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

Rimita Lahiri¹, Tiantian Feng¹, Rajat Hebbar¹, Catherine Lord², So Hyun Kim³, Shrikanth Narayanan¹

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

rlahiri@usc.edu

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 phenotype. 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.

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.
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.

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 administeredADOS (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 $f(·)$ on the basis of CNNs. Subsequently, the transformer encoder $g(·)$ 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:

Positive Pairs: $c_i \approx c'_i$ \hspace{1cm} (1)

Negative Pairs : $c_i \neq c_k, c_i \neq c'_k$ \hspace{1cm} (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 $\tau$, the loss function $L_{NTXent}$ for the positive audio pairs $x_i$ and $x'_i$ within a batch of $B$ input audio is:

$$-\log \frac{\exp(\text{sim}(c_i, c'_i)/\tau)}{\sum_{k \neq i} \exp(\text{sim}(c_i, c_k)/\tau) + \sum_{k \neq i} \exp(\text{sim}(c_i, c'_k)/\tau)}$$

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Duration (mean ± s.d.)</th>
<th>Child-speaking fraction (mean ± s.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-training</td>
<td>14.05 ± 2.08</td>
<td>n.a</td>
</tr>
<tr>
<td>ADOSMod3</td>
<td>3.23 ± 1.61</td>
<td>0.46 ± 0.18</td>
</tr>
<tr>
<td>Simons</td>
<td>19.05 ± 12.86</td>
<td>0.40 ± 0.08</td>
</tr>
</tbody>
</table>

3558
Table 2: Number of trainable parameters for the pre-training experiments based on unfrozen transformer layers

<table>
<thead>
<tr>
<th>Number of unfrozen transformer layers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.2M</td>
<td>13.5M</td>
<td>20.8M</td>
<td>27.1M</td>
<td>33.8M</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>ADOSMod3-RNN</th>
<th>ADOSMod3-CNN</th>
<th>Simons-RNN</th>
<th>Simons-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>W2V2 - Base</td>
<td>67.92</td>
<td>70.59</td>
<td>63.41</td>
<td>64.13</td>
</tr>
<tr>
<td>W2V2 - PT1</td>
<td>69.31</td>
<td>72.41</td>
<td>65.19</td>
<td>66.28</td>
</tr>
<tr>
<td>W2V2 - PT2</td>
<td>71.55</td>
<td>72.95</td>
<td>65.87</td>
<td>65.12</td>
</tr>
<tr>
<td>W2V2 - PT3</td>
<td>72.23</td>
<td>74.38</td>
<td>68.81</td>
<td>65.44</td>
</tr>
<tr>
<td>W2V2 - PT4</td>
<td>74.01</td>
<td>74.89</td>
<td>67.63</td>
<td>66.79</td>
</tr>
<tr>
<td>W2V2 - PT5</td>
<td>72.19</td>
<td>74.05</td>
<td>65.01</td>
<td>65.39</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>ADOSMod3-RNN</th>
<th>ADOSMod3-CNN</th>
<th>Simons-RNN</th>
<th>Simons-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>WavLM - Base</td>
<td>72.73</td>
<td>73.09</td>
<td>71.78</td>
<td>70.25</td>
</tr>
<tr>
<td>WavLM - PT1</td>
<td>74.29</td>
<td>74.93</td>
<td>72.64</td>
<td>71.11</td>
</tr>
<tr>
<td>WavLM - PT2</td>
<td>76.66</td>
<td>75.81</td>
<td>72.88</td>
<td>72.74</td>
</tr>
<tr>
<td>WavLM - PT3</td>
<td>75.95</td>
<td>76.37</td>
<td>72.31</td>
<td>71.09</td>
</tr>
<tr>
<td>WavLM - PT4</td>
<td>75.18</td>
<td>75.92</td>
<td>72.01</td>
<td>71.59</td>
</tr>
<tr>
<td>WavLM - PT5</td>
<td>75.48</td>
<td>73.17</td>
<td>71.47</td>
<td>70.17</td>
</tr>
</tbody>
</table>

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 Networks (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 – 13yr s) 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.

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 trainable 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
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


