LEARNING CONVERSE-LEVEL MULTIMODAL EMBEDDING TO ASSESS SOCIAL DEFICIT SEVERITY FOR AUTISM SPECTRUM DISORDER

Chin-Po Chen[1,3], Susan Shur-Fen Gau[2], Chi-Chun Lee[1,3]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
stu94116@gapp.nthu.edu.tw, gaushufe@ntu.edu.tw, cclee@ee.nthu.edu.tw

ABSTRACT

Developing algorithms to automatically assess mental constructs using media data of human behaviors is becoming important, especially relevant for mental health applications. In this work, we focus on a critically prevalent neurodevelopment disorder, i.e., the autism spectrum disorder (ASD). While researchers have worked on automatic differentiation of ASD from healthy control using a variety of behavior modalities, few works have modeled the severity of ASD behavior symptoms in the existing clinical practice. Thus, we propose to learn a converse-level multimodal (speech and text) embedding derived during a severity assessment interview, i.e., the Autism Diagnosis Observation Schedule (ADOS), that considers the intricate interaction behaviors between the investigator and the participant. Further by fusing two attentional GRUs with this multimodal embedding, our approach achieves an averaged regression score of 0.567 on four items of socio-communicative constructs in the ADOS. Our analysis results suggest that the number of words uttered by both the investigator and the participant is a major predictor.

Index Terms— autism, GRU, BERT, ADOS

1. INTRODUCTION

The advancement of technology has enabled the large-scale collection of media data using commodity devices, such as audio-video recorders, in settings where human's behaviors, e.g., speech and body gestures, can be expressed spontaneously and naturally in real life. This provides a new venue for analyzing human behaviors objectively and in deriving relevant analytics for domain experts [1]. In fact, developing automated diagnostic algorithms for clinical decision and health condition monitoring has already become increasingly popular. For example, Liu and his colleagues proposed a computer-aided system for automatic cirrhosis diagnosis based on ultrasound images, and Lin et al. designed an algorithm for eyeball region segmentation to assist diagnosis of ophthalmic diseases [2, 3]. Also there are studies aimed at improving personal healthcare based on monitoring people's daily activities using multimodal media data [4, 5].

A recent position paper has laid out the importance in developing data-driven behavior computing algorithms especially for mental health applications, where assessing disordered symptoms, while critical but consistently suffers from issues of observational subjectivity [1, 6]. In this work, we focus on one of the most prevalent neurodevelopmental disorders, autism spectrum disorder (ASD). One in 160 children were diagnosed with ASD according to a study done in 2012 [7]. This high level of prevalence coupling with the fact that ASD is a heterogeneous neurodevelopmental disorder clinically characterized by different types of symptoms [8] makes development of computational assessment for ASD continues to be challenging yet critical for early detection, diagnosis and even intervention strategy.

Most of the previous studies concentrate on developing algorithms for automatic diagnosis of ASD, e.g., classification between ASD and typically developing children [9, 10] or differentiation between subgroups of ASD [11, 12]. Several studies have also quantified the atypicality of ASD as expressed in their acoustic, linguistic, and facial expressions [13, 10]. However, these previous works do not model the behaviors of ASD within the existing clinically-validated setting. In this work, our goal is to develop automatic algorithms in assessing behavior constructs evaluated in the Autism Diagnostics Observation Schedule (ADOS) [14]. ADOS has been regarded as a golden standard severity assessment scoring system for ASD. Although ADOS helps the medical professionals to track the efficacy of ASD intervention outcome and generally use as a clinical assessment for severity, the effort in training a qualified examiner and the natural subjec-

Thanks to Ministry of Science and Technology and National Health Research Institute (NHRI-EX104-10404PI, NHRI-EX105-10404PI) for funding

tivity in the manual observation make it time-consuming and create unwanted inconsistency.

In this work, we propose to learn a multimodal converse-level embedding to predict four different social constructs rated during an ADOS administration using an attentional GRU-DNN fusion architecture. Due to the nature of back-and-forth social interaction in ADOS, our multimodal embedding, i.e., capturing both acoustics and lexical modalities, is designed on a 'converse-level', i.e., a give-and-take spoken turn segment between the investigator and the subject, to model the intricate dependencies between the two interlocutors. Our framework demonstrates the importance of joint interlocutor modeling when predicting these socio-communicative scores. Further, our analysis reveals a distinct question-answering pattern, specifically obvious in the turn length, for those subjects with high severity ratings versus low severity ratings.

2. DATABASE

We collected audio-video recordings of ADOS interview sessions with their associated clinical ratings. In total, we collected digital data of 88 adolescent ASD patients at the National Taiwan University Children Hospital[1]. The ADOS instrument is a semi-structured assessment that contains a series of activities with two people involved: the investigator and the participant under assessment. All interactions happened in Mandarin, and each sentence was segmented and manually transcribed. The demographics of the participants are listed in Table1

In this study, we selected four ADOS items as an outcome to be predicted by our model. These items include: 'Offers information (OINF)' and 'Conversation (CONV)' from the communication dimension and 'Amount of Reciprocal Social Communication (ARSC)' and 'Insight (INS)' from the social reciprocity dimension. In these ADOS items, a higher score indicates a more severe communication impairment or social deficits. The distribution of the ADOS codes are listed in Table1. The four ADOS socio-communicative constructs used in our study are briefly described in the following.

Table 1: The rows show the mean and standard deviation of the participants' age as well as the scores distribution of the four ADOS algorithm items used as outcome to be predicted including OINF, CONV, INS, ARSC.

Age	16.59 /(3.95)				
Language	Mandarin				
ADOS codes	0 (normal)	1	2	3 (severe)	Total
OINF	46	29	13	X	88
CONV	25	43	18	2	
INS	8	33	40	7	
ARSC	45	24	19	X	

- OINF: To evaluate if the participant could provide essential information such as the subject of the event, the time order of the event, and the location of event occurrence.
- CONV: To evaluate if the participant could initiate and maintain conversations that are sensitive to social context. Qualified conversations are that the participant commented on what the investigator just said and maintained this conversation for at least four rounds.
- ARSC: If qualified conversation (described in CONV), which is restricted to certain topics, happened many times, the participant will be coded with a score of 0.
- INS: To evaluate if the participant could recognize social relationships. First of all, the investigator might talk about school, friends, or teachers. Then the investigator code this item base on how the participant describe them.

3. METHOD

Our framework involves three parts (Fig 1): first, we split the interaction into units of 'converse-level' segments; then, we utilize two different embedding approaches for text (BERT encoding) and speech acoustics (statistical encoding); finally, these multimodal converse-level embeddings are fed into gated recurrent unit networks (GRUs) followed by dense layers with attention mechanism to perform regressions.

3.1. Converse-level Unit Definition

We separated each dialogue into multiple turn-exchanging blocks. Each block was treated as a simple unit of turn containing one or many utterances spoken by only one person, and the next block contained only the other person's utterances. A converse-level unit, then, takes a pair of investigator's and participant's turns to be the basic unit in this work (an example of a converse-level unit is shown in Fig 1 under 'Preprocessing' tag).

3.2. Converse-level Embedding (Lex)

We choose Bidirectional Encoder Representation from Transformers (BERT) to be our lexical encoder. BERT is a stack of 12-layers transformer, where the layer at the end encodes higher level of semantic information [15]. We first take a pre-trained BERT, that is pre-trained on the setting of bert-based-multilingual-uncased containing 104 languages. The input to the BERT is either turn or converse-level sequence of words. The output vector corresponding to the '[CLS]' token is used as the

[1] Approved by IRB: REC-10501HE002 and RINC-20140319

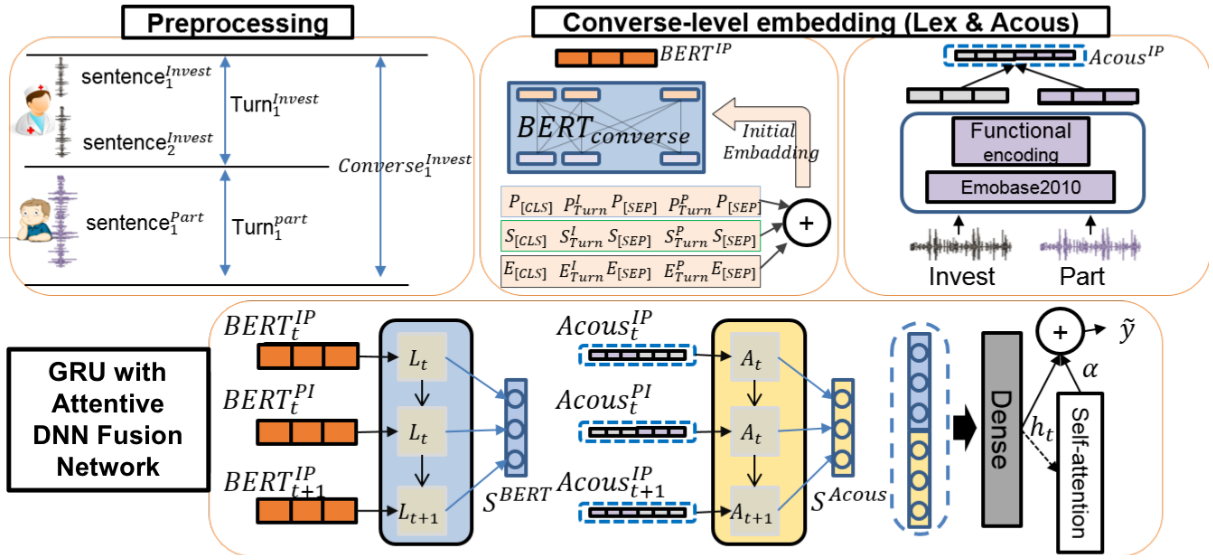

Fig. 1. An overview of our method. First (at the top left) we preprocess our data by splitting the corpus into several turns. Then we use two different methods to encode the acoustic and lexical modalities(at the top right). At the bottom shows GRUs with an attentive DNN multimodal fusion network used in predicting the ADOS codes.

turn/converse-level representation of lexical embedding, which is similar to past works [16, 17]. A brief description of the details are described below.

3.2.1. Lexical Embedding using BERT

A special token '[CLS]' is appended in the front of each input turn (a sequence of words from a single speaker before floor change), whereas '[SEP]' is appended at the back. The new word sequence ([CLS]+$Turn^{Part}$+[SEP]) is then transformed into the initial embedding. The initial embedding is the summation of three kinds of simple word embeddings that contain: word index number, word position, segment id's. The initial embedding is further fed into the BERT network to form a word embedding sequence with each word token corresponding to a BERT vector (dimension 768). Then, the BERT vector corresponding to the token of '[CLS]' from the last output layer is used as our turn-level lexical embedding (the segment id's are set to 0's for turn-level embeddings).

Furthermore, we extend turn-level embedding to converse-level embedding by concatenating investigator's turn and participant's turn separated by the token '[SEP]' (Fig 1 the left box of Converse-level embedding). This results in the following pattern: [CLS]+$Turn^{Invest}$+[SEP]+$Turn^{Part}$+[SEP]. Besides, we indicate the relationship between the two turns by modifying the segment id. We use a sequence of zeros as the investigator's segment id (S_{Turn}^{I} $= 0$ in Fig 1) and a sequence of ones as the participant's segment id ($S_{Turn}^{P} = 1$) in order to derive a final converse-level feature $BERT_{IP}$. On the contrary, $BERT_{PI}$ puts the two turns in a reverse order ([CLS]+$Turn^{Part}$+[SEP]+$Turn^{Invest}$+[SEP] and $S_{Turn}^{P} = 0$, $S_{Turn}^{I} = 1$).

The fine-tuning of the initial BERT into our ADOS interaction dataset is implemented using the following procedure. We optimize the output of '[CLS]' token to correctly recognize consecutive turns within our ADOS corpus. We finetune the pre-trained BERT to our corpus for 20 epochs. The evaluation accuracy of this fine-tuning is based on masked LM and next sentence prediction (introduced in [15]) tasks, and this fine-tuned BERT achieves accuracy of 0.71 and 0.91 respectively.

3.3. Converse-level Embedding (Acous)

We first compute low-level acoustic descriptors (LLDs) of dimension 52, which contains 12-dimensional Mel-Frequency Cepstral Coefficients(MFCCs), 7 dimensional Line Spectral Pairs (LSPs), intensity, loudness, F0, F0 envelope, zero-crossing rate, voice probability and their first-order derivatives using the OPENSMILE toolbox [18]. The LLD's are derived within each 60ms window and the step size is set to be 10 ms. We further apply session-wise Z-score normalization on these LLDs for each ADOS interaction sample.

These acoustic features are then split into converse-level blocks to align with the converse-level lexical features described in section.3.2. Within each turn, 15 dimensional functional encoding that includes: maximum, minimum, mean, median, standard deviation, 1^{st} percentile, 99th percentile, the difference between 1^{st} and 99^{th} percentile, skewness, kurtosis, relative position of

Table 2: Spearman ranking correlation results between the model prediction and the true scores. The rows are four different tasks 'OINF',CONV', 'ARSC','INS' with the average correlation obtained, and the columns are four different levels of embeddings. The subscript shows the input features of that particular level of embedding.

	S_A	S_{W2VQT}	T_A	T_{W2VQT}	$T_{W2VQT+A}$	T_{BERT+A}	J_{BERT+A}	C_{BERT+A}
OINF	0.004	0.460	0.105	0.519	0.556	0.511	0.578	0.582
CONV	-0.048	0.493	0.170	0.557	0.599	0.531	0.580	0.519
INS	0.011	0.428	0.019	0.304	0.293	0.369	0.226	0.523
ARSC	-0.037	0.492	0.188	0.530	0.535	0.594	0.571	0.645
AVERAGE	-0.018	0.468	0.121	0.478	0.496	0.501	0.489	0.567

minimum value, relative position of maximum position, the first and third quartiles, and interquartile range (see the right box of Converse-level embedding in Fig 1). Finally, the turn-level acoustic features (780 dimensions) from the investigator and the participant are concatenated to form the converse-level acoustic embedding.

3.4. GRU with Attentive DNN Fusion Network

After obtaining converse-level lexical embedding and acoustic features, we further train two modality-specific GRU networks that are further fused using a DNN to perform ADOS scores regressions. GRU networks are defined as below:

$$r_t = \sigma(W_r x_{t-1} + U_r x_t) \tag{1}$$

$$z_t = \sigma(W_z x_{t-1} + U_z x_t) \tag{2}$$

$$\hat{s}_t = tanh(W(r_t \circ s_{t-1}) + U x_t) \tag{3}$$

$$s_t = (z_t \circ s_{t-1}) + (1 - z_t) \circ \hat{s}_t \tag{4}$$

where r_t, z_t, $\hat{s}_t$, s_t are reset gate, update gate, candidate current state, current state, and the operator $\circ$ represents Hadamard product. Referring to eq.3 and eq.4, the candidate current state ($\hat{s}_t$) is determined by the current input x_t and a proportion of previous state controlled by reset gate(r_t). Finally, the current state (s_t) is determined by linear combination of previous and candidate current state weighted by the update gate (z_t). x_t indicates either acoustic or lexical converse-level embedding. We take the outputs of the two GRU networks to perform fusion with a dense layer. We further apply attention mechanism in this fusion dense layer:

$$h_t = W_d[s_t^{Acous}; s_t^{Bert}] + b_d \tag{5}$$

$$\hat{y_i} = \Sigma_{j=1}^{T_i} \alpha_{ij} h_j \tag{6}$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\Sigma_{j=1}^{T_i} exp(e_{ij})} \tag{7}$$

$$e_{ij} = W_a h_t + b_a \tag{8}$$

The outputs s_t^{Acous} and s_t^{BERT} are concatenated and passed into a dense layer (eq.5). The output of the dense layer h_t is multiplied by attention weight α_{ij} and summed up to be the final output $\hat{y_i}$, where T_i is the total number of turns of one session. The attention layer is implemented using a dense layer (eq.8) with a softmax layer (eq.7).

4. EXPERIMENTS

In this work, we compared different levels (sentence, turn, joint, converse) of embeddings. Sentence-level embedding means the unit of a time step is a sentence (eg. $sentence_1^{Invest}$,$sentence_2^{Invest}$ refer to Fig 1) instead of a turn. Turn-level embedding treats each turn block as a single unit. Joint-level embedding is implemented by simple feature concatenation of investigator's turn and participant's turn. The basic unit of joint-level embedding is the same as converse-level embedding. Finally, converse-level embedding is our proposed framework. We also compared BERT to Quick Thought vector[19] proposed in 2018 as the lexical encoder. Our Quick Thought comparison model is pretrained using the following steps: we first pre-trained the Word2Vec encoder and Quick Thought encoder using Chinese Wikipedia, and fine-tune both of them onto our ADOS corpus. The description of the comparison models (refer to the columns in Table2) are listed below:

- S_A: Sentence-level acoustic feature baseline.
- S_{W2VQT}: Sentence-level Word2Vec with Quick Thought embedding baseline.
- T_A: Turn-level acoustic feature baseline.
- T_{W2VQT}: Turn-level Word2Vec with Quick Thought embedding baseline.
- $T_{W2VQT+A}$: Turn-level fusion of acoustic and Word2Vec with Quick Thought embedding.
- T_{BERT+A}: Turn-level fusion of acoustic and BERT embedding.
- J_{BERT+A}: Joint-level fusion of acoustic and BERT embedding.
- C_{BERT+A}: Proposed Converse-level fusion of acoustic and BERT embedding.

A five fold cross-validation scheme using metric of spearman correlation is used for evaluation. We conduct each

Table 3: The table shows the average turn length of the lowest, middle, highest groups. We reported the average turn lengths of both the participant and the investigator separately in each group.

Analysis of Turn Length								
	OINF		CONV		INS		ARSC	
	Investigator	Participant	Investigator	Participant	Investigator	Participant	Investigator	Participant
Lowest	7.909	15.341	8.705	15.409	10.409	15.227	9.250	15.795
Middle	12.068	9.909	12.977	9.795	13.295	9.318	11.432	9.818
Highest	19.273	4.932	16.659	5.091	14.318	6.227	19.205	4.659

experiment five times and report the average correlation in Table 2. The hyperparameters of the network are listed below: all of the GRU networks contain 1 layer GRU cell. The dropout rate and the hidden nodes are selected to be 30% and 8 respectively. The model is optimized by minimizing MSE loss using stochastic gradient descent with the optimizer, ADAMW [20, 21].

4.1. Results

We first compare the turn-level embedding to sentence-level embedding (the left most Table 2). The average scores of the turn-level embedding models are better than the sentence-level embedding models, and the results of acoustic modality only (S_A, T_A) are worse than the lexical modality (S_{W2VQT}, T_{W2VQT}). We also observe that results of fusing acoustic and lexical features at turn-level ($T_{W2VQT+A}$) already improves the average spearman correlations across the four codes from 0.478 to 0.496. Furthermore, by replacing Quick Thought vector with BERT in the lexical modality, the correlation improves to 0.501. Finally, to model the interaction of both interlocutor, we implemented Joint-level and Converse-level fusion models. While Joint-level fusion model is not better than turn-level fusion model, our proposed Converse-level fusion model obtained the best average score suggesting the importance of joint interlocutor modeling, specifically using the GRUs with attentive dense fusion layer.

Our proposed converse-level fusion model obtains results of 0.582, 0.519, 0.523, 0.645 (spearman correlation) when regressing ADOS codes of OINF, CONV, INS, ARSC. Out of these four social communicative ratings, we obtain the best accuracy when using converse-level model in all of them except for CONV. CONV code is coded on whether the participating subject could initiate or try to maintain a conversation. Turn-level has the highest accuracy may indicate that in this scenario, either the participant's or the investigator's turns by themselves already provide adequate information without the need of sophisticated modeling such as our proposed converse-level fusion in attempting to capture the social exchange phenomenon.

4.2. Analysis and Discussion

We further conduct a turn length analysis to demonstrate the insights derived from using our proposed framework. One of the key intuitive indicators in predicting the final score that we have identified is that the number of words used in each turn (referred to as turn length in this paper). We analyze the turn length as a function of the low, mid, high three scoring groups for each ADOS codes used in this paper. The three groups are defined with the following criteria: 'Lowest' means the samples of the lowest 25% regressed values (closer to normal); 'Highest' means the highest 25% regressed values (higher severity); 'Middle' indicates the subgroups in between.

4.3. Analysis of Turn Length

The results are summarized in Table 3. The participant columns show an obvious pattern that the lowest group has the highest turn length value. It is quite intuitive that the lower the social-communicative symptom severity the more words that the participant would say to engage in smooth conversation with the investigator. Interestingly, we see a clear opposite trend for the investigator. It demonstrates an interesting dependency between the two interlocutors that is conditioned on the severity of the subject, i.e., the better-abled participant leads to a situation where the investigator says less, and the more severely impaired subjects speaks less leading to a situation where the investigator needs to say more. This phenomenon has also been identified in the ADOS administrated in English according to a previous study by Bone et al.[6]. While we identify that word length as one of the key intuitive indicators in analyzing the difference between the three scoring subgroups, many of these are also encoded in the semantics (what) and style (how), that are further captured with our proposed converse-level BERT model fused with acoustic descriptors.

4.4. Conclusions

In this work, we propose a converse-level multimodal (speech and text) embedding with a GRU-DNN attention networks to automatically regress on four ADOS items (socio-communication related measure, i.e., OINF, CONV, INS, ARSC). Our proposed model obtains spearman correlation of 0.582, 0.519, 0.523, 0.645 for the tasks OINF, CONV, INS, ARSC outperforming a variety of baseline models. Finally, our analysis implies that longer participant's turn length and shorter investigator's turn length seems more likely to be correlated to a lower severity symptoms. More detailed analysis, for example, the exact semantic patterns and the acoustic manifestation, which are two additional key components modeled in our framework will be included in our future work.

5. REFERENCES

[1] Shrikanth Narayanan and Panayiotis G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," Proceedings of the IEEE, vol. 101, no. 5, pp. 1203–1233, 2013.

[2] Xiang Liu, Zhiqin Zhan, Ming Yan, Jingwen Zhao, Jialin Song, and Yan Qiu Chen, "Computer-aided cirrhosis diagnosis via automatic liver capsule extraction and combined geometry-texture features," in 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017, pp. 865–870.

[3] Fanchao Lin, Chuanbin Liu, Hongtao Xie, Zheng-Jun Zha, and Yongdong Zhang, "Semantic-embedding and shape-aware u-net for ultrasound eyeball segmentation," in 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019, pp. 892–897.

[4] Yasmina Andreu-Cabedo, Pedro Castellano, Sara Colantonio, Giuseppe Coppini, Riccardo Favilla, Danila Germanese, Giorgos Giannakakis, Daniela Giorgi, Marcus Larsson, Paolo Marraccini, et al., "Mirror mirror on the wall... an intelligent multisensory mirror for well-being self-assessment," in 2015 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2015, pp. 1–6.

[5] Duc Nhan Tran, Hyukzae Lee, and Changick Kim, "A robust real time system for remote heart rate measurement via camera," in 2015 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2015, pp. 1–6.

[6] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," IEEE Signal Processing Magazine, vol. 34, no. 5, pp. 196–195, 2017.

[7] Mayada Elsabbagh, Gauri Divan, Yun-Joo Koh, Young Shin Kim, Shuaib Kauchali, Carlos Marcín, Cecilia Montiel-Nava, Vikram Patel, Cristiane S Paula, Chongying Wang, et al., "Global prevalence of autism and other pervasive developmental disorders," Autism research, vol. 5, no. 3, pp. 160–179, 2012.

[8] Anne Masi, Marilena M DeMayo, Nicholas Glozier, and Adam J Guastella, "An overview of autism spectrum disorder, heterogeneity and treatment options," Neuroscience bulletin, vol. 33, no. 2, pp. 183–193, 2017.

[9] Ming Li, Dengke Tang, Junlin Zeng, Tianyan Zhou, Huilin Zhu, Biyuan Chen, and Xiaobing Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," Computer Speech & Language, vol. 56, pp. 80–94, 2019.

[10] Sunghye Cho, Mark Liberman, Neville Ryant, Meredith Cola, Robert T Schultz, and Julia Parish-Morris, "Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations," Proc. Interspeech 2019, pp. 2513–2517, 2019.

[11] Yun-Shao Lin and Chi-Chun Lee, "Using interlocutor-modulated attention blstm to predict personality traits in small group interaction," in Proceedings of the 2018 on International Conference on Multimodal Interaction. ACM, 2018, pp. 163–169.

[12] Chin-Po Chen, Susan Shur-Fen Gau, and Chi-Chun Lee, "Toward differential diagnosis of autism spectrum disorder using multimodal behavior descriptors and executive functions," Computer Speech & Language, vol. 56, pp. 17–35, 2019.

[13] Yuan Tian, Xiongkuo Min, Guangtao Zhai, and Zhiyong Gao, "Video-based early asd detection via temporal pyramid networks," in 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019, pp. 272–277.

[14] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael 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, no. 3, pp. 205–223, 2000.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805v2, 2018.

[16] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger, "On measuring social biases in sentence encoders," arXiv:1903.10561v1, 2019.

[17] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu, "Understanding the behaviors of bert in ranking," arXiv:1904.07531v4, 2019.

[18] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010, pp. 1459–1462.

[19] Lajanugen Logeswaran and Honglak Lee, "An efficient framework for learning sentence representations," arXiv:1803.02893v1, 2018.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980v9, 2014.

[21] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," arXiv:1711.05101v3.