Interspeech 2024
1-5 September 2024, Kos, Greece

Towards Classifying Mother Tongue from Infant Cries - Findings Substantiating Prenatal Learning Theory

Tim Polzehl[1,2], *Tim Herzig*[1], *Friedrich Wicke*[2], *Kathleen Wermke*[3], *Razieh Khamsehashari*[2], *Michiko Dahlem*[3], *Sebastian Möller*[1,2]

[1]Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI)
[2]Quality and Usability Lab, Technische Universität Berlin, [3]University Hospital Würzburg

{tim.polzehl,tim.herzig}@dfki.de, WermkeK@ukw.de,
{f.wicke,razieh.khamsehashari,sebastian.moeller}@tu-berlin.de, dr.michiko.dahlem@gmx.de

Abstract

In this work we introduce automatic mother tongue classification based on infant cries. We use data of 63 German and Japanese healthy, term-born neonates, and model their cries with the help of data augmentation and Pre-trained Audio Neural Networks (PANNs), leveraging transfer learning methods suited to the very limited data at hand. Applying small models on top of PANNs we obtain F1 scores of 85% and above on a held-out test set. We conduct several experiments analyzing model reliability, all of which indicate the network focuses on the actual infant cries rather than on confounding factors. We visualize the network focus to adhere to pitch contour and harmonics thereof, rendering prosody central for our model prediction. Eventually, our models add a novel objectively obtained perspective to neonate crying analysis, while our results substantiate an extremely early vocal learning indication.

Index Terms: infant cry, language acquisition, PANN, augmentation, LOSO

1. Introduction

The human infant is dependent upon close social contact with its mother from birth. Crying, i.e. acoustic signaling, is the most powerful signaling available to the neonate. The basic acoustic signature of neonatal cries is primarily determined by processes that characterize the sound source (vocal fold oscillation) as the vocal tract is still immature [1]. Neonate populations worldwide do not differ in terms of the anatomical structures and functional mechanisms of phonation and the essential acoustic attributes of neonatal cries are similar everywhere [2]. It might sometimes seem difficult to imagine that neonates' natural crying is laying the foundations for later language. [3] demonstrated that neural memory traces are formed by auditory learning prior to birth and that auditory experiences during the fetal period have a remarkable influence on the neonates' auditory pitch discrimination accuracy. [4] show that infants can differentiate between languages in the womb a month before they are born, i.e. prenatal understanding starts before birth. As a consequence, the fetal auditory system becomes especially attuned to the dominant salient input of the intra-uterine acoustic environment. Fetal learning includes input characteristics that seem to affect output characteristics. For example, it has been found that prenatal exposure to both, non-tonal and tonal languages seems to shape neonates' cry melody. Having had ample opportunity to become familiar with their mother's speech prosody in the womb, neonates have been found to exhibit salient pitch-based elements in their own cry melodies [5]. While most pre-speech studies have focused on infants during babbling or even later stages, there is increasing evidence that auditory learning commences already very early in utero.

Crying is further believed to be a way of processing or expressing a wide range of emotions, and there is research suggesting that infants use crying as a means of communication [6]. However, a serious methodological limitation of most previous studies is the used cry data basis. For example, to reliably differentiate between "sick" and "healthy" infants requires consideration of a variety of individual developmental and medical information which is generally not provided with the cry signals. Classifications of assumed cry causes, like hunger or discomfort, are likewise fuzzy as infants crying is a result of an emotional state and may not always express a single discrete cause. Verifying the ground truth data labels seems almost impossible. Further, acoustic features of infants' natural crying dramatically change with age. Very often, cry signals were pooled for experiments that comprise age intervals of several months or even a year. Importantly, in our work, these limitations were avoided by analyzing natural crying of healthy, term-born newborns aged 2-6 days from monolingual mothers, having access to objective ground truth with respect to the mothers' language. It is the first work to compare tone-accent and intonation languages and, to the best of the authors' knowledge, remains unprecedented in this respect.

In terms of cry representation modeling, recent advances in self-supervised learning (SSL) have further substantiated the quality of speech representations for different purposes. Exploring SSL capacity and answering the need to cope with the very limited amount of cry examples existing, we apply pretrained neural networks, i.e. networks pre-trained on large-scale datasets that have shown to generalized well to several other tasks in computer vision, natural language processing, and audio pattern analysis. Specifically, Pre-trained Audio Neural Networks [7] have successfully been applied to several related tasks such as audio tagging, acoustic scene classification, music classification, speech emotion classification, and sound event detection. Recently, generalization power of PANNs have been investigated with respect to "non-semantic" tasks, e.g. paralinguistic tasks like speaker recognition, language recognition, medical diagnosis, voice emotion recognition, etc [8]. Specifically, the "TRILLsson" model family has been released, showing superior performance on the *Non-Semantic Speech Benchmark* (NOSS) [9]. In our work we apply TRILLsson models as feature extractor and further adapt its output through a small number of neuronal layers, in order to map the pre-trained features to our task. The dataset that we use comprises infant cry recordings of German and Japanese babies, which are finally the two classes of our downstream task. Eventually, the objective of this study is to propose the introduction of a new element into the discourse surrounding the influence of the maternal language prosody in infant cry analysis. For the first time, our work provides objective evidence that features of a tone-accent lan-

10.21437/Interspeech.2024-1000

guage are reflected in infant crying. We open-source our code[1], and hoping to stimulate discussion.

2. Literature Review

There is a broad body of work on modeling and analyzing infant cries. Works often analyze acoustic features, prosodic features, or both and are applied to pathological cry identification [10], identification of the cry reasons, detection of the presence of cries and other applications, cf. [11] for a detailed overview. Chaiwachiragompol et al. [12] use a discrete wavelet transform for feature extraction, followed by a single layer neural network in order to classify the infants' cries according to what they want to communicate, like hunger, sleepiness and discomfort. Cries are classified by an *extreme learning machine (ELM)* neural network using hidden layers of 10, 20 and 30 nodes. The classification type of infants sound are "*Eh*", "*Eairh*", "*Heh*", "*Neh*" and "*Owh*", reaching up to 98% accuracy in type extraction.

A multichannel CNN architecture is used for infant cry detection by Severini et al. [13], and a 1-D CNN is able to outperform SVMs at the same task [14]. There is also work on CNNs with transfer learning applied to infant cry classification [15]. The authors use existing pre-trained convolutional neural network ResNet50 in order to classify between different types of baby cries from the Baby Chillanto Database [16]. Using a 5-fold cross-validation their models achieve accuracies of more than 90%. Recurrent Neural Networks are also often used, for example in the form of Long-Short-Term-Memory Neural Networks [17] or combined with CNNs [18]. Classifying (i) neutral from positive mood sounds, (ii) fussing sounds, and (iii) crying sounds [17] reached 68.7% unweighted average recall in the Interspeech 2018 Cry Challenge [19]. In [18], the authors study the problem of classifying five different types of emotion or needs expressed by infant cry, namely *hunger, sleepiness, discomfort, stomachache,* and indications that the infant wants to *burp*. Applying 5-folds cross-validation their CNN-RNN model reached accuracy up to 94.97%. Very recently, [20] achieved 98.34% applying classification of five types of crying perception using DeepSVMs. [21] explores diarization of family members including babies/toddlers (aged 3-24 months), parents and siblings and further classifies babies/toddlers vocalizations as *cry*, *fuzz* and *babble*. [22] builds upon the Whisper encoder [23] to classify infants cries (from newborn to 9 months of age) for needs and healthy classes like, *healthy-normal*, *healthy-hungry* and *healthy-pain* along 4 discrete pathology classes. Scoring F1 of up to 99.1% the authors conclude that a whisper-based classification outperforms a MFCC-based classification, when using CNN and Bi-LSTM on top of the bases. Finally, Manfredi et al. [24] uses 12 different melodic shapes with an commercial software tool and reports up to 95% differences among French, Arabic and Italian mother-tongue healthy full term newborns using Random Forest and 4 neuro-fuzzy classifiers. The shapes examined and their observed frequency of occurrence remain open to interpretation linguistically, as, e.g. the typical French "rising pattern" [5] occurs more frequently in the Italian cries in the authors work.

3. Experiment Setup

3.1. Data

We used the University Hospital Würzburg Infant Cry database, which is a collection cleaned from cry vocalizations that exhibited marked phonatory noise phenomena (e.g., creaky sounds) and/ or phenomena like sudden pitch shifts (register changes) or marked vibrato-like phenomena [25]. It consists of 1711 cries (1173 Japanese, 538 German) from 63 infants (31 Japanese, 32 German), aged 2 to 6 days. On average, the data comprises 27.15 samples per infant, with maximum of 95 and minimum of 6 samples per infant overall and a duration of 1.794 seconds on average. The samples were recorded using portable recording devices with a distance of $10-15$ cm from the infants' head. The German samples were recorded in hospitals in Berlin and Würzburg, the Japanese samples were recorded in a hospital in Hiroshima. All subjects were healthy, full-term neonates with normal hearing from a strictly monolingual family background.

3.2. Data Augmentation

To counter unwanted effects and learning from short-cut cues in the recordings not originating from the infant cries we further apply augmentation techniques that have successfully been used in many domains such as image recognition, object classification, and speech recognition.

Basic augmentation: In order to account for duration differences we stretch the signal by a random factor between 0.8 and 1.2. Next, zero-padding is added to both ends of the signal, using a random factor between 0.0 and 0.1 on both ends. In order to additionally realize an increase of data for training and balancing of class distributions, we create several samples out of each of the original signals by applying these basic augmentation methods. Eventually, we add to the original dataset (*'org.'*) another augmented set with 3000 samples per language (*'aug.'*) and compare results on the two sets.

MixUp in between classes: We use the *MixUp* regularization technique [26] empirically exploring the configuration of α, the upper bound of the mixing weights. It controls the degree of contrast between the inputs to be mixed. We draw uniformly distributed weights with α in range $[0.0, 0.4]$ for the mixed-in cross-class example, keeping it the background of the (foreground) sample at hand. Similar to the application in [27] and [28] we utilize *MixUp* to create background sound perturbation by randomly picking a cross-class sample from the dataset. This way, the models are forced not to concentrate on the respective mixed-in background characteristics, which is expected to mainly correspond to acoustic properties such as the microphone frequency characteristics, room/reverberation patterns originating from the environment, and else background noise present in different recording rooms.

Signal-filtering for room and microphone acoustics: To further validate the independence towards room- and microphone acoustics, we conduct preliminary experiments preparing our dataset like they were recorded in a set of 10 different rooms and/or using a total of 10 different microphone characteristics. For each sample in our dataset, we randomly impose room and/or microphone characteristics applying [29] in order to prevent the model from focusing on these characteristics. We selected 10 popular contemporary microphones offering diverse frequency response patterns, and room characteristics perceptively exceeding the recording scenario conditions in the hospitals by room size and reverb.

3.3. Evaluation

We create different *train* and *validation* sets in addition to a held-out test set unseen to the training procedure. We divide the sets in a speaker-independent way, making sure that different recordings from individual infants do not spread across the train, validation and test set borders. The global *test* set comprises 4 Japanese and 4 German infants completely unseen

[1] https://github.com/timherzig/infant_cry

throughout training. Augmentation are applied on train and validation sets. We use the F1-metric for evaluation, i.e. harmonic mean of precision and recall, primarily on basis of a *80%:20%* split of *train:validation* sets respectively.

LOSO: In order to obtain another additional perspective on the reliability of the performance for individual infants, we add a leave-one-speaker-out (LOSO) evaluation, with one infant per language left out from the training set placed into the validation set per loop. By nature, this evaluation method often shows lower but more robust estimates of performance, as we align a size-maximized training set towards a single infant per language in training. The generalization of these models is then again tested on the global test set. We finally report on the average performance of LOSO models along individually maximum and minimum scores to show the range of performance.

3.4. Model architecture

With the TRILLsson model release, the authors follow a knowledge distillation approach, where models are distilled on public data only, ensuring further public release. The authors train small student models on several fixed-length input architectures, including ResNets, EfficientNets, and Transformers, and match them to arbitrary-length input of their teacher CAP12 Transformer embeddings [30]. CAP12 is a 606M parameter (2.2GB) Conformer model trained with a modified Wav2Vec2.0 self-supervised loss on a 900M+ hour speech dataset derived from YouTube [31]. The authors use 58K hours of publicly available speech data from Libri-light [32] and AudioSet [33] for distillation. For the present work we use their TRILLsson-1 AST [34] student model, yielding best results for several downstream tasks potentially related to infant cry classification including detection of speaker, language, command, synthetic speech, dysarthria, and emotion. Being the smallest of the distilled models, size results in 729kB and 5.5M parameters respectively. Eventually, the distillation uses only 7% of data used in training of CAP12 and achieves between 90% and 96% of the larger CAP12 accuracy on 6 of 7 NOSS benchmark tasks. We adapt the 1024-dimensional output of TRILLsson-1 for experimentation using different configurations before the classification head: (1) drop-out layer with drop-out ratio [0.0, 0.2, 0.4]; (2) BiLSTM layer of size 64; (3) 2 x fully connected feed forward layer of size [1024, 512]. We train with up to 100 epochs applying early stopping on Nvidia RTXA6000, training batch size 8 samples, with a reduced learning rate on plateaus.

4. Results

Table 1 shows results, which are to be interpreted against a random guessing baseline. Best models achieve up to 89.9% F1 on the held out test set, which well demonstrates the capability to discriminate between the two mother tongue classes from the cries. In more detail, when looking at dropout, we observe that lower or zero dropout rates seem to be beneficial, indicating no additional generalization effect as well as no overfitting to the train set data for the present experiments. When looking at MixUp, preferably paired with zero dropout, the method does not decrease the performance drastically. In other words, even when applying strong mixing of backgrounds, e.g. to a degree of 40% which is a point close to transitioning background into foreground sounds, F1 results stay reasonable, i.e. above 83.5%, with a small decrease of 6.4% F1 absolute. Introducing MixUp-regularization thus indicates that our model focuses primarily on the infants' cries, not on respective mixed-in confounding factors from background characteristics.

Set	MixUp	Drp	$F1(\%)\uparrow$	LOSO $F1(\%)\uparrow$	
				mean	max/min
org.	0.0	0.0	**89.9**	**85.5**	**92.0 / 47.7**
		0.2	86.3	80.9	89.2 / 46.6
		0.4	83.5	74.8	87.5 / 47.7
	0.2	0.0	**86.3**	83.0	92.0 / 47.7
		0.2	85.2	77.5	87.5 / 47.7
		0.4	82.3	68.5	86.8 / 47.7
	0.4	0.0	**88.6**	81.6	90.9 / 47.7
		0.2	81.2	74.8	88.1 / 47.7
		0.4	73.8	64.7	81.3 / 47.7
aug.	0.0	0.0	**85.8**	**84.5**	**87.5 / 77.8**
		0.2	84.1	81.4	85.8 / 65.9
		0.4	84.7	77.1	84.1 / 58.5
	0.2	0.0	**84.1**	82.0	87.5 / 60.2
		0.2	84.1	77.1	85.2 / 47.7
		0.4	64.8	72.5	86.4 / 47.7
	0.4	0.0	**83.5**	80.0	86.9 / 47.7
		0.2	84.7	76.6	86.9 / 47.7
		0.4	72.7	70.7	84.1 / 47.7

Table 1: *Results of infant mother language classification for different settings of MixUp and dropout ('drp') for the original ('org') and augmented ('aug') datasets. Mean F1 on test set a) when trained on fixed train and validation splits; b) when applying LOSO evaluation allowing for max and min insights.*

Table 1 also shows the LOSO evaluation results. Again, zero dropout seems beneficial, and the decrease of mean F1 scores caused by MixUp results in a maximum of 5.5% F1 absolute when retaining zero dropout. As expected, the mean overall F1 scores obtained with LOSO evaluation are generally lower than in the experiment above, however, we now also see a smoothing effect when looking a the max and min scores for the augmentation dataset. While for the top-performing model of 85.5% F1 we observe a huge range between 92.0% and 47.7% of F1 when training the models with individual infants in the validation set splits, we can reduce the risk of obtaining very low results by applying basic augmentation. Scarifying 1% of overall best F1 performance on base of the LOSO performance method (corresponding to 4.1% F1 loss of fixed train-validation set performance method) we can safeguard worst results to climb up from 47.7% F1 to 77.8% F1.

In another experiment further iterating through views on the reliability of our obtained scores we apply signal-based filtering to superimpose explicit room and/or microphone characteristics to the infant recordings. The general observations follow the results reported above. Best recognition scores were obtained with zero dropout, while scaling MixUp from 0% to 20% and 40% results in F1 scores of 88.1%, 81.3%, and 85.2% respectively. Again, scaling the background noise MixUp does not decrease the performance drastically, fostering previous indications that our models learn predominantly from the infant cries rather than from background characteristics.

Eventually analyzing which input parts the model focuses on we apply Gradient-weighted Class Activation Mapping (Grad-CAM) [35]. Inserting a Conv2D layer with 1 filter of kernel size (3x3) for GradCam visualization after the BiLSTM layer GradCam utilizes gradients from a target concept flowing into the final convolutional layer to produce a coarse localiza-

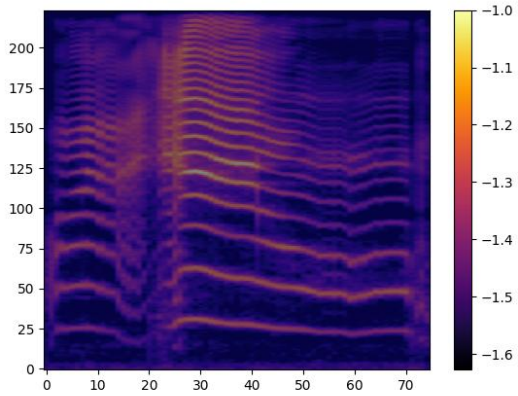

Figure 1: *GradCam visualization of a cry with lighter color corresponding to more model focus*

tion map highlighting the important regions in the input that contribute to predicting the concept. By employing Grad-CAM visualization as exemplified in Figure 1, we find further indication that the model focus is on pitch and harmonics thereof, indicating that the model learns from expected pitch values, the pitch contour, and pitch-related voice quality characteristics. An analysis towards more specific insights, i.e. which language shows which patterns or characteristics of pitch and harmonics potentially including their relative distribution and energy patterns cannot be concluded from this level of visualization. However, for the present experiments, random noise or ill-conditioned recordings can be excluded from the list of predominant potential shortcut features based on the plot.

Overall, the obtained classification scores ranging from above 80% to the highest of 89.9% indicate that the models were able to principally differentiate the mother tongue from the cries. However, the absolute best scores are not of highest importance in this work. More importantly, the figures demonstrate, that the models have well begun to learn to differentiate between the mother tongues based on the infant cries, which in turn substantiates a neonatal language pattern acquisition in utero for the given languages and data sets.

5. Discussion and Outlook

Our infant cry classification experiments provide indicative results to foster and demonstrate the shaping effect of the surrounding language on neonates' crying. From a clinical perspective, the finding is very useful as it corroborates the functional maturity of the neonatal larynx and phonatory system elements responsible for crying.

As a major limitation factor, these results demonstrate the found effect on the pair of Japanese and German language infant cries only. Specifically, Japanese is regarded as mainly tone-accent language, which is characterized by stronger melodic variations of the speech melody compared to German. While there is general proof of distinctiveness of prosody and melody in between languages [36], varying degrees of inter- and intra-similarities in between languages may strongly limit generalization of our findings for other languages.

More thorough analyses as well as the comparison with other infant-cry datasets would contribute to the discussion. However, as comparable data (especially with respect to infant age) is hardly available, the obtained results should be interpreted as an indication rather than a proof. First, the presented results show that noise patterns like background noise or recording device characteristics do not play an important role for the models. Second, Grad-CAM visualizations shown the model focuses on the cries pitch. Eventually, other unknown factors may still be intertwined, e.g. dialects in the mother language.

Future work will focus on a more systematic exploration of model parameter space, augmentation, and model architecture beyond simple layers. In current experiments, the basic augmentations applied seem to help reduce the risks of incurring very low results with individual infants. Follow-up work will analyze the distribution of performances across individual infants in more detail. In addition, augmentations are also expected to help improve mean overall recognition scores, which could not be observed with the current scope of basic augmentations. Future work will therefore focus on adding methods to widen augmentation in order to further facilitate the learning process. Further, foundation or general purpose models like BYOL-A [28], WavLM [37] or Whisper [23] are interesting for future work, as well as examining the impact of infant age. Finally, the complexity of the underlying neurophysiological mechanisms as well as the high speed at which the vocal control mechanisms act, make neonatal cry properties a suitable indicator for the identification of neurophysiological dysfunctions [38]. We further intend to target neurophysiological dysfunctions, along with inclusion and comparison with other languages, if data will be made available.

6. Conclusions

In this work, we successfully applying deep transfer learning for the task of classifying the mother tongue from infant cries. Thus, our work substantiates the theory of neonatal language acquisition from a new objective angle, by classifying experiments on Japanese and German infant cries. We focus our analysis on reliability of the proposed method rather than task-specific advanced modeling methods. Still, our best models reach overall reasonable performances of $> 85\%$ F1 using PANNs and a few adaptation layers. Further experimentation shows that major assumable confounding factors related to recording conditions do not significantly influence the classification success. Further, a supportive visualization of model focus indicates the expected importance of pitch as a prosodic descriptor for the model performance. We therefore argue that our models have already begun to learn how to differentiate infants' mother language from the cries, opening a discussion about objectively analyzing prenatal learning from postnatal observation. We carefully discuss limitations and interpretation, and outline follow-up work in our discussion.

7. References

[1] R. D. Kent and H. K. Vorperian, "Development of the craniofacial-oral-laryngeal anatomy." Singular, 1995.

[2] E. H. Buder, L. B. Chorna, D. K. Oller, and R. B. Robinson, "Vibratory Regime Classification of Infant Phonation," *Journal of voice : official journal of the Voice Foundation*, vol. 22, no. 5, pp. 553–564, Sep. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2575878/

[3] E. Partanen, T. Kujala, M. Tervaniemi, and M. Huotilainen, "Prenatal music exposure induces long-term neural effects," *PloS one*, vol. 8, no. 10, 2013.

[4] U. Minai, K. Gustafson, R. Fiorentino, A. Jongman, and J. Sereno, "Fetal rhythm-based language discrimination: A biomagnetometry study," *Neuroreport*, vol. 28, no. 10, p. 561, 2017.

[5] B. Mampe, A. D. Friederici, A. Christophe, and K. Wermke, "Newborns' Cry Melody Is Shaped by Their Native Language," *Current Biology*, vol. 19, no. 23, pp. 1994–1997, Dec. 2009.

[6] M. J. Corwin, B. M. Lester, and H. L. Golub, "The infant cry: What can it tell us?" *Current Problems in Pediatrics*, vol. 26, no. 9, pp. 313–334, Oct. 1996. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0045938096800120

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020, iEEE.

[8] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Interspeech 2020*, Oct. 2020, pp. 140–144.

[9] J. Shor and S. Venugopalan, "TRILLsson: Distilled Universal Paralinguistic Speech Representations," in *Interspeech*. ISCA, Sep. 2022, pp. 356–360.

[10] S. Möller and R. Schönweiler, "Analysis of infant cries for the early detection of hearing impairment," *Speech communication*, vol. 28, no. 3, pp. 175–193, 1999.

[11] 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.

[12] A. Chaiwachiragompol and N. Suwannata, "The features extraction of infants cries by using discrete wavelet transform techniques," *Procedia Computer Science*, vol. 86, pp. 285–288, 2016.

[13] M. Severini, D. Ferretti, E. Principi, and S. Squartini, "Automatic detection of cry sounds in neonatal intensive care units by using deep learning and acoustic scene simulation," *IEEE Access*, pp. 51 982–51 993, 2019.

[14] K. Manikanta, K. Soman, and M. S. Manikandan, "Deep learning based effective baby crying recognition method under indoor background sound environments," in *IC on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*. IEEE, 2019, pp. 1–6.

[15] 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 *IC on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*. IEEE, 2019, pp. 106–110.

[16] M. Moharir, M. Sachin, R. Nagaraj, M. Samiksha, and S. Rao, "Identification of asphyxia in newborns using gpu for deep learning," in *2017 2nd International Conference for Convergence in Technology (I2CT)*. IEEE, 2017, pp. 236–239.

[17] M. Huckvale, "Neural network architecture that combines temporal and summative features for infant cry classification in the interspeech 2018 computational paralinguistics challenge," in *Interspeech*. ISCA, 2018, pp. 137–141.

[18] T. N. Maghfira, T. Basaruddin, and A. Krisnadhi, "Infant cry classification using cnn–rnn," in *Journal of Physics: Conference Series*. IOP Publishing, 2020.

[19] B. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats," in *Interspeech*. ISCA, Sep. 2018, pp. 122–126.

[20] 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 Transaction on Asian and Low-Resource Language Information Processing*, Jan. 2023.

[21] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Towards robust family-infant audio analysis based on unsupervised pretraining of wav2vec 2.0 on large-scale unlabeled family audio," *arXiv preprint arXiv:2305.12530*, 2023.

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

[23] 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.

[24] C. Manfredi, R. Viellevoye, S. Orlandi, S. Orlandi, A. A. Torres-García, G. Pieraccini, and C. A. Reyes-García, "Automated analysis of newborn cry: relationships between melodic shapes and native language," *Biomed. Signal Process. Control.*, vol. 53, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:181762846

[25] K. Wermke, J. Teiser, E. Yovsi, P. J. Kohlenberg, P. Wermke, M. Robb, H. Keller, and B. Lamm, "Fundamental frequency variation within neonatal crying: Does ambient language matter?" *Speech, Language and Hearing*, vol. 19, no. 4, pp. 211–217, 2016.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[27] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[28] D. Niizumi, D. Takeuchi, N. Harada, Y. Ohishi, and K. Kashino, "BYOL for Audio: Exploring Pre-Trained General-Purpose Audio Representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 137–151, 2022.

[29] A. Ratnarajah, S.-X. Zhang, M. Yu, Z. Tang, D. Manocha, and D. Yu, "Fast-rir: Fast neural diffuse room impulse response generator," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 571–575.

[30] A. Jansen, D. S. Park, J. Shor, W. Han, and Y. Zhang, "Universal Paralinguistic Speech Representations Using Self-Supervised Conformers," *ICASSP*, 2022.

[31] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. C. Sim, B. Ramabhadran, T. N. Sainath, F. Beaufays, Z. Chen, Q. V. Le, C.-C. Chiu, R. Pang, and Y. Wu, "BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, Oct. 2022.

[32] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fügen, T. Likhomanenko, G. Synnaeve, A. Joulin, M. I. Abdelrahman, and E. Dupoux, "LIBRI-LIGHT: a benchmark for asr with limited or no supervision." IEEE, May 2020, p. 7669.

[33] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017, pp. 776–780.

[34] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," *Interspeech 2021*, pp. 571–575, Aug. 2021, conference Name: Interspeech 2021.

[35] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," *IEEE Winter C. on Applications of Computer Vision (WACV)*, pp. 839–847, 2018.

[36] S. E. Trehub, "Musical Predispositions in Infancy," *Annals of the New York Academy of Sciences*, vol. 930, no. 1, pp. 1–16, Jun. 2001.

[37] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[38] L. L. LaGasse, A. R. Neal, and B. M. Lester, "Assessment of infant cry: Acoustic cry analysis and parental perception: Assessment of Infant Cry," *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 11, no. 1, pp. 83–93, Feb. 2005.