Interspeech 2018
2-6 September 2018, Hyderabad

An Interlocutor-Modulated Attentional LSTM for Differentiating between Subgroups of Autism Spectrum Disorder

Yun-Shao Lin[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
astanley18074@gmail.com, gaushufe@ntu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Recalling and discussing personal emotional experiences is one of the key procedures in assessing complex affect processing of individuals with Autism Spectrum Disorder (ASD). This procedure is a standard subpart of a diagnostic interview to assess ASD - the Autism Diagnostic Observation Schedule (ADOS). Previous work has demonstrated that the behavior features computed from this procedure in ADOS possess discriminative information between the three distinct ASD subgroups: Autistic Disorder (AD), High Functioning Autism (HFA), and Asperger Syndrome (AS). In this work, we propose an interlocutor-modulated attentional long short term memory network (IM-aLSTM) that models the ASD individual's acoustic features with a novel interlocutor-modulated attention mechanism. Our IM-aLSTM achieves ASD subgroup categorization accuracy of 66.5%, which is a 14% absolute improvement over baseline method on the same database. Our analyses further indicate that the attention weights are concentrated more on interaction segments where the ASD individual is being asked to recall and discuss his/her own negative emotional experiences.
Index Terms: behavioral signal processing (BSP), autism spectrum disorder, dyadic interaction, attention mechanism

1. Introduction

Self-disclosure is a dynamic process where people reveal and reflect on personal information, including thoughts, feelings, and experiences about themselves to another person [1]. Face-to-face spoken communication is an interactive and useful mean for carrying out such a process [2]. In fact, in the clinical application of psychotherapy, it is imperative for therapists and patients to engage in dyadic interviews; research has demonstrated that appropriate therapeutic strategy leading to patients' self-disclosure during the interactions is positively correlated to the success of the therapy [3, 4]. This back-and-forth interactive procedure is not only being used in clinical intervention but also used in the assessment of socio-emotional and socio-communicative skill, particularly for individuals with Autism Spectrum Disorder (ASD).

Aside from inadequate socio-communicative skill, ASD individuals exhibit further deficits in complex emotion processing [5], e.g., they have difficulties in accurately recognizing other's emotional states [6, 7]; ASD children also react differently to personal negative emotional experiences than typically-developing (TD) children, suggesting an impaired mechanism in self-awaring negative emotional episodes [8]. As part of the standard ASD diagnostic interview instrument, i.e., Autism Diagnostic Observation Schedule (ADOS), the investigator would also engage the participant in spoken conversation in order for the subject to self-disclose (talk about) his/her past emotion experiences (this assessment is often termed as the the *Emotion* part in the ADOS interview). The spontaneous and interactive nature of the Emotion part has further made this the focal point of recent computational studies into modeling communicative aspect of ASD. For example, Bone et al. analyzed the subtle "atypical" prosodic variation and the synchronized patterns between the investigator and the participant as a function of the severity of autism [9]. They further examined the ASD severity manifested in the acoustic-prosodic and turn-taking dynamics during the Emotion part of the ADOS [10].

In this work, we concentrate our analyses also on the Emotion part toward differential diagnosis between the three subgroups of ASD: Autistic Disorder (AD), High Functioning Autism (HFA), and Asperger Syndrome (AS). Several supporting research suggests that the differences between these three ASD subgroups is currently indistinguishable for clinicians [11, 12, 13], and the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders version 5 [14]) has merged the three subgroups into a single spectrum. However, competing research demonstrates contradictory evidence [15, 16]. A deeper understanding between these three subgroups is important not only in helping to find etiology and cause of ASD but also for developing a more targeted treatment [17, 18, 19]. Recent work presented by Chen et al. has shown initial empirical evidence that by computing low-level behavior descriptors of the participant, the interlocutor, and interaction between the two during the *Emotion* part of the ADOS, it can differentiate the three subgroups of ASD [20].

We propose an interlocutor-modulated attentional LSTM (IM-aLSTM) network architecture to perform the same recognition task by modeling the participant's acoustic features during the *Emotion* part. Specifically, we introduce a novel *interlocutor-modulated* attention mechanism where the participant's LSTM is learned by jointly integrating discriminative information of the dyad together (both the investigator and the participant). Our IM-aLSTM achieves a promising unweighted recall of 66.5% in three subgroup categorization, which outperforms Chen et al. by 14% absolute on the same dataset [20]. IM-aLSTM shows an improvement of 11.57% relative over using participant-only attention mechanism, which reinforces the importance of integrative modeling of the interlocutors. Lastly, our analysis shows an interesting result that the learned attention weights are concentrated in regions where the participant is being asked to recall and describe *negative* emotion experiences an indication that the difference between the three subgroups may be related to the behavior exhibited during the interactive spoken interaction of self-disclosing *negative* emotion episodes.

The rest of the paper is organized as follows: Section 2 introduces our framework along with the database and detail

10.21437/Interspeech.2018-1288

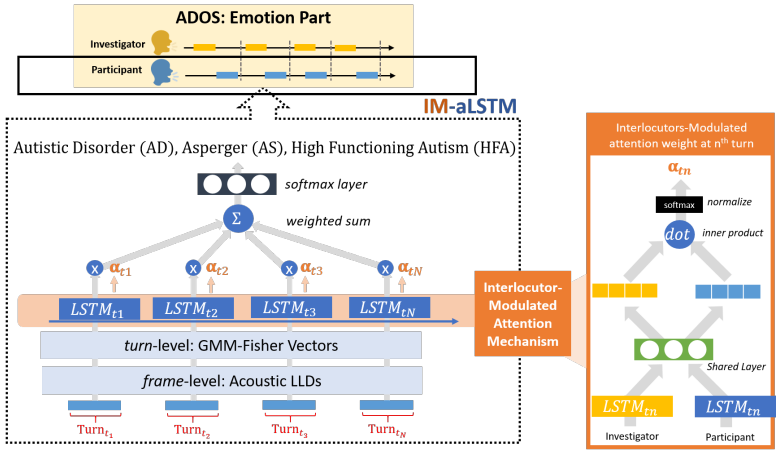

Figure 1: *A Schematic of our Interlocutor-Modulated Attentional LSTM: Our IM-aLSTM introduces an Interlocutor-Modulated Attention Mechanism to emphasize the important turn-feature during dyadic face-to-face interaction in the Emotion part of the ADOS. The turn-level feature is a fixed high-dimensional acoustic feature encoded using GMM-based Fisher scoring. We model the progress of the turn-level features using LSTM with the Interlocutor-Modulated Attention Mechanism. Finally, the learnable weight α_{tn} with LSTM can be used to differentiate between the three ASD subgroups.*

ASD Clinical Diagnosis			
Diagnosis (Number)	AD (28)	AS (20)	HFA (12)
Age (Avg/Std)	15.04/3.07	15.95/3.28	18.58/4.42
Module (M3/M4)	23/5	11/9	3/9

Table 1: *Information about ASD participants and their clinical diagnosis in our dataset*

methodology. Section 3 summarizes our experimental results and discussions. Section 4 is a conclusion and future work.

2. Research Methodology

2.1. The ADOS Audio-Video Database

Our ADOS audio-video database[1] is collected at the Department of Psychiatry of the National Taiwan University Hospital (NTUH).The ADOS session is a semi-structured dyadic interview between the clinical investigator and the ASD participant. To elicit targeted socio-communicative behaviors from the participant, the design of ADOS includes a series of activities, e.g., communication, social interaction, socio-emotional questions, imaginative use of materials, etc. In this work, we utilize the *Emotion* part of the ADOS session as our analysis data. The *Emotion* part includes a spontaneous conversation between the investigator and the participant; the investigator utilizes a semi-structured method in guiding the participant to discuss their past emotional experiences in daily life - specifically focusing on the four basic emotional experiences: happy, angry, fear, and depressed. Each session of the *Emotion* part lasts about 5-7 minutes. The semi-structured format of the *Emotion* part usually involves the investigator engage the participant in a conversation as follows:

Investigator: Do you feel the [emotion] sometimes?
Participant: [Yes or No], when I ..... .
Investigator: What happens, when you are [emotion] ?
Participant: I ..... .
Investigator: Can you describe the feeling of the [emotion]?
Participant: I ..... .

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

The database includes audio recordings collected using two separate wireless lapel microphones, i.e., one each for the investigator and the participant. Table 1 summarizes the database information. In total, we have collected ADOS interviews of 60 ASD subjects: 28 of them are diagnosed as AD, 20 of them are AS and 12 of them are HFA. The diagnostic outcome is determined based on a combination of clinical diagnosis by SSG, a senior child psychiatrist, ADOS, and Autism Diagnosis Interview-Revised (ADIR) [21], and other relevant clinical interviews and assessments. This database is also one of the largest clinically-valid research-level audio-video databases of ADOS interaction sessions.

2.2. Interlocutor-Modulated Attentional LSTM

Figure 1 shows our proposed interlocutor attentional LSTM (IM-aLSTM) architecture. This LSTM is learned from the input of the ASD participant's vocal features. The time step is at every *turn*, i.e., a complete speaking portion of the participant before the speaker floor is changed to the investigator. In this following sections, we will describe the extraction of *turn*-level acoustic inputs for LSTM and the proposed interlocutor-modulated attention mechanism.

2.2.1. Turn-level Acoustic Features

The turn-level acoustic features are computed on the speaking portion of the participants. First, we segment the *Emotion* part into multiple turn-taking event regions. Due to the question-answer nature of the *Emotion* part, each region includes data of a complete floor exchange in the form of "the investigator - the participant." Within each *turn*, we extract *frame*-level acoustic low-level descriptors (LLDs) including pitch, intensity, harmonic-to-noise ratio (HNR), MFCC, and their delta and delta-delta using the Praat toolkit [22]. Pitch, intensity, MFCC and the HNR are all extracted at a framerate of 10ms; these LLDs are z-normalized with respect to each speaker.

Each of the turns includes a varying number of LLD sequences. We further encode this sequence of LLDs using a Gaussian Mixture Model (GMM) based Fisher scoring [23] to derive a fixed high-dimensional acoustic at the *turn*-level. This particular method has been shown to be useful in speech-related

Models	Participant-only Methods			Interlocutor-Modulated Attention Mechanism	
	M0 (Baseline [20])	M1 (PO-LSTM)	M2 (PO-aLSTM)	M3 (IB-aLSTM)	M4 (IM-aLSTM)
UAR	0.525	0.576	0.596	0.628	**0.665**
AD	0.619	0.571	0.571	0.750	0.679
AS	0.500	0.550	0.550	0.550	0.650
HFA	0.455	0.667	0.667	0.583	0.667

Table 2: *Comparison of model performance from M0 to M4. It shows the unweighted average recall (UAR). The overall result show that the Interlocutor-Modulated Attention Mechanism outperforms the Participant-only methods. Our proposed IM-aLSTM achieves 66.5% UAR on differentiating the three ASD subgroups and outperform the past work by 14%.*

recognition tasks of emotion [24] and para-linguistic attribute [25], also in the assessment of impromptu speech [26].

2.2.2. *Interlocutor-Modulated Attention Mechanism*

We utilize forward long short term memory neural network (LSTM) [27] as our model to process time-dependent progression of turn-level vocal features of the participant to recognize the three subgroups of ASD. LSTM is an improvement over recurrent neural network (RNN), where the introduction of forget gate enables LSTM to capture longer time-steps and richer contextual information mitigating issues of gradient vanishing.

For the "i-th" participant's session, we input the participant's turn-level feature sequence x (section 2.2.1) to obtain a corresponding output sequence of LSTM's hidden states, h:

$$\{h_{i1}, \quad \dots \quad, h_{iT}\} = LSTM_{\text{part}}(\{x_{i1}, \quad \dots \quad, x_{iT}\}) \tag{1}$$

Furthermore, the use of attention mechanism in LSTM time-series modeling [28] has been shown to be effective across a variety of recognition tasks, e.g., motion recognition [29], emotion recognition [30], prominent counselor and client behaviors during addiction counseling [31], etc. The attention mechanism is achieved by placing a learnable weight in the network to emphasize the important parts of the time series. In our work, we also utilize the attention mechanism in our LSTM architecture to emphasize the turns within the *Emotion* part in our three subgroup recognition tasks. In specifics, we propose to learn a novel *interlocutor-modulated* attention weights, α_i, instead of conventional attention weights.

The *interlocutor-modulated* attention weights intend to capture the time-dependent interactive relationship between the interlocutors (the architecture of this attention mechanism is shown in Figure 1). We first additionally train an investigator's LSTM using the same set of turn-level acoustic features. Then, the hidden state sequences of g_i (the investigator) and h_i (the participant) are:

$$\{g_{i1}, \quad \dots \quad, g_{iT}\} = LSTM_{\text{invt}}(\{y_{i1}, \quad \dots \quad, y_{iT}\}) \tag{2}$$

$$\{h_{i1}, \quad \dots \quad, h_{iT}\} = LSTM_{\text{part}}(\{x_{i1}, \quad \dots \quad, x_{iT}\}) \tag{3}$$

In order to learn the non-linear relationship between these hidden state sequence of g_i and h_i. We add a shared fully-connected layer to these two hidden states to map these two sequences to a shared space:

$$f(h_{it}) = Relu(W_u h_{it} + b_u) \tag{4}$$

$$f(g_{it}) = Relu(W_u g_{it} + b_u) \tag{5}$$

The parameters W_u and b_u from this shared interaction layer are trained jointly. We then assign a modified attention weight u_{it} for the "i-th" participant at "t-th" time-step using:

$$u_{it} = \langle f(g_{it}), f(h_{it}) \rangle \tag{6}$$

Next, we obtain the time-normalized attention weight α_{it}:

$$\alpha_{it} = \frac{exp(u_{it})}{\sum_t^T exp(u_{it})} \tag{7}$$

These *interlocutor-modulated* attention weights are combined to the participant's LSTM's hidden vectors h_{it} using the following equation:

$$s_i = \sum_t \alpha_{it} h_{it} \tag{8}$$

The final recognition of the three groups of ASD, y_i, can be obtained using softmax function as the output layer, i.e.,

$$y_i = softmax(s_i) \tag{9}$$

3. Experimental Setup and Results

3.1. Experimental Setup

3.1.1. *Models Comparison*

We compare our proposed framework with four different models in task of differentiating between the three ASD subgroups: AD, AS, and HFA.

- M0-Baseline:
  The method previously proposed by Chen et al. [20] to perform recognition by computing dyadic low level behavior descriptors on the same dataset [20]
- M1-Participant-only LSTM:
  Using the participant's vocal LSTM with average pooling to differentiate between the three ASD subgroups without attention mechanism.
- M2-Participant-only Attentional LSTM:
  Using the participant's vocal LSTM to differentiate between the three ASD subgroups with standard attention mechanism.
- M3-Interlocutor-based Attentional LSTM:
  Using the participant's vocal LSTM with the "interlocutor-modulated attentional mechanism"

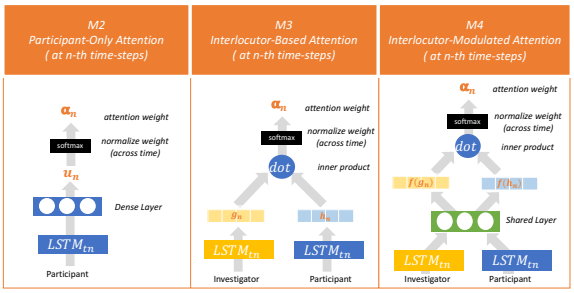

Figure 2: *It shows the three model structures (M2, M3, M4) in utilizing attention mechanisms for ASD subgroup recognition.*

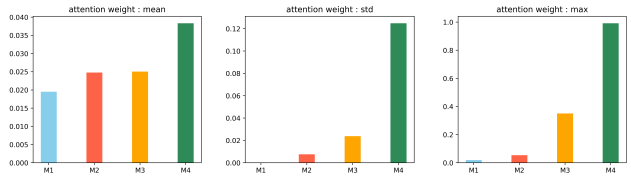

Figure 3: *Statistics of attention weights: Comparison between M1,M2, M3 and M4*

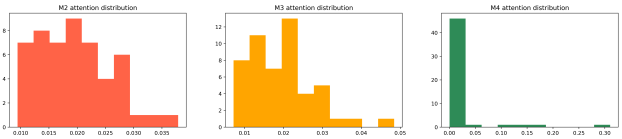

Figure 4: *Attention weights distribution: Comparison between M2, M3 and M4*

but without the shared dense layer to differentiate between the three ASD subgroups.

- M4-Interlocutor-Modulated Attentional LSTM: Using the participant's vocal LSTM with our complete proposed "interlocutor-modulated attentional mechanism" to differentiate between the three ASD subgroups.

Firgure 2 shows the three different M* models.

3.1.2. Other Experimental Parameters

The LSTM is trained with a fixed length (51 time-steps), which is the maximum number of turn-takings occurred between the investigator and the participant in our dataset; we zero-pad those sessions without 51 turn-takings. The number of hidden nodes in the LSTM is eight, and the shared dense layer in the interlocutor-modulated mechanism also has eight units. The experiment is carried out using leave-one-participant-out cross validation with the metric of unweighted average recall (UAR). We choose batch size 25, learning rate 0.01 with ADAM optimizer [32], cross-entropy as our loss function with 5 epochs of learning for our proposed network structure.

3.2. Experimental Results and Analyses

3.2.1. Analysis on Model Performance

Table 2 summarizes our complete recognition results. Our proposed IM-aLSTM obtain the best overall classification accuracies (66.5% UAR). This method outperforms the previous method by 14% absolute. The use of LSTM in time-series modeling provides improved modeling power as evident in the improvement of M1 over M0, and the attention mechanism provides yet another improvement over straightforward LSTM (M2 vs. M1).

One important observation is that by integrating dyadic interaction information in the computation of attention weights for LSTM is critical in achieving further improved recognition results (comparing between M3 M4 and M2) - reinforcing the importance in modeling the social-communicative interaction dynamics of the dyad jointly. Lastly, the final shared dense layer in the computation of "interlocutor-modulated" attention mechanism indeed better capture the subtle and complex when computing the weights, in specific our proposed method of M4 outperforms M3 by 3.7% absolute.

	AD (27)		AS (19)		HFA (11)	
	number	ratio	number	ratio	number	ratio
happy	4	0.15	1	0.05	0	0.00
fear	8	0.30	7	0.37	2	0.18
angry	**11**	**0.41**	3	0.16	**7**	**0.64**
sad	4	0.15	**8**	**0.42**	2	0.18

Table 3: *The number of maximum attention weights placed on certain emotional topics for different subgroups of ASD*

3.2.2. Analyses of Attention Weights

We first visualize the learned attention weights of M1, M2, M3, and M4 models. Figure 3 shows mean, standard deviation, and maximum of the attention weights for each model, and Figure 4 shows the distribution of the weights. Our proposed IM-aLSTM learns attention weights that are higher overall with larger standard deviation, and their distributions are also much more concentrated compared to the other models. Our results seem to indicate that the more distinct this particular pattern exhibits in the attention weights the higher the recognition accuracy.

We further analyze which topical segments within the *Emotion* part that our IM-aLSTM places the attention on. We manually segment the *Emotion* part into the four distinct emotion topical segments: happy, sad, angry, and fear. Table 3 summarizes the number of times that the maximum value of attention weights occurred within each segment for each participant. We find that regardless of ASD subgroups, the topic of "happy," which is a positive emotion, gains much less attention as compared to the more negative emotional topic, e.g., "sad," "angry," and "fear." Our result suggests that vocal characteristics of the ASD participants when discussing and revealing about their past negative emotional experiences during interaction might include unique subtle behavior differences between these three ASD groups. This observation may also be related to the findings obtained from the past psychology experiment indicating the impaired process in the self-awareness of negative emotion episodes for the ASD population [8].

4. Conclusions and Future work

The heterogeneity exists in the behavioral manifestation of ASD present challenging scenarios in understanding different important subtleties among ASD subgroups (AD, HFA, AS). In this work, we propose an IM-aLSTM framework that models the vocal behaviors in the *Emotion* part of the ADOS sessions to improve differential categorization between the three groups. Our IM-aLSTM jointly consider the dyadic interaction and embed such dynamics in our proposed interlocutor-modulated attention mechanism. Our method achieves a promising accuracy of 66.5%. Additional analyses not only provides a visualization on the learned attention weights distribution, it also demonstrates an interesting pattern that segments within the *Emotion* part of the ADOS contain scenarios more on participants being asked to discuss and talk about their past negative emotional experiences compared to positive ones.

In our immediate future work, we plan to extend our framework to include other behavior modalities, e.g., facial expressions and lexical content. By continuously engaging in inter-disciplinary collaboration with the Autism researchers, we would bring additional insights into understanding the behavioral differences between the three complexly-intertwined syndromes of ASD by developing advanced technical frameworks in modeling their expressive behavioral signals [33, 34].

5. References

[1] V. J. Derlega and J. H. Berg, *Self-disclosure: Theory, research, and therapy.* Plenum Press, 1987.

[2] C. Antaki, R. Barnes, and I. Leudar, "Self-disclosure as a situated interactional practice," *British journal of social psychology*, vol. 44, no. 2, pp. 181–199, 2005.

[3] B. A. Farber, "Patient self-disclosure: A review of the research," *Journal of Clinical Psychology*, vol. 59, no. 5, pp. 589–600, 2003.

[4] A. E. Kelly, "Helping construct desirable identities: A self-presentational view of psychotherapy." *Psychological Bulletin*, vol. 126, no. 4, p. 475, 2000.

[5] L. Capps, N. Yirmiya, and M. Sigman, "Understanding of simple and complex emotions in non-retarded children with autism," *Journal of Child Psychology and Psychiatry*, vol. 33, no. 7, pp. 1169–1182, 1992.

[6] P. Mundy, M. Sigman, J. Ungerer, and T. Sherman, "Defining the social deficits of autism: The contribution of non-verbal communication measures," *Journal of child psychology and psychiatry*, vol. 27, no. 5, pp. 657–669, 1986.

[7] E. K. Farran, A. Branson, and B. J. King, "Visual search for basic emotional expressions in autism; impaired processing of anger, fear and sadness, but a typical happy face advantage," *Research in Autism Spectrum Disorders*, vol. 5, no. 1, pp. 455–462, 2011.

[8] C. Rieffe, M. M. Terwogt, and K. Kotronopoulou, "Awareness of single and multiple emotions in high-functioning children with autism," *Journal of autism and developmental disorders*, vol. 37, no. 3, pp. 455–465, 2007.

[9] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[10] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. S. Narayanan, "Acoustic-prosodic and turn-taking features in interactions with children with neurodevelopmental disorders." in *Interspeech*, 2016, pp. 1185–1189.

[11] S. R. Leekam, S. J. Libby, L. Wing, J. Gould, and C. Taylor, "The diagnostic interview for social and communication disorders: algorithms for icd-10 childhood autism and wing and gould autistic spectrum disorder," *Journal of Child Psychology and Psychiatry*, vol. 43, no. 3, pp. 327–342, 2002.

[12] M. Mordre, B. Groholt, A. K. Knudsen, E. Sponheim, A. Mykletun, and A. M. Myhre, "Is long-term prognosis for pervasive developmental disorder not otherwise specified different from prognosis for autistic disorder? findings from a 30-year follow-up study," *Journal of autism and developmental disorders*, vol. 42, no. 6, pp. 920–928, 2012.

[13] C. Lord, E. Petkova, V. Hus, W. Gan, F. Lu, D. M. Martin, O. Ousley, L. Guy, R. Bernier, J. Gerdts *et al.*, "A multisite study of the clinical diagnosis of different autism spectrum disorders," *Archives of general psychiatry*, vol. 69, no. 3, pp. 306–313, 2012.

[14] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[15] U. Frith, "Emanuel miller lecture: Confusions and controversies about asperger syndrome," *Journal of child psychology and psychiatry*, vol. 45, no. 4, pp. 672–686, 2004.

[16] T. Bennett, P. Szatmari, S. Bryson, J. Volden, L. Zwaigenbaum, L. Vaccarella, E. Duku, and M. Boyle, "Differentiating autism and asperger syndrome on the basis of language delay or impairment," *Journal of autism and developmental disorders*, vol. 38, no. 4, pp. 616–625, 2008.

[17] C. Ecker, W. Spooren, and D. Murphy, "Developing new pharmacotherapies for autism," *Journal of internal medicine*, vol. 274, no. 4, pp. 308–320, 2013.

[18] S. Odom, K. Hume, B. Boyd, and A. Stabel, "Moving beyond the intensive behavior treatment versus eclectic dichotomy: Evidence-based and individualized programs for learners with asd," *Behavior Modification*, vol. 36, no. 3, pp. 270–297, 2012.

[19] L. Schreibman, "Intensive behavioral/psychoeducational treatments for autism: Research needs and future directions," *Journal of autism and developmental disorders*, vol. 30, no. 5, pp. 373–378, 2000.

[20] C.-P. Chen, X.-H. Tseng, S. S.-F. Gau, and C.-C. Lee, "Computing multimodal dyadic behaviors during spontaneous diagnosis interviews toward automatic categorization of autism spectrum disorder," in *Proc. Interspeech*, 2017.

[21] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 24, no. 5, pp. 659–685, 1994.

[22] P. Boersma and D. Weenink, "Praat-a system for doing phonetics by computer [computer software]," *The Netherlands: Institute of Phonetic Sciences, University of Amsterdam*, 2003.

[23] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[24] Y.-S. Lin and C.-C. Lee, "Deriving dyad-level interaction representation using interlocutors structural and expressive multimodal behavior features," *Proc. Interspeech 2017*, pp. 2366–2370, 2017.

[25] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[26] S.-W. Hsiao, H.-C. Sun, M.-C. Hsieh, M.-H. Tsai, Y. Tsao, and C.-C. Lee, "Toward automating oral presentation scoring during principal certification program using audio-video low-level behavior profiles," *IEEE Transactions on Affective Computing*, 2017.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[29] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[30] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[31] J. Gibson, D. Can, P. Georgiou, D. C. Atkins, and S. S. Narayanan, "Attention networks for modeling behaviors in addiction counseling," in *Proc. Interspeech*, 2017.

[32] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] S. Narayanan and P. 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.

[34] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. 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.