic analysis of different pathologies. Figure 2(a) represents a normal Healthy cry, whereas Figures 2 (b), (c), (d), and (e) represents the Asphyxia, Asthma, Deaf, and HIE pathological cries respectively. HIE (Hypoxic-Ischemic Encephalopathy) is a brain injury caused by a lack of supply of oxygen to the brain, asphyxia is a condition in which the body is deprived of oxygen due to a lack of air or an obstruction of the airway, and asthma is a chronic respiratory disorder in which the airways become inflamed and restricted, making breathing difficult. While some symptoms may overlap, these illnesses have different underlying causes and therefore, different approaches to diagnosis and treatment. Hyperphonation is generally defined as energy distribution with high F_0. It can be observed that *hyperphonation* is found in every class except asphyxia. However, it is more prominent in HIE. HIE, occurs when brain lack blood and oxygen supply. This leads to high-pitched cry, or difficulty in crying and hence, supporting the proposed hypothesis. Further, it can also be observed that there is no sudden rise/fall/break in cry patterns (primarily due to F_0) for HIE, unlike other pathologies. It was further observed that dysphonation and inhalation cry modes were found only in asphyxia. Hence, we see similar breathing pattern as neonate struggles to inhale oxygen. Moreover, *transient downfall* over the *entire* frequency-axis was observed in asthmatic cry. From Figure 2 it is observed in all pathologies, that there is sudden or steep rise in the pitch source harmonics indicating efforts made by infants to cry.

4.2. Comparison with Existing Feature Sets

Experiments in this work were performed using *10*-fold Cross Validation *(CV)*. The reported accuracy is averaged over all folds.

4.2.1. Overall Performance for Binary Classifications

In this subsection, we present results obtained on, binary classification of healthy *vs.* pathological cries on three binary datasets considered. Table 2 reports the results obtained for all datasets for Whisper Encoder features and MFCC features. It is apparent form the table that whisper encoder features out performs baseline MFCC, more so, on both classifiers indicating no bias on classifier architecture. This might be due to the fact that, Whisper based features are known to capture sequential information from the signals. According to study reported in [4], infants have melodic structure in their cries. However, CNN is found to perform better than Bi-LSTM. This might be due to the effect of data partitioning in each fold.

Table 2: *Overall Performance (in % Accuracy) of Baselines and the Proposed Features on the Three Datasets*

Classifier ↓	**Dataset →**	**D1**	**D2**	**D3**
CNN	**MFCC**	95.72	88.31	91.24
	Whisper	**97.31**	**96.22**	**95.86**
Bi-LSTM	**MFCC**	97.17	95.88	96.79
	Whisper	**97.31**	**96.81**	**97.48**

4.2.2. *Multi-Class Classification*

As studied in section 4.2.1, it was observed that whisper based models outperformed MFCC Feature. Here we discuss the results obtained on multiple pathological detection. A new dataset ***(D4)*** was prepared, where each pathology from both ***(D1 & D2)*** datasets were considered for this study summing up to 4 classes in total. Figure 3 reports the % classification accuracy on CNN and Bi-LSTM classifiers. It can be observed form figure that proposed whisper based features outperforms MFCC features on both classifiers by increment of 1.67% on CNN and 1.05% on Bi-LSTM classifiers. Here, we observe significant increment in performance on both classifiers. As seen from the spectrographic analysis in section 4.1, various pathologies have various cry patterns/melodic structure and each having unique characteristics. This add to the proposition that whisper encoder features capture the sequential information better than MFCC based features. Additionally we can observe that the performance of Bi-LSTM classifier is comparable to that of CNN classifier.

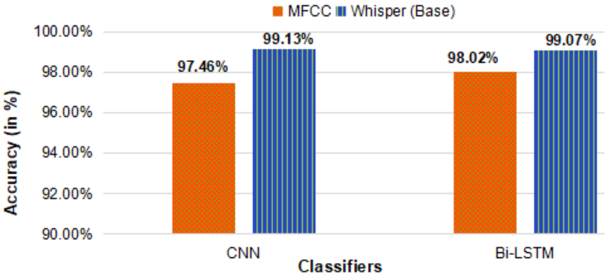

Figure 3: *Results for Multi-Class Classification*

4.3. Statistical Measures

Experiments were performed to assess the performance using statistical metrics, such as F1-score, Jaccard's index, MCC score, and Hamming loss as reported in Table 3. The results for all set of experiments are averaged over *10*-folds on the CNN classifier. On the whole, it is clearly observed that the proposed Whisper Encoder features clearly outperforms baseline MFCC features for all the metrics.

Table 3: *Performance Measures for Classification Experiments on Various Datasets using CNN*

Datasets ↓	Feature Set	F1 Score	MCC	Jaccard Index	Hamming Loss
D1	**Whisper**	**0.973**	**0.946**	**0.948**	**0.0269**
	MFCC	0.957	0.916	0.918	0.0428
D2	**Whisper**	**0.962**	**0.919**	**0.928**	**0.0378**
	MFCC	0.881	0.730	0.793	0.1168
D3	**Whisper**	**0.959**	**0.919**	**0.921**	**0.0414**
	MFCC	0.912	0.824	0.839	0.0876
D4	**Whisper**	**0.991**	**0.986**	**0.983**	**0.0087**
	MFCC	0.975	0.960	0.952	0.0253

4.3.1. *Analysis of Precision through Retraining*

Figure 4 illustrates the test accuracy variation over *10*-folds, where on each fold the model was trained for 10 epochs. It was performed to check robustness of classifier to the quoted precision for proposed whisper encoder for 5 trials. The results obtained were consistent across all trials indicating the robust behaviour for multiclass pathology classification task. However, trivial variance in accuracy, depicted in figure can be attributed to the randomness inherent in the DNN and seed values.

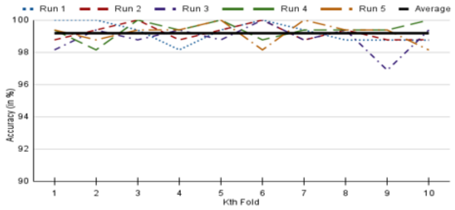

Figure 4: *Analysis of Precision for Retraining on CNN.*

4.3.2. *Feature Space Visualization Using t-SNE Plots*

Multi-class pathology classification capability is also validated using scatter plots obtained using t-SNE plots. Here Whisper and MFCC features were projected to 2-D space for various pathological class. Figure 5(a) and Figure 5(b) presents the scatter plots for whisper Encoder-based features and MFCC features, respectively. It is observed from figure 5 that the inter-class distance between the clusters of different classes is greater for Whisper encoder features. Also, Whisper model can clearly distinguish between two closely co-related classes, namely, Asphyxia and HIE. However, MFCC features fail to differentiate between them.

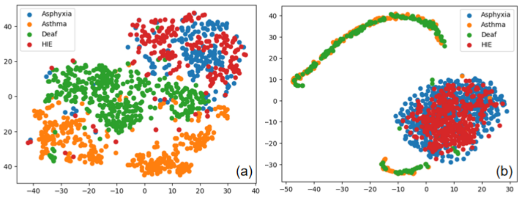

Figure 5: *Scatter Plots Obtained using t-SNE Plot for 4 Pathologies (a) Whisper encoder features, and (b) MFCC Features. Best viewed in colour.*

5. Conclusions

This work investigated significance of novel whisper encoder features based on transfer learning framework for socially relevant research problem on infant cry classification. This study aimed at both for pathological detection as well as classification of normal *vs.* pathological cries. Proposed methodology was found to outperform baseline MFCC for various evaluation scenarios, such as use of two different datasets, along with combined dataset, and different classifier structures. This might be due to sequential capturing characteristic of whisper model as infants are known to have melodic structure in cries. Further discrimination capability of proposed features were analysed using t-SNE plots, and statistical measures, such as F1-score, MCC, Jaccard index and Hamming loss. Furthermore, analysis of precision for retraining the model was also checked to ensure the precision of proposed whisper encoder features on pathology detection. However, some the limitations of this work would be to check the effect of different whisper models along with effect of data augmentation and comparison with other encoder based features such as TERA, and other acoustical features such as GeMAPS, eGeMAPS, and ComParE. Our future work will be aimed towards investigating the relevance whisper model for analysis and classification of different voiced pathologies and augmented cries.

6. Acknowledgements

The authors sincerely thank MeitY, Govt. of India, for the project 'Speech Technologies in Indian Languages' BHASHINI', (Grant ID: 11(1)2022-HCC (TDIL)).

7. References

[1] C. C. Onu, I. Udeogu, E. Ndiomu, U. Kengni, D. Precup, G. M. Sant'Anna, E. Alikor, and P. Opara, "Ubenwa: Cry-based diagnosis of birth asphyxia," 31^{st} *Conference on Neural Information Processing Systems (NIPS), Long Beach, CA*, 2017, https://www.ubenwa.ai/research.html {Last Accessed: 30/11/2022}.

[2] C. C. Onu, J. Lebensold, W. L. Hamilton, and D. Precup, "Neural transfer learning for cry-based diagnosis of perinatal asphyxia," *International Conference on Learning Representations (ICLR) Workshop, Graz, Austria*, 2019.

[3] K. Manickam and H. Li, "Complexity analysis of normal and deaf infant cry acoustic waves," in 4^{th} *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2005.

[4] L. Armbrüster, W. Mende, G. Gelbrich, P. Wermke, R. Götz, and K. Wermke, "Musical intervals in infants' spontaneous crying over the first 4 months of life," *Folia Phoniatrica et Logopaedica*, vol. 73, no. 5, pp. 401–412, 2021.

[5] Q. Xie, R. K. Ward, and C. A. Laszlo, "Determining normal infants' level-of-distress from cry sounds," in *Proceedings of Canadian Conference on Electrical and Computer Engineering*. IEEE, 1993, pp. 1094–1096.

[6] H. A. Patil, "Cry baby": Using spectrographic analysis to assess neonatal health status from an infant's cry," in *A. Neustien (Ed.) Advances in Speech Recognition, Springer*, 2010, pp. 323–348.

[7] H. F. Alaie, L. Abou-Abbas, and C. Tadj, "Cry-based infant pathology classification using gmms," *Speech Communication*, vol. 77, pp. 28–52, 2016.

[8] 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, pp. 1–17, 2021.

[9] K. Rezaee, H. Ghayoumi Zadeh, L. Qi, H. Rabiee, and M. R. Khosravi, "Can you understand why i am crying? a decision-making system for classifying infants' cry languages based on deepsvm model," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

[10] H.-N. Ting, Y.-M. Choo, and A. A. Kamar, "Classification of asphyxia infant cry using hybrid speech features and deep learning models," *Expert Systems with Applications*, vol. 208, p. 118064, 2022.

[11] G. Anjali, S. Sanjeev, A. Mounika, G. Suhas, G. P. Reddy, and Y. Kshiraja, "Infant cry classification using transfer learning," in *TENCON 2022*. IEEE, Seoul, South Korea, 2022, pp. 1–7.

[12] H.-t. Xu, J. Zhang, and L.-r. Dai, "Differential time-frequency log-mel spectrogram features for vision transformer based infant cry recognition," *Proc. INTERSPEECH, Incheon Songdo ConvensiA, Korea*, pp. 1963–1967, 2022.

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022, {Last Accessed: March 6, 2023}.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in NIPS, Long Beach California, United States of America*, vol. 30, 2017.

[15] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[16] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015, {Last Accessed: February 25, 2023}.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[18] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García, "Validation of the cry unit as primary element for cry analysis using an evolutionary-neural approach," in *2008 Mexican International Conference on Computer Science*. IEEE, 2008, pp. 261–267.

[19] A. Rosales-Pérez, C. A. Reyes-García, J. A. Gonzalez, O. F. Reyes-Galaviz, H. J. Escalante, and S. Orlandi, "Classifying infant cry patterns by the genetic selection of a fuzzy model," *Biomedical Signal Processing and Control,*, vol. 17, pp. 38–46, 2015.

[20] N. Buddha and H. A. Patil, "Corpora for analysis of infant cry," *Oriental Cocosda, Vietnam*, 2007.

[21] 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.

[22] O. M. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," in *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction, Norwich, United Kingdom*, Norwich, United Kingdom, 30-31 August, 2004.

[23] S. Bock and M. Weiß, "A proof of local convergence for the adam optimizer," in *2019 International Joint Conference On Neural Networks (IJCNN)*, 2019, pp. 1–8.

[24] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018, {Last Accessed: Feb 6, 2023}. [Online]. Available: http://arxiv.org/abs/1803.08375

[25] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in NIPS*, vol. 31, 2018, Montreal Canada.

[26] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[27] M. Bouchard, A.-L. Jousselme, and P.-E. Doré, "A proof for the positive definiteness of the Jaccard index matrix," *International Journal of Approximate Reasoning*, vol. 54, no. 5, pp. 615–626, 2013.

[28] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[29] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "Regret analysis for performance metrics in multi-label classification: the case of Hamming and subset zero-one loss," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 280–295.