



Towards Robust Family-Infant Audio Analysis Based on Unsupervised Pretraining of Wav2vec 2.0 on Large-Scale Unlabeled Family Audio

Jialu Li^{1,2}, Mark Hasegawa-Johnson^{1,2}, Nancy L. McElwain^{2,3}

¹Department of Electrical and Computer Engineering, University of Illinois

²Beckman Institute for Advanced Science and Technology, University of Illinois

³Department of Human Development and Family Studies, University of Illinois

jialuli3@illinois.edu, jhasegaw@illinois.edu, mcelwn@illinois.edu

Abstract

To perform automatic family audio analysis, past studies have collected recordings using phone, video, or audio-only recording devices like LENA, investigated supervised learning methods, and used or fine-tuned general-purpose embeddings learned from large pretrained models. In this study, we advance the audio component of a new infant wearable multi-modal device called LittleBeats (LB) by learning family audio representation via wav2vec 2.0 (W2V2) pretraining. We show given a limited number of labeled LB home recordings, W2V2 pretrained using 1k-hour of unlabeled home recordings outperforms oracle W2V2 pretrained on 52k-hour unlabeled audio in terms of parent/infant speaker diarization (SD) and vocalization classifications (VC) at home. Extra relevant external unlabeled and labeled data further benefit W2V2 pretraining and fine-tuning. With SpecAug and environmental speech corruptions, we obtain 12% relative gain on SD and moderate boost on VC. Code and model weights are available.

Index Terms: family-infant audio analysis, speaker diarization, vocalization classification, unsupervised learning, wav2vec 2.0

1. Introduction

In the U.S., 1 in 6 children aged 2–8 years has a diagnosed mental, behavioral, or developmental disorder [1], but such disorders are often neglected. Child mental health problems begin in early childhood, and daily interactions with family members that are repeated and reinforced over time are critical to children’s emotional well-being. According to attachment theory: the primary caregivers who respond quickly and consistently to an infant’s needs allow the child to develop a sense of security. Children may develop insecure attachment styles if parents are often unavailable, intrusive, or respond inconsistently to the child’s cues, particularly signs of distress [2]. Additionally, previous psychological studies indicate that parents and infants are more likely to show coordinated physiological activities when their vocal and physical behaviors are mutually responsive during play [3], and that such mutually responsive behaviors help maintain or increase shared positive interactions and emotions [4]. Therefore, to better support child mental health outcomes, it is essential to detect if parent and infant establish coordinated behaviors at an early stage in daily activities at home. Although previous work emphasizes the importance of the mother-infant dyad, father-infant [5] and sibling-infant [6] interactions also play a crucial role in infant development. Thus, we consider the larger family context in this work to better understand infant emotional and behavioral development during the first years of life. In the past, to analyze family interactions,

This work was supported by funding from NIDA (R34DA050256), NIMH (R21MH112578), NIFA (ILLU-793-368), NSF (1725729), and Beckman Graduate Fellowship.

researchers or parents have recorded family audio at home or laboratory using a cell phone, video camera, or an audio-only recording device like the Language Environment Analysis device (LENA) [7]. In this study, we test a new infant wearable multi-modal device we developed called LittleBeats™ (LB), and we aim to advance the LB audio pipeline such that it automatically provides reliable labels of SD and VC for family members, including infants, parents, and siblings, at home.

For automatic family audio analysis, previous studies used supervised machine learning models on adult-child diarization [8, 9], detecting infant cry [10, 11], and classifying parent/infant vocalizations [12] with a limited amount of labeled vocalization data because audio annotation is a time-consuming and labor-intensive task. To tackle challenges of data sparsity, past studies explored transfer learning techniques on infant cry detection by incorporating external relevant labeled datasets for additional training [13, 14], computing node similarities in graphical convolutional networks using both labeled and unlabeled datasets [15], and fine-tuning [16] or leveraging embeddings [17] from models pretrained on large-scale image or audio datasets. Recently, a self-supervised learning model, W2V2 [18], which uses unsupervised pretraining from ~52k-hours of unlabeled audio, excels at multiple downstream speech processing tasks, including speech-to-text [19], speech emotion recognition [20], and speaker diarization [21]. Fine-tuning pretrained W2V2 general-purpose audio representations has shown robust performance in some audio-related healthcare applications with resource-poor data, such as detecting stuttering [22] and classifying COVID-19 vocalization [23], but in other applications, such as autism detection in child speech [24] and pathological speech recognition [25], W2V2 did not always outperform supervised learning models potentially due to the domain mismatch between pretraining and fine-tuning. Limited research has assessed pretraining effects of W2V2 to learn task-specific audio representation; one example [26] showed task-adaptive pretraining on speech emotion recognition to bridge the gap between W2V2 pretraining and a target domain by continuing pretraining on a target dataset after initial pretraining.

In this paper, we explore benefits of learning family audio representation in W2V2 pretraining using large-scale day-long home recordings collected from LB and LENA. To the best of our knowledge, this is the first study that examines the pretraining effects of a self-supervised model on a family audio analysis task. We demonstrate that given a limited amount of labeled LB home recordings, W2V2 pretrained using 1100 hours of unlabeled LB home recordings outperforms oracle W2V2 pretrained using 52k-hour audio in the fine-tuned downstream tasks of SD and VC. We also discover that pretraining and fine-tuning steps can further benefit from an additional 3200 hours of LENA unlabeled home recordings and from relevant labeled

Table 1: Overview of data distribution prepared after preprocessing steps. Last six columns are Total duration (including silence), Vocalized duration (all participants), and vocalized duration broken down into (CRY/FUS/BAB) for key child (CHN), (CDS/ADS/LAU/SNG) for adult female (FAN) and male (MAN), and total non-key child/sibling (CXN) in hours (h). In:in-domain/Out:out-of-domain. m:month/y:year.

Device	Type	Domain	Context	Age	# of families	Total dur	Vocal dur	CHN	FAN	MAN	CXN
LB	Labeled	In	Home	< 14m	22	10.61h	4.78h	.18/.48/1.1	1.0/.68/.04/.11	.24/.34/.01/.03	.57
LENA	Labeled	Out	Home	< 24m	30	14.59h	9.05h	.58/.75/.84	1.11/1.97/.08/.50	1.24/.82/.05/.09	1.02
LB	Labeled	Out	Virtual visit	< 10m	11	1.35h	0.8h	.02/.14/.14	.38/.03/.01/.08	-	-
Camera	Labeled	Out	Lab visit	< 14m	105	9.94h	3.4h	.24/.69/.23	1.75/.01/.1/.38	-	-
LB	Unlabeled	In	Home	< 5y	110	1100h	-	-	-	-	-
LENA	Unlabeled	Out	Home	< 5y	275	3200h	-	-	-	-	-

LENA home and laboratory recordings, respectively. This work is well-aligned with the Interspeech 2023 theme of inclusive speech technology. The current standard method for diagnosing behavioral disorders is observation in the clinic. The need for clinical observation can make it difficult for working parents, and for parents in rural settings, to obtain an accurate diagnosis. At-home diagnosis supported by LB technology would better serve poor and rural populations, and thus increase the inclusiveness of healthcare.

2. Data

For this study, we recruited families with study flyers distributed at multiple local community organizations and online family forums. All study procedures were approved by the Institutional Review Board at the University of Illinois at Urbana-Champaign (UIUC). To encourage participants to permit the recording of their private family life, our consent form specifies that data will not be shared outside of our research team. Our consent form further specifies that most of the recordings will be processed automatically without human intervention (the unlabeled data), and that human coders will only listen to small samples of the data (the labeled data). Many families noted that they were willing to participate only because the vast majority of recordings are processed automatically without human auditing. We aim to improve performance on LB home recordings, so we consider LB home recordings as in-domain data and other relevant data as out-of-domain data. Table 1 summarizes the distributions of data used in this experiment after preprocessing steps (see Section 2.3).

2.1. Unlabeled data

We collected a large amount of unlabeled day-long home recordings from families with children under 5 years of age. Child participants wore either the LB or LENA device at home during the day for two or three days. Families who participated in LB and LENA home recordings do not overlap.

2.2. Labeled data

To manually annotate in-domain LB home recordings, we separated each daylong recording into 10-min segments. As continuous manual annotation of the audio recordings is time- and labor-intensive, human coders only annotated a few 10-min segments for each family, selected based on the highest active vocalization rates computed by a statistical voice activity detector (VAD). Human coders manually labeled key child (CHN), female adult (FAN), male adult (MAN), and other child/sibling (CXN) vocalizations using Praat [27], with cross-coder validation at a precision of 0.2s. Ten percent of selected segments were double-coded, and inter-coder reliability (Cohen’s kappa score) was 0.89 for CHN, 0.86 for FAN, 0.81 for MAN, and 0.80 for CXN. Child vocalizations were manually labeled as *cry* (CRY), *fuss* (FUS), and *babble* (BAB); adult vocaliza-

tions (MAN and FAN) were manually labeled as *child-directed speech* (CDS), *adult-directed speech* (ADS), *laugh* (LAU), and *singing/rhythmic speech* (SNG). The Cohen’s kappas were 0.76 for CHN, 0.87 for FAN, and 0.71 for MAN. In total, we obtained 70 labeled 10-mins segments from 22 families.

We also labeled out-of-domain datasets from two studies of infant (3-12 months) and toddler (18-24 months) development. We followed similar data collection and annotation protocol for out-of-domain datasets. One of the studies included annotations of LENA home recordings. Another study recorded mother and infant completing two semi-structured interactive tasks in the laboratory (Lab visit) or at home using LB during the COVID-19 pandemic (Virtual visit). The kappas for out-of-domain data are similar to in-domain data. Families who participated in in- and out-of-domain studies do not overlap.

2.3. Data preprocessing

LB audio was sampled at 15832Hz or 22756Hz at two hardware versions. Camera and LENA audio were sampled at 48k Hz and 16k Hz respectively. To make audio data compatible with W2V2 training, we resampled LB and camera audio to 16kHz using librosa (v0.9.2) [28]. To prepare unlabeled data for pre-training, we first applied VAD to remove silent portions for all unlabeled home recordings, and then divided non-silent audio into 10s segments. To prepare labeled data for fine-tuning, we labeled the audio stream in intervals of 2s starting every 0.2s. The label of each 2s interval was determined by the temporal majority of human annotations on the centered 1s interval (timestamps 0.5-1.5s), if two or more vocalization labels were present. If only one vocalization label was present, its label was applied to the whole interval if its duration was 0.2s or greater, as we observed that children tend to make short vocalizations. To reduce SD errors, intervals labeled with more than one speaker were discarded. To obtain better quality out-of-domain labeled data, non-silent examples with energy below the minimum energy of in-domain CHN vocalizations were discarded. In this way, we intentionally select vocalizations that are close to the CHN and ignore those that have lower energy or can be easily confused with background noise. Note that we didn’t apply energy thresholding for in-domain labeled data in order to avoid altering its data distribution.

For in-domain labeled data partition used in fine-tuning and testing, we followed a leave-one-family-out scheme to ensure training, development and testing sets did not have overlapped families. We divided 70 recordings into four groups based on infant age ranges (1.1-4m, 4-9m, 9-13m, and 13-14m). We randomly selected recordings from one family per age group as the testing set, a small number of recordings as the development set, and the rest of the recordings form the training set. In total, we have 52/6/12 recordings from 15/3/4 families for training/development/testing sets respectively.

2.4. Data augmentation

Data augmentation has proven beneficial for several speech processing tasks. In this study, we used 5 data augmentation techniques implemented using Speechbrain (v0.5.7) [29], as described in [30], including specAugment [31], random chunks of audio dropping, speed perturbation by resampling audio with slightly different sampling rates, reverberation (convolve speech with a room impulse response), noise (add noise to speech with random signal-to-noise ratio ranging from 0-15 dB), and reverberation+noise. We used a room impulse response dataset [32] for reverberation corruption. We explored two noise datasets, MUSAN [33] and CHiME-Home (CH) [34]. MUSAN contains 6-hour noises from a variety of sources, such as thunder, paper rustling, fax machine noises, animal noises, etc. CH contains 6.8-hour audio of 4s clips from domestic environment, including child and adult vocalizations, TV and kitchen noises, etc. We also attempted to mine noises from LB and LENA home audio but obtained little improvement, perhaps because W2V2 already learned the distribution of LB/LENA domestic noise.

3. Experimental Setup

W2V2 first encodes audio waveforms into latent feature embeddings using several convolutional layers. These latent embeddings are then fed to both a quantizer and a transformer network. The quantizer assigns each latent embedding to a specific learned speech unit from an inventory of quantized units. Some of the latent representations are masked before they are fed into the transformer network. During training, a contrastive loss is applied to transformer outputs and used to predict target audio segment from a context of surrounding audio segments. Detailed pretraining procedures are described in [18]. Two versions of W2V2, base (encoded feature size 768 with 12 transformer layers) and large (encoded feature size 1024 with 24 transformer layers), are available, and W2V2 base model is used in our study.

Figure 1 shows the overall model architecture for fine-tuning W2V2. Three 2-layer feed-forward networks (FFNs) are used as output tiers, including a SD tier and two VC tiers, CHN and ADU (FAN and MAN). The SD tier learns to detect speaker as silent or one of CHN, FAN, MAN, or CXN; if not silent, the corresponding vocalization tier learns to classify the vocalization type. We compare W2V2 features extracted from the last transformer layer vs. all 12 transformer layers. The former applies mean pooling (MP) over time dimension before feeding to FFNs, while the latter applies MP over output $f_{i,t}$ of each transformer layer i for every time step t , then a weighted average (WA) layer learns a weight α_i for each output f_i , as described in Equation 1:

$$f_{out} = \frac{\sum_{i=1}^{12} \alpha_i ((\sum_{t=1}^T f_{i,t})/T)}{\sum_{i=1}^{12} \alpha_i} \quad (1)$$

We also test applying WA layer first followed by MP but it yields inferior performance. The loss objective for fine-tuning W2V2 is the average of cross-entropy loss over three output FFNs. If the training dataset includes out-of-domain data, one of two methods is used to train the network to make any necessary distinction between the processing of in- vs. out-of-domain data: concatenating one-hot learnable embeddings to W2V2 features, or multi-task learning in which a fourth FFN learns a binary domain label. For the former, we test concatenating one-hot embeddings before or after WA layer and find adding domain embeddings before WA layer gives better results. For the latter, the loss objective is defined as

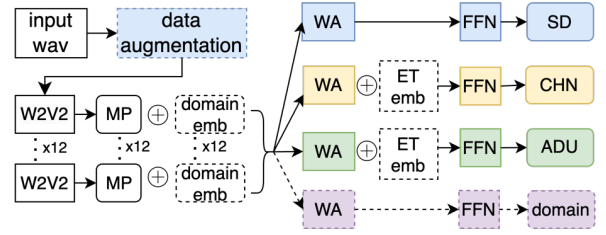


Figure 1: Overview of model architecture. W2V2=wav2vec 2.0, MP=mean pooling, emb=embedding, WA=weighted average, ET=ECAPA-TDNN speaker identity vector, FFN=feed-forward network, SD=speaker diarization, CHN=child vocalization classifier, ADU=adult vocalization classifier, domain $\in \{in, out\}$ is a boolean marker for out-of-domain data augmentation via embeddings or multi-task learning. Dashed outline indicates optional modules in some experimental tests.

$$L = \alpha_1 L_{SD} + \alpha_2 L_{CHN} + \alpha_3 L_{ADU} + \alpha_4 L_{domain} \quad (2)$$

where $\alpha_1 = \alpha_2 = \alpha_3 = 0.33$ and $\alpha_4 = 0$ if domain FFN is not present, otherwise $\alpha_1 = \alpha_2 = \alpha_3 = 0.32$ and $\alpha_4 = 0.03$. To further improve overall performance, we explore introducing additional ECAPA-TDNN (ET) [35] speaker embeddings and data augmentation techniques as described in Section 2.4. As ET has shown great performance on adult speaker diarization/recognition tasks, we concatenate ET speaker embeddings pretrained from our labeled data with W2V2 features to provide extra speaker information to potentially improve our model.

We pretrain W2V2 on fairseq (v0.12.2) [36] using UIUC HAL cluster [37] with 4 NVIDIA V100 GPUs for 3 weeks until convergence for each experiment. We adapt the pretraining recipe of W2V2 base model on 960h Librispeech data with minor changes on minimum (1s) and maximum (10s) audio lengths to save computational memory. We implement fine-tuned model and ET using SpeechBrain. Each fine-tuning experiment is trained 10 epochs with batch size 32 on single NVIDIA GTX 1080 Ti for about 10/40 hours without/with out-of-domain data respectively. With data augmentation, the total training time roughly increases 5 times when there are 5 types of augmentation. We evaluate our model using unweighted F1-scores for each tier over all classes. The epoch with the best average score over three tiers on in-domain development set is used for final evaluation on in-domain testing set. Adam optimizer with learning rates of output FFNs and W2V2 model starting from $1e-4$ and $1e-5$ respectively is used; scheduler with new-bob technique is used to anneal learning rates based on development set performance after each epoch. ET speaker embedding size is 192. FFN hidden node size is set as 384 (half of W2V2 feature dimension), and one-hot learnable domain embedding size is set as 256. Between the two FFN layers, 1D batch normalization, Leaky Relu activation, and dropout with 0.1 probability are applied sequentially. In total, W2V2 and 3 FFN have 95M and 8M learnable parameters respectively. Our code and model weights are available.

4. Results & Discussions

4.1. Comparisons across different unsupervised pretrained models fine-tuned on in-domain labeled data

To compare unsupervised pretraining effects, we test 4 unlabeled datasets: *base* (oracle version): $\sim 52k$ -hour audio

https://huggingface.co/lijialudew/wav2vec_LittleBeats_LENA

Table 2: *F1*-scores (%) for SD, CHN, ADU, and the average of three tiers among W2V2 models pretrained using different unlabeled datasets and fine-tuned on in-domain labeled LB data only, FT: Fine-tune W2V2 and 3 output FFNs/FR: Freeze W2V2 and fine-tune 3 output FFNs

model	setting	features	SD	CHN	ADU	Avg
base	FT	all layers	62.8	54.9	59.2	59.0
Libri960h	FT	all layers	62.5	46.5	53.5	54.2
LB1100h	FR	layer 12	45.5	45.3	29.6	40.1
LB1100h	FR	all layers	67.9	66.2	56.4	63.5
LB1100h	FT	layer 12	70.8	66.6	60.0	65.8
LB1100h	FT	all layers	68.2	66.3	63.0	65.8
LL4300h	FT	layer 12	74.9	67.4	60.3	67.5
LL4300h	FT	all layers	71.5	68.9	66.9	69.1

Table 3: *F1*-scores (%) trained on both in- and out-of-domain data comparing systems w/o domain tagging, w/one-hot domain embedding, or w/multi-task learning of a domain classifier

features	domain tagging	SD	CHN	ADU	Avg
layer 12	-	68.7	68.1	67.6	68.1
all layers	-	68.9	69.6	70.6	69.7
layer 12	one-hot	70.6	70.4	71.9	71.0
all layers	one-hot	69.6	70.6	68.7	69.6
layer 12	multi-task	66.1	68.3	69.7	68.0
all layers	multi-task	68.4	68.6	67.3	68.1

and music, *Libri960h*: oracle version followed by fine-tuning with 960h Librispeech, *LB1100h*: 1100h LB home audio, and *LL4300h*: 1100h LB+3200h LENA home audio. Each pre-trained models was fine-tuned with in-domain labeled LB data. Table 2 presents the results. W2V2 models pretrained on unlabeled home recordings outperformed original oracle models with/without fine-tuning on Librispeech for all three tiers, and *LL4300h* has the overall best performance. Thus, we use *LL4300h* for the rest of the experiments unless noted otherwise. The results suggest W2V2 effectively learns family audio representation on large-scale unlabeled home recordings across several age groups of infants and toddlers under 5 years old at initial pretraining stage. For fine-tuning, we compare fine-tuning W2V2 model vs. freezing W2V2 and fine-tuning 3 output FFNs only. We find the former significantly boosts the overall performance compared with the latter. We suspect fine-tuning the entire W2V2 makes W2V2 robust to noisy home audio, which often contains clothing rustling, TV noises, toy banging, etc. For W2V2 pretrained on *LB1100h*, we obtain comparable performances by using features from either last transformer layer (layer 12) or all 12 layers. For W2V2 pretrained on *LL4300h*, features extracted from all layers benefit VC tiers while features extracted from the last layer help the SD tier. Empirically, features of the top layer encode speaker information and features of lower layers encode vocalization type.

4.2. Effects of adding out-of-domain labeled data

Table 3 shows the results of fine-tuning both in- and out-of-domain labeled data using binary domain learning techniques. Compared with fine-tuning on in-domain data only (see Table 2 last row), we observe that adding out-of-domain data helps improve VC tiers (mostly ADU) but slightly hurts SD tier (see Table 3), perhaps because out-of-domain microphones cause domain shift in the distribution of speaker diarization cues. For binary domain learning, one-hot domain embeddings provide marginal benefits while multi-task learning isn't helpful.

Table 4: *F1*-scores (%) with and w/o ET speaker embeddings

data	model	features	ET emb	SD	CHN	ADU	Avg
In	LB1100h	layer 12	-	70.8	66.6	60.0	65.8
In	LB1100h	layer 12	In	73.2	68.4	62.9	68.2
In+Out	LL4300h	all layers	-	68.9	69.6	70.6	69.7
In+Out	LL4300h	all layers	In+Out	70.2	69.6	70.8	70.2

Table 5: *F1*-scores (%) using in- and out-of-domain data augmented by SpecAug, reverberation from RIR corpus, additive noise from M:MUSAN/C:CH corpora, and reverberation+noise; features are weighted average of all layers

noise	domain tagging	ET emb	SD	CHN	ADU	Avg
-	-	-	68.9	69.6	70.6	69.7
C(SD)	-	-	76.5	69.2	67.7	71.1
M(SD)	-	-	77.2	70.8	71.5	73.2
M(VC)	-	-	75.2	71.8	67.8	71.6
M(All)	-	-	76.2	72.0	68.7	72.3
M(SD)	one-hot	-	77.4	68.5	64.3	70.1
M(SD)	one-hot	In+Out	78.2	70.1	68.8	72.4

4.3. Effects of introducing ECAPA-TDNN speaker embeddings and data augmentation

We pretrain ET using in- only or in- and out-of-domain labeled data, and we achieve unweighted F1-scores of 66.6% and 67.6% on labeled LB testing data on SD task respectively. We attempt to concatenate ET embeddings to all three tiers for training but obtain large degradation, which may due to the incompatibility of W2V2 features and ET embeddings learning speaker information simultaneously. Thus, we only concatenate ET embeddings on VC tiers. Table 4 presents results of ET embedding. We find that ET embeddings are helpful when W2V2 is trained on a limited number of family recordings, such as fine-tuning on *LB1100h* using in-domain labeled data (relative gain 3.3% for SD, 2.7% for CHN, and 4.8% for ADU). To save computational time, we use features from the last transformer layer in *LB1100h*, which performs similarly to using all 12 layers. If W2V2 is trained on a relatively larger number of family recordings, such as fine-tuning on *LL4300h* using all labeled data, ET embeddings provide limited benefits.

We apply 5 types of data augmentation (see Section 2.4), on SD tier only, VC tiers only, or all tiers. Table 5 summarizes relevant results. We discover that applying data augmentation on SD tier only using MUSAN for additive noise achieves the best performance overall, which shows relative improvement of 12% on SD tier, 1.7% on CHN tier, and 1.3% on ADU tier. We observe data augmentation with CH domestic noises degrades VC performances. Probably corrupting home audio with CH child/adult vocalizations injects undesirable background speaker acoustics to target speaker vocalizations. When we combine data augmentation with binary domain learning and ET embeddings, we achieved the optimal performance on SD tier with 13.5% relative improvement.

5. Conclusions & Future Work

This study shows the effectiveness of family audio pretraining and data augmentation to reduce domain mismatch for family audio analysis. Pretraining ET speaker embeddings are useful when a limited number of family recordings are used for training. Data augmentation largely helps SD and moderately benefits VC. In the future, we aim to collect home recordings from more families and explore active learning for quickly adapting current models to new families with minimal labeling effort.

6. References

- [1] R. H. Bitsko, J. R. Holbrook, L. R. Robinson, J. W. Kaminski, R. Ghandour, C. Smith, and G. Peacock, "Health care, family, and community factors associated with mental, behavioral, and developmental disorders in early childhood—united states, 2011–2012," *Morbidity and Mortality Weekly Report*, vol. 65, no. 9, pp. 221–226, 2016.
- [2] N. L. McElwain and C. Booth-LaForce, "Maternal sensitivity to infant distress and nondistress as predictors of infant-mother attachment security," *Journal of Family Psychology*, vol. 20, no. 2, p. 247, 2006.
- [3] Y. Hu, N. L. McElwain, and D. Berry, "Mother-child mutually responsive orientation and real-time physiological coordination," *Developmental psychobiology*, vol. 63, no. 7, p. e22200, 2021.
- [4] G. Kochanska, D. R. Forman, N. Aksan, and S. B. Dunbar, "Pathways to conscience: early mother-child mutually responsive orientation and children's moral emotion, conduct, and cognition," *Journal of Child Psychology and Psychiatry*, vol. 46, no. 1, pp. 19–34, 2005.
- [5] V. Sethna, E. Perry, J. Domoney, J. Iles, L. Psychogiou, N. E. Rowbotham, A. Stein, L. Murray, and P. G. Ramchandani, "Father-child interactions at 3 months and 24 months: Contributions to children's cognitive development at 24 months," *Infant mental health journal*, vol. 38, no. 3, pp. 378–390, 2017.
- [6] W. Oh, B. L. Volling, and R. Gonzalez, "Trajectories of children's social interactions with their infant sibling in the first year: A multidimensional approach," *Journal of Family Psychology*, vol. 29, no. 1, p. 119, 2015.
- [7] X. D., J. A. Richards, and J. Gilkerson, "Automated analysis of child phonetic production using naturalistic recordings," *Journal of Speech, Language & Hearing Research*, vol. 57, no. 5, pp. 1638–1650, 2014.
- [8] J. Xie, L. P. Garcia-Perera, D. Povey, and S. Khudanpur, "Multi-lda diarization on children's speech," in *Interspeech*, 2019, pp. 376–380.
- [9] A. Cristia, S. Ganesh, M. Casillas, and S. Ganapathy, "Talker diarization in the wild: The case of child-centered daylong audio-recordings," in *Interspeech 2018*, 2018, pp. 2583–2587.
- [10] J. Xie, X. Long, R. A. Otte, and C. Shan, "Convolutional neural networks for audio-based continuous infant cry monitoring at home," *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27 710–27 717, 2021.
- [11] T. Jian, Y. Peng, W. Peng, and Z. Yang, "Research on lstm+ attention model of infant cry classification," *Journal of Robotics, Networking and Artificial Life*, 2021.
- [12] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," in *Interspeech*, 2020.
- [13] A. Gujral, K. Feng, G. Mandhyan, N. Snehil, and T. Chaspari, "Leveraging transfer learning techniques for classifying infant vocalizations," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.
- [14] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations," *Speech Communication*, vol. 133, pp. 41–61, 2021.
- [15] J. Chunyan, M. Chen, L. Bin, and Y. Pan, "Infant cry classification with graph convolutional networks," in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, 2021, pp. 322–327.
- [16] H. Xu, J. Zhang, and L. Dai, "Differential time-frequency log-mel spectrogram features for vision transformer based infant cry recognition," *Proc. Interspeech 2022*, pp. 1963–1967, 2022.
- [17] X. Yao, M. Micheletti, M. Johnson, E. Thomaz, and K. de Barbaro, "Infant crying detection in real-world environments," in *ICASSP 2022*, 2022, pp. 131–135.
- [18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on NeurIPS Systems*, 2020, pp. 12 449–12 460.
- [19] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech 2019*, 2019.
- [20] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, pp. 3400–3404, 2021.
- [21] X. Zheng, C. Zhang, and P. Woodland, "Tandem multitask training of speaker diarisation and speech recognition for meeting transcription," *Proc. Interspeech 2022*, pp. 3844–3848 – 3848, 2022.
- [22] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Introducing ecapa-tdnn and wav2vec2. 0 embeddings to stuttering detection," *arXiv preprint arXiv:2204.01564*, 2022.
- [23] A. Mallol-Ragolta, S. Liu, and B. Schuller, "Covid-19 detection exploiting self-supervised learning representations of respiratory sounds," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2022, pp. 1–4.
- [24] N. A. Chi, P. Washington, A. Kline, A. Husic, C. Hou, C. He, K. Dunlap, and D. P. Wall, "Classifying autism from crowd-sourced semistructured speech recordings: Machine learning model comparison study," *JMIR Pediatrics and Parenting*, 2022.
- [25] L. P. Violeta, W.-C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," *Interspeech 2022*, 2022.
- [26] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition," *arXiv preprint arXiv:2110.06309*, 2021.
- [27] P. Boersma, "Praat: doing phonetics by computer," <http://www.praat.org/>, 2006.
- [28] 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 *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.
- [29] M. Ravanelli, T. Parcollet, P. Plantinga, and et al, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.
- [30] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "Ecapa-tdnn embeddings for speaker diarization," *arXiv preprint arXiv:2104.01466*, 2021.
- [31] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019.
- [32] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *Proceedings of ICASSP 2017*, pp. 5220–5224, 2017.
- [33] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [34] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.
- [35] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Proc. Interspeech 2020*, 2020.
- [36] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.
- [37] V. Kindratenko, D. Mu, Y. Zhan, J. Maloney, S. H. Hashemi, B. Rabe, K. Xu, R. Campbell, J. Peng, and W. Gropp, "Hal: Computer system for scalable deep learning," in *Practice and Experience in Advanced Research Computing*, 2020, pp. 41–48.