INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Robust Self Supervised Speech Embeddings for Child-Adult Classification in Interactions involving Children with Autism

Rimita Lahiri[1], *Tiantian Feng*[1], *Rajat Hebbar*[1], *Catherine Lord*[2], *So Hyun Kim*[3], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, USA
[2]Semel Institute of Neuroscience and Human Behavior, University of California, USA
[3]School of Psychology, Korea University, Korea

rlahiri@usc.edu

Abstract

We address the problem of detecting who spoke when in child-inclusive spoken interactions i.e., automatic child-adult speaker classification. Interactions involving children are richly heterogeneous due to developmental differences. The presence of neurodiversity e.g., due to Autism, contributes additional variability. We investigate the impact of additional pre-training with more unlabelled child speech on the child-adult classification performance. We pre-train our model with child-inclusive interactions, following two recent self-supervision algorithms, Wav2vec 2.0 and WavLM, with a contrastive loss objective. We report $9-13\%$ relative improvement over the state-of-the-art baseline with regards to classification F1 scores on two clinical interaction datasets involving children with Autism. We also analyze the impact of pre-training under different conditions by evaluating our model on interactions involving different subgroups of children based on various demographic factors.

Index Terms: speech, child-adult classification, self-supervision, autism

1. Introduction

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder, characterized by deficits in social and communicative abilities along with restrictive repetitive behavior [1, 2]. Individuals with ASD tend to show symptoms of anomalies in language, non-verbal comprehension, expressions and vocal prosody patterns [3, 4]. In the United States, the prevalence of ASD in children has steadily increased from 1 in 150 [5] in 2002 to 1 in 44 in 2022. It is critical to develop early ASD diagnosis to create timely interventions. One of the most common observation tools supporting ASD diagnostic and intervention efforts includes clinically-administered semi-structured dyadic interactions between the child and a trained clinician [6, 7]. Computational analysis of such interactions provides evidence-driven opportunities for the support of behavioral stratification as well as diagnosis and personalized treatment.

However, with regards to behavioral feature extraction and analysis for these dyadic interactions, prior works have primarily relied on human-annotated data segmentation by speaker labels, which is expensive and time-consuming to obtain, especially for large corpora. Computational modeling of naturalistic conversations has gained a lot of attention in the past few decades because of its potential in rich human behavioral phenotyping. Hence, it is desirable to conduct automatic analysis of these interactions using signal processing and machine learning. Specifically, one fundamental module for supporting automated processing of child-adult interactions is the task of child-adult speech classification i.e., distinguishing the speech regions of the child from those of an interacting adult. Analysis of child speech is more challenging than adult speech because of the wide variability and idiosyncrasies associated with child [8, 9, 10]. An additional layer of complexity arises while analyzing speech for the clinical domain, as different clinical conditions may lead to unique patterns in language and speech, making it challenging for current computational approaches to capture.

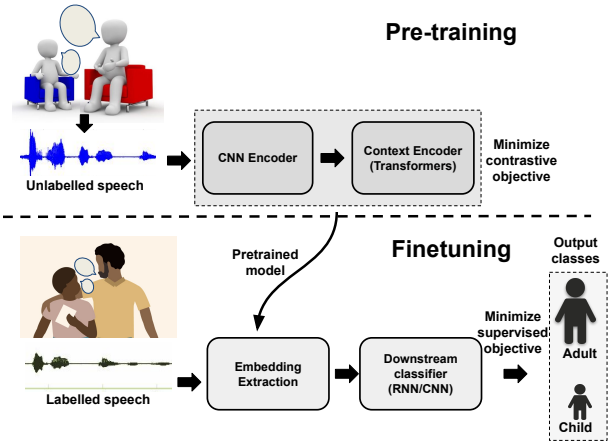


Figure 1: *Schematic overview of the proposed two-step recipe for child-adult classification*

Training a robust child-adult classifier is challenging for two main factors: the scarcity of reliably labeled datasets containing child speech and the larger within-class variability due to the changes in child speech based on demographic factors like age, gender, and developmental status including any clinical symptom severity [11]. Most recent works addressing the problem of speaker diarization have primarily targeted fine-tuning the pre-trained models by optimizing a supervised objective. So far, *Self-Supervised Learning* (SSL) algorithms are largely under-explored for leveraging unlabelled child speech for developing speaker discriminative embeddings, especially in real-world settings such as clinical diagnostic and monitoring sessions. Specifically, there is a limited understanding of how the performance of these models varies across children with different demographics, including age and gender.

Contributions of this paper: We address the above questions by evaluating the impact of including more child speech, during pre-training on the child-adult speaker classification. We choose Wav2vec 2.0 (W2V2)-base and WavLM-base+ as the backbone models. The detailed contributions of this work are summarized as:

- Our work represents one of the first attempts to **leverage unlabelled child speech in pre-training** for developing speaker discriminative embeddings, especially due to the vast inherent heterogeneity in the data arising from developmental differences.

10.21437/Interspeech.2023-1447

- We experimentally substantiate the effectiveness of our method for downstream child-adult speaker classification tasks using W2V2 and WavLM and **report over 13% and 9% relative improvement over the base models in terms of F1 scores** in two datasets, respectively.
- We also illustrate and analyze the performance of the proposed method among different subgroups of children based on demographic factors.

2. Background

2.1. Self-supervision in speech

The need for building speech processing frameworks in low/limited resource scenarios has spurred significant efforts on unsupervised, semi-supervised and weakly supervised learning strategies to reduce reliance on labeled datasets. The success of SSL [12] in natural language processing, notably due to its generalizability and transferability, has also inspired its adoption within the speech domain. Early studies explored SSL in speech with generative loss [13, 14], while more recent ones have focused on discriminative loss [15, 16] and multi-task learning objectives [17, 18]. The current approach in this realm follows a two-step process: first pre-train a model in a self-supervised manner on large amounts of unlabeled data to encode general-purpose knowledge, and next specialize the model on various downstream tasks through fine-tuning. Past studies have reported the efficacy of SSL algorithms by leveraging the pre-trained embeddings on downstream tasks including ASR [15], speaker verification [19], speaker identification [20], phoneme classification [21], emotion recognition [22], spoken language understanding [22], and TTS [23].

2.2. Child-adult classification in the ASD domain

Child-adult classification is among the more difficult tasks within speaker diarization, due to the challenges related to "in the wild" child speech in naturalistic conversational settings including short speaker turns, varied noise sources and a larger fraction of overlapping speech. Early diarization solutions involving child speech used traditional feature representations (MFCCs, PLPs) [24, 25]. In [26], the authors introduced several methods for processing audio collected from children with autism using a wearable device. Later, deep speech representations, i-vectors [26] and x-vectors [27] were studied for this task. A variety of challenges, both from signal processing and limited data availability, have been identified and addressed. In [11], the authors have proposed an adversarial training strategy to address the large within- and across-age and gender variability due to developmental changes in children. Alternatively, in [28], pre-trained x-vectors were fine-tuned for child/adult speaker diarization using a meta-learning paradigm, namely prototypical networks. Moreover, the role of the amount of child speech in building deep neural speaker representations was studied in [29] and their experimental results confirm that including more child data indeed enhances the task performance in a supervised setup.

3. Datasets

Our child-inclusive data come from interactions in a clinical setting, specifically obtained during the administration of two clinical protocols related to developmental disorders. The first protocol is the gold standard *Autism Diagnostic Observation Schedule* (ADOS) [6], used for diagnostic purposes. The second protocol is a recently proposed outcome-measure focused instrument *Brief Observation of Social Communication Change (BOSCC)* [7] for tracking changes in social and communicative skills during the course of treatment. A typical ADOS session lasts $40-60$ minutes and contains multiple (usually $10-15$) semi-structured activities for addressing specific symptoms related to ASD. Usually these interactions aim to elicit spontaneous responses from children under different circumstances to obtain a diagnostic score for classifying children with and without ASD. A BOSCC session is usually 12 minutes long, consisting of two $2min$ conversational talk sessions and two $4min$ play sessions where the child plays with a toy.

Table 1: *Session-level statistics of child-adult corpora.*

Dataset	**Duration** (mean $\pm std$)	**Child-speaking fraction** (mean $\pm std$)
Pre-training	14.05 ± 2.08	n.a
ADOSMod3	3.23 ± 1.61	0.46 ± 0.18
Simons	19.05 ± 12.86	0.40 ± 0.08

In our pre-training experiments, we use a dataset consisting of 369 recordings of unlabelled BOSCC sessions comprising approximately $100K$ utterances. For the fine-tuning experiments, we use two different corpora, ADOSMod3 and Simons. The ADOSMod3 corpus was collected across 2 clinical sites. These data are from administrations of the ADOS Module-3 designed for verbally fluent children, with a focus on the Social Difficulties and Annoyance and Emotional sub-tasks for this work. The data consist of total 346 sessions collected from 165 children (86 ASD, 79 Non-ASD). The Simons corpus used in our study consists of a combination of clinically administered ADOS ($n = 6$) and BOSCC ($n = 33$) sessions collected across 4 sites and these sessions were labeled by trained annotators to extract speaker timestamps. The details of datasets are reported in Table 1.

4. System Description

4.1. Pre-training

Our research aims to adapt the existing self-supervised approaches to the child-adult interaction domain through contrastive learning. Similar to [30], our contrastive learning framework is based on the assumption that neighboring segments from audio samples are highly likely to contain identical information. For instance, it is probable that adjacent audio frames are produced by the same speaker and are expected to contain similar semantic meaning, linguistic content, as well as acoustic characteristics. To elaborate, we define the dataset of audio samples as N, where each audio sample is denoted as x_i. The corresponding neighboring audio segment is represented as x_i', and is defined as any audio sample that has a time shift of half a second or less from the original sample x_i.

As outlined in the previous section, transformer-based models first transform the input speech sample x to intermediate features z using the feature encoder $f(\cdot)$ on the basis of CNNs. Subsequently, the transformer encoder $g(\cdot)$ maps the features z to contextualized representations c. Consequently, we can create similar pairs of contextualized representations c_i and c_i' from the neighboring audio segments x_i and x_i', with the remaining pairs being considered as negative pairs:

$$\text{Positive Pairs} : c_i \approx c_i' \tag{1}$$

$$\text{Negative Pairs} : c_i \neq c_k \text{ , } c_i \neq c_k', \text{where } i \neq k \tag{2}$$

Motivated by SimCLR [31], we apply the NTXent contrastive loss [32] as the pretraining objective with the adult-child conversational corpora. Given the temperature value τ, the loss function L_{NTXent} for the positive audio pairs x_i and x_i' within a batch of B input audio is:

$$-log \frac{exp(sim(c_i, c_i')/\tau)}{\sum_{\substack{k=0 \\ k \neq i}}^{B} exp(sim(z_i, z_k)/\tau) + \sum_{k=0}^{B} exp(sim(z_i, z_k')/\tau)} \tag{3}$$

Table 2: *Number of trainable parameters for the pre-training experiments based on unfrozen transformer layers*

Number of unfrozen transformer layers				
1	2	3	4	5
6.2M	13.5M	20.8M	27.1M	33.8M

Table 3: *Child-adult classification F1 score using W2V2. PT corresponds to pre-training and the following number represents the number of layers used for pre-training.*

Model	ADOSMod3		Simons	
	RNN	CNN	RNN	CNN
W2V2 - Base	67.92	70.59	63.41	64.13
W2V2 - PT1	69.31	72.41	65.19	66.28
W2V2 - PT2	71.55	72.95	65.87	65.12
W2V2 - PT3	72.23	74.38	**68.81**	65.44
W2V2 - PT4	**74.01**	**74.89**	67.63	**66.79**
W2V2 - PT5	72.19	74.05	65.01	65.39

4.2. Downstream Classifier Architectures

We use two different neural network models for child-adult speaker classification based on [33]. Both the classifiers include a self-attention based projector module, whereas one of them uses CNNs to capture speaker characteristics and the other uses *Recurrent Neural Netowrks* (RNN) to model the temporal dependencies present in the signal.

The RNN-based classifier consists of a stacked sequence of a *Feed Forward Layer* (FFL), a bidirectional *Long Short Term Memory* (LSTM) layer, a self-attention based projector layer and an output layer comprised of 2 FFLs, separated by a non-linear activation. The CNN classifier architecture is comprised of a weighted feature extraction module, followed by a convolutional module having 3 1D convolutional layers, each with a dropout and a non-linear activation in between, a self-attention based projector layer and an output layer comprised of 2 FFLs, separated by a non-linear activation. For all the experiments we use *Rectified Linear Unit (ReLU)* as the non-linear activation and a dropout ratio of 0.3.

5. Experimental Setup

5.1. Child Adult Classification

In this study, we hypothesize leveraging unlabelled child-speech for pre-training can guide models to learn the heterogeneous child speech and interaction patterns, leading to enhanced performance of downstream child-adult speaker classification. Instead of training from scratch, we pre-train the existing W2V2 and WavLM models with additional unlabelled child speech by unfreezing and updating specific transformer layers using a contrastive loss described in Sec 4.1. We report the child-adult classification macro F1-score on two labeled child-adult interaction corpus described in section 3. We report the results in Table 3 and Table 4, where the first row denotes downstream child-adult classification performance using the model solely relying on pre-trained embeddings. The subsequent rows denote downstream task performance using the models pre-trained with additional child speech, where the number indicates the number of trainable transformer layers involved in the pre-training task.

5.2. ADOSMod3 Experiments on Demographics

Our study also investigates the model performance across age-groups in the ADOSMod3 corpus. Prior works [11, 8] have reported age as an important variability factor impacting speech characteristics. Based on this hypothesis, we conduct an experiment by partitioning the ADOSMod3 corpus ($3-13yrs$) into three different age-groups (Age-group 1: 43-90 months, Age-group 2: 91-118 months, Age-group 3: 119-158 months), such that each group contains equal number of sessions. For each of these groups, we report the child-adult classification F1 score using the pre-trained base models of W2V2 and WavLM and also the best-performing pre-trained models of those two categories.

Table 4: *Child-adult classification F1 score using WavLM pre-training. PT corresponds to pre-training and the following number represents the number of layers used for pre-training.*

Model	ADOSMod3		Simons	
	RNN	CNN	RNN	CNN
WavLM-Base	72.73	73.09	71.78	70.25
WavLM - PT1	74.29	74.93	72.64	71.11
WavLM - PT2	**76.66**	75.81	**72.88**	**72.74**
WavLM - PT3	75.95	**76.37**	72.31	71.09
WavLM - PT4	75.18	75.92	72.01	71.59
WavLM - PT5	75.48	73.17	71.47	70.17

Not only age, analyses of developmental changes in speech have revealed sex ("gender") differences in speech characteristics, especially post puberty [8]. In this work, we also report gender-based child-adult speaker classification performance on ADOSMod3 dataset, with recordings from 244 male and 84 female individuals. Similar to age-focused experiments, the dataset is partitioned into male and female subsets, and comparisons are drawn between the base model and the best-performing pre-trained models for both W2V2 and WavLM.

5.3. Experimental details

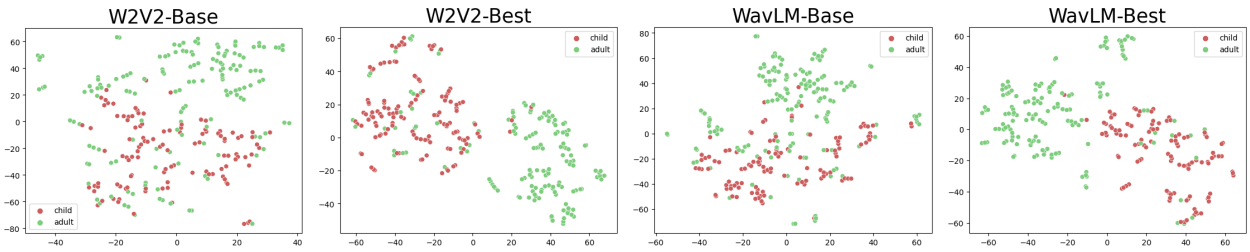
For both the pre-training and fine-tuning experiments, $Adam$ optimizer is used with a batch size of 32 samples and temperature is set to 0.1. The number of tunable parameters for the pre-training experiments is reported in Table 2. The initial learning rate is set to 1e-5 and the models are trained for 30 epochs with an early stopping callback on validation loss, patience being 5 epochs. For the downstream child-adult classification task, the model is trained to minimize the binary cross-entropy loss for a maximum of 50 epochs, while the initial learning rate for this experiment is 2e-4 with a weight decay of 1e-4. For both the datasets, we use 70% for training, 15% for validation and 15% for testing. We use the model checkpoints from HuggingFace [34]. We pre-train the models using a single NVIDIA GeForce GPU 1080 Ti and each experiment took less than two days.

6. Results and Discussion

6.1. Classification Evaluation

In this subsection, we analyze the experimental results reported in Table 3 and Table 4 to address the following questions:

Does pre-training with more child speech improve the classification? The results reveal that pre-training with additional child speech improves the child-adult classification F1 score over the base model. This underscores the models' ability to account for the heterogeneity that is inherent in children's speech. It can be observed that WavLM-based pre-trained models show better performance compared to W2V2 across all the experiments. Both the classifiers show comparable performance, with the RNN-based classifier yielding the best score in the majority of the experiments. Among the datasets, the experimental results reveal better F1 scores in ADOSMod3 compared to the Simons corpus. One possible reason might be related to the session recording length difference between the datasets. The



(a) *Session A: W2V2-Base* (b) *Session A: W2V2-PT4* (c) *Session B: WavLM-Base* (d) *Session B: WavLM-PT2*

Figure 2: *t-SNE plots of the most discriminative 2 components of the embedding space corresponding to the classes*

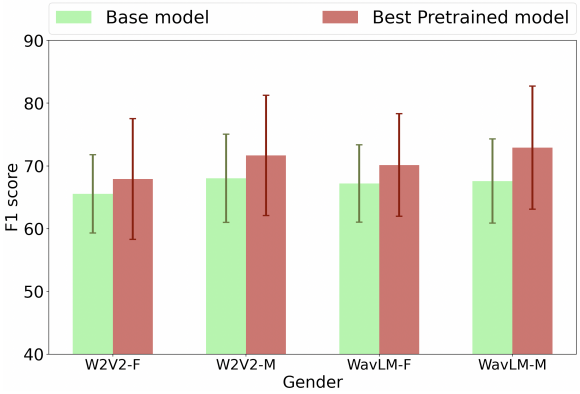


Figure 3: *Gender based Child-adult classification F1 scores.*

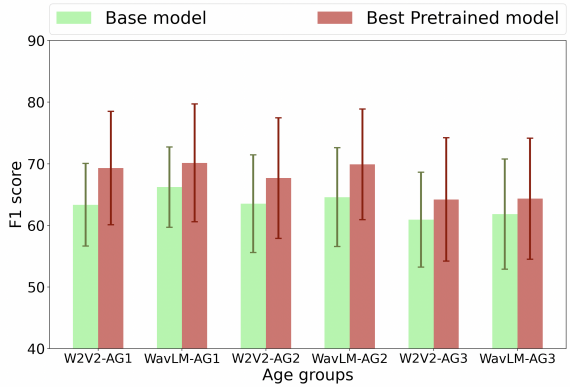


Figure 4: *Age based Child-adult classification F1 scores.*

average duration of sessions in Simons corpus is much higher than ADOSMod3, resulting in greater potential variability and heterogeneity, which may have degraded the F1 scores.

Does pre-training with more transformer layers improve the classification? It is interesting to note that, while in W2V2-based pre-training, the classification F1 keeps improving by tuning more transformer layers, in the case of WavLM, the performance improvements reach the maximum with tuning fewer transformer layers. One possible explanation is that the WavLM model is trained with an objective function to capture speaker related information, helping the model to achieve the optimum performance with lesser training. However, in both the scenarios, the model performance starts to degrade by adding more than four transformer layers. As these models are designed to provide generalized speech representations, tuning larger portions of these pre-trained models on a relatively smaller dataset might lead to the loss of generalizability, causing the performance to decrease for the classification task. However, our results provide compelling evidence that it is beneficial to adapt the last few transformer layers for the adult-child classification.

Qualitative analysis We present t-SNE visualizations of pre-trained embeddings for 2 output classes from two sessions in Figure 2. We plot the embeddings with and without additional pre-training. In both cases, it is evident from the plots that our method increases the discriminative information between them.

6.2. Result Evaluation based on Demographics

For the gender-focused experiments, the relative improvement in F1 scores are 6.39% and 3.14% for the male and female subsets. Possibly due to both inherent speech pattern differences and inherent data distribution biases (see Section 5.2), the models yield higher F1 scores in the male population than the female population. For the age-focused experiment, the relative improvements of 9.42%, 8.23%, and 4.06% are seen for the three age-groups (youngest to oldest). The results imply that it is intrinsically challenging to model children of AG1 and AG2 due to the developing vocal tract behaviors among these ages. As a consequence, adding more children speech in training data provides greater benefits to the model to capture more relevant information, resulting in more improvement in the younger age-group than the older ones.

7. Conclusion

Past work has demonstrated the promise of deploying self-supervised algorithms in a variety of downstream tasks like ASR, speaker diarization, and speaker verification [35, 19]. In this work, we investigate the utility of additional pre-training with more child speech, even in the presence of the inherent heterogeneity and variability, to improve child-adult speaker classification in clinical recordings involving interactions with children with autism. The experimental results with the proposed models support our hypothesis of benefiting from incorporating child speech based additional pre-training, across both age and gender dimensions of variability.

In this work, we used the manually-annotated ground truth labels for identifying and evaluating the speech and non-speech regions. In the future, we plan to build a child-adult diarization framework with an integrated *Voice Activity Detection* (VAD) system to further reduce the need of human effort. In addition, we plan to extend this study with an additional emphasis on early vocalization and speech (from toddlers and infants) in the interaction. Unlike verbally fluent children, toddler speech contains significant amounts of pre-verbal sounds and non-verbal vocalizations, which pose additional challenges for automated processing.

8. Acknowledgements

This work is supported by funds from USC Hearing, Communication and Neuroscience (HCN) pre-doctoral fellowship, NIH and Simons foundation.

9. References

[1] J. Volden and C. Lord, "Neologisms and idiosyncratic language in autistic speakers," *Journal of autism and developmental disorders*, vol. 21, no. 2, pp. 109–130, 1991.

[2] S. V. Huemer and V. Mann, "A comprehensive profile of decoding and comprehension in autism spectrum disorders," *Journal of Autism and Developmental Disorders*, vol. 40, pp. 485–493, 2010.

[3] S. H. Kim, R. Paul, H. Tager-Flusberg, and C. Lord, "Language and communication in autism," *Handbook of Autism and Pervasive Developmental Disorders, Fourth Edition*, 2014.

[4] T. Sorensen, E. Zane, T. Feng, S. Narayanan, and R. Grossman, "Cross-modal coordination of face-directed gaze and emotional speech production in school-aged children and adolescents with asd," *Scientific reports*, vol. 9, no. 1, p. 18301, 2019.

[5] Centers for Disease Control and Prevention (CDC), "Mental health in the united states: parental report of diagnosed autism in children aged 4-17 years–united states, 2003-2004," *MMWR. Morbidity and mortality weekly report*, vol. 55, no. 17, pp. 481–486, 2006.

[6] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, pp. 205–223, 2000.

[7] R. Grzadzinski, T. Carr, C. Colombi, K. McGuire, S. Dufek, A. Pickles, and C. Lord, "Measuring changes in social communication behaviors: preliminary development of the brief observation of social communication change (boscc)," *Journal of autism and developmental disorders*, vol. 46, pp. 2464–2479, 2016.

[8] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[9] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022.

[10] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech Language*, vol. 72, p. 101289, 2022.

[11] R. Lahiri, M. Kumar, S. Bishop, and S. Narayanan, "Learning domain invariant representations for child-adult classification from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6749–6753.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[14] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[17] S. Pascual, M. Ravanelli, J. Serra, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[18] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6989–6993.

[19] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[20] 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.

[21] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[22] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[23] D. Álvarez, S. Pascual, and A. Bonafonte, "Problem-agnostic speech embeddings for multi-speaker text-to-speech with samplernn," *arXiv preprint arXiv:1906.00733*, 2019.

[24] M. Najafian and J. H. Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 114–120.

[25] 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.

[26] T. Zhou, W. Cai, X. Chen, X. Zou, S. Zhang, and M. Li, "Speaker diarization system for autism children's real-life audio data," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016, pp. 1–5.

[27] J. Xie, L. P. Garcia-Perera, D. Povey, and S. Khudanpur, "Multi-plda diarization on children's speech." in *Interspeech*, 2019.

[28] N. R. Koluguri, M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, "Meta-learning for robust child-adult classification from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8094–8098.

[29] S. Krishnamachari, M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, "Developing neural representations for robust child-adult diarization," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 590–597.

[30] V. Sachidananda, S.-Y. Tseng, E. Marchi, S. Kajarekar, and P. Georgiou, "Calm: Contrastive aligned audio-language multirate and multimodal representations," *arXiv preprint arXiv:2202.03587*, 2022.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[32] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[33] T. Feng, R. Hebbar, and S. Narayanan, "Trustser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition," *arXiv preprint arXiv:2305.11229*, 2023.

[34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[35] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.