A Multimodal Interlocutor-Modulated Attentional BLSTM for Classifying Autism Subgroups During Clinical Interviews

Yun-Shao Lin, *Student Member, IEEE*, Susan Shur-Fen Gau, and Chi-Chun Lee, *Member, IEEE*

***Abstract*—The heterogeneity in Autism Spectrum Disorder (ASD) remains a challenging and unsolved issue in the current clinical practice. The behavioral differences between ASD subgroups are subtle and can be hard to be manually discerned by experts. Here, we propose a computational framework that is capable of modeling both vocal behaviors and body gestural movements of the interlocutors with their intricate dependency captured through a learnable interlocutor-modulated (IM) attention mechanism during dyadic clinical interviews of Autism Diagnostic Observation Schedule (ADOS). Specifically, our multimodal network architecture includes two modality-specific networks, a speech-IM-aBLSTM and a motion-IM-aBLSTM, that are combined in a fusion network to perform the final three ASD subgroups differentiation, i.e., Autistic Disorder (AD) vs. High-Functioning Autism (HFA) vs. Asperger Syndrome (AS). Our model uniquely introduces the IM attention mechanism to capture the non-linear behavior dependency between interlocutors, which is essential in providing improved discriminability in classifying the three subgroups. We evaluate our framework on a large ADOS collection, and we obtain a 66.8% unweighted average recall (UAR) that is 14.3% better than the previous work on the same dataset. Furthermore, based on the learned attention weights, we analyze essential behavior descriptors in differentiating subgroup pairs. We further identify the most critical self-disclosure emotion topics within the ADOS interview sessions, and it shows that *anger* and *fear* are the most informative interaction segments for observing the subtle interactive behavior differences between these three sub-types of ASD.**

***Index Terms*—Behavioral signal processing, autism spectrum disorder, multimodal BLSTM, attention mechanism.**

I. Introduction

SELF-DISCLOSURE is a dynamic and unique process that is naturally carried out in face-to-face interaction settings [1]. The process of message exchange between communicators during self-disclosure differs greatly from other general conversations. Self-disclosure involves people revealing personal private information including opinions, feelings, emotion and experiences, to another person [2], [3], which tends to induce insecure feelings for the revealing person [3], [4]. This process is, however, an effective mean for people to enhance intimacy and trust between the interacting partners [5]. In fact, the interpersonal relationship between the two communicators plays an important role in the levels of disclosed information [6]. The better-liked partner [7] and the manner of the listener's response [8] are two of the most important factors in influencing the level of disclosed information. For example, disclosure behavior in a highly intimate relationship like spouses can easily lead to an exchange of more private thoughts; in contrast, for the low intimate relationship like strangers, self-disclosure often results in sharing non-private self-descriptive information.

The self-disclosure process is also critical as a clinically-valid interaction setting carried out mostly in the psychotherapy session between the subject and the therapist. Not only is it important to *know* about the subject's story through their self-disclosure, but studies have also shown that the appropriate level of self-disclosing behavior from the therapist is effective in building trust with patients, which would lead to a higher chance of successful therapy [9], [10]. Hence, instead of considering disclosure interaction as a unidirectional process from the discloser to the recipient, this process is often considered as a reciprocal social exchange process [11], [12], i.e., both the communicators continuously adjust and adapt their behaviors according to one another during their interactions. This mutually-dependent and adaptive behavior shapes the overall disclosing process dynamically and constructs the conversation context for the on-going interaction. Past research has shown that entrainment (a.k.a. interaction synchrony) is a critical accommodation process of human communication [6], [13] especially relevant for psychotherapy. Koole *et al.* depict an Interpersonal Synchrony (In-Sync) model demonstrating that the coordination processes resulting from the basic low-level movement to the high-level coordination of language and thought are all crucial factors during psychotherapy in achieving effective treatment [14].

In this work, we concentrate on studying subjects of Autism Spectrum Disorder (ASD) during interactive clinical interview sessions, targeting the self-disclosing part of the interview. ASD is a neural developmental disorder characterized mainly by the associated socio-communicative deficit [15], [16]. Expressive symptoms such as restricted and repetitive behaviors [17], lack of eye contact [18] and poor language skills [19] often appear

Manuscript received May 6, 2019; revised October 12, 2019; accepted January 10, 2020. Date of publication January 30, 2020; date of current version April 8, 2020. The work was supported in part by Ministry of Science and Technology under Grants 109-2634-F-007-012 and 108-2634-F-007-005.

Y.-S. Lin and C.-C. Lee are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan, and also with the MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei 10617, Taiwan (e-mail: astanley18074@gmail.com; cclee@ee.nthu.edu.tw).

S. S.-F. Gau is with the Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taipei 10002, Taiwan (e-mail: gaushufe@ntu.edu.tw).

Digital Object Identifier 10.1109/JSTSP.2020.2970578

to be the distinctive abnormal characteristics of ASD subjects when they interact with others. In order to quantitatively assess the severity of their socio-communicative deficits, researchers have developed a series of instruments to elicit and measure the subject's behaviors with respect to different intended functions. Among these instruments, the Autism Diagnostic Observation Schedule (ADOS) [20] has been recognized as the gold standard instrument, which is based on a semi-structured diagnostic interview protocol, in assessing the severity of the autistic symptoms of individuals. During the administration of an ADOS, a certified clinician (investigator) plays the role of an interaction partner, who is trained to elicit the participant's behaviors and acts as an experienced observer at the same time to rate the severity. This spontaneous dyadic setting during an ADOS session provides a well-defined and verified protocol in investigating socio-communicative behavior symptoms of ASD.

Aside from the apparent deficit in socio-communicative skills, a series of emotion-related issues are commonly identified as essential features of ASD, i.e., ASD subjects are unable to correctly recognize other's emotional states [15], [21]. Specifically, compared to typically-developing subjects (TD), individuals with ASD display more frequently their negative emotions, like anger and anxiety, often with stronger intensity [22]–[24], and they also show poorer ability in understanding complex socially-derived emotion [25]. Researchers have further identified that ASD subjects have impaired mechanism in self-awaring their own negative emotional experiences in daily life [26]. This impaired mechanism in self-assessing and recalling one's own negative emotion not only concerns their emotion regulation problem [27], [28] but also leads to more maladaptive behaviors when experiencing negative emotion and displaying internal experiences expressively [29].

Emotion part during the administration of an ADOS is designed specifically to assess the emotion deficit in ASD. *Emotion* part consists of an interactive conversation session, where the investigator would use alternated questions to ask the ASD participants in order to facilitate the ASD subjects to self disclose and recall their own emotional experiences in daily life and further describe/express their feelings about it. This procedure provides not only a suitable environment to investigate emotion-related issues for clinicians but also an environment for computational researchers to develop algorithms that can automatically assess varying socio-communicative abilities of ASD due to its back-and-forth conversational nature. For example, Bone *et al.* observe an increase in subtle turn-end pitch slope and abnormal voice quality on both the investigator and the participant that is related to the severity of ASD symptoms and the acoustic-prosodic coordination is also observed between communicators during the *Emotion* part of the ADOS [30]. Moreover, Bone *et al.* extend their studies to further investigate the relationship between the levels of ASD severity and the resulting turn-taking dynamics during the *Emotion* part of the ADOS [31].

In this work, we also leverage the interactive nature of the *Emotion* part in order to compute the socio-communicative behaviors expressed by both interlocutors for the task in differentiating the three diagnostic categories of ASD. The three unique subgroups, Autistic Disorder (AD), High-Functioning Autism (HFA), and Asperger Syndrome (AS), defined by DSM-4, are included in our experiment. In the current clinical practice, many empirical pieces of evidence suggest that the differences between these three ASD subgroups are indistinguishable given the available clinical measurements [32]–[34]; hence, the newly-revised DSM-5 (Diagnostic and Statistical Manual of Mental Disorders version 5 [35]) merges these autism subgroups into a single umbrella term, ASD. It re-defines the categorical system used in DSM-4 with the dimensional model, i.e., heterogeneity exists in ASD individuals can be viewed as a combination of different deficit dimensions instead of distinct categories. However, many researchers present an opposite view on this new system, which they believe has inevitably enlarged the subgroups differences further as compared to the previous categorical system [36], [37]. Many argue that although the difference between these subgroups can be divergent and hard to observed directly given current clinical instruments, differential diagnosis remains a fundamental and a necessity in order to identify the etiology and further advance the development of a more targeted treatment for ASD [38]–[40].

In fact, a couple of recent research have shown initial empirical evidence on developing effective computational methods as an automatic version of the instrument to distinguish ASD. It computes behaviors of the interlocutors directly from the recorded signals (audio and video) and the experiment result shows that these quantitative descriptors are capable of providing discriminative power in differentiating subgroups of ASD. Specifically, recent works published by Chen *et al.* have shown supporting evidence that by computing directly the low-level descriptors on motion and vocal behaviors from both the participant and the investigator *jointly* during the *Emotion* part of the ADOS, these behavior features can be used to differentiate the three subgroups. It achieves 52 % unweighted classification accuracy, which is substantially better than the chance baseline [41], [42]. These stuedies point to the potential of deriving novel signal processing and machine learning-based methods for behavior measurements of the ASD subjects that are capable of capturing information beyond current clinically available instruments.

Our previous work has advanced Chen *et al.* research technically by introducing a network architecture of interlocutor-modulated attention network (IM-aLSTM) that learns to integrate both interlocutors' vocal information during *Emotion* part of an ADOS conversation to perform ASD subgroup classification [43]. Inspiring from the IM-aLSTM framework, in this work, we extend further the use of *interlocutor-modulated* attention mechanism, where the participant's BLSTM is learned by jointly integrating discriminative information of the dyad together, toward developing a complete multimodal (speech and gesture information) neural network architecture for differentiating between the three different groups of ASD. Specifically, our contributions in this work beyond our previous work [43] is listed below:

1) Development of a multimodal (speech acoustics and body gestures) interlocutor-modulated (IM) attention network architecture to differentiate between the three ASD subgroups;

2) Integrative embedding of the interlocutor relationship as an IM-attention mechanism to better model jointly both the interlocutor's behavior dynamics with fusion modeling of the two behavior modality; and
3) Additional analysis on the working of IM-mechanism in understanding the ASD subgroup differences during *Emotion* disclosure as a function of the two behavior modalities expressed in a dyadic interaction setting from a large scale real-world audio-video ADOS interview dataset

Our multimodal IM-aBLSTM model achieves the best unweighted average recall of 66.8% in a three subgroup categorization, which is 14.3% absolute improvement on the exact same dataset by Chen *et al.* [42], by modeling the participant's vocal behaviors and the investigator's gestural behaviors. Our further analysis shows a consistent result that the learned attention weights for both modalities are concentrated heavily in the regions where the ASD participant is being asked to reveal their own *negative* emotional experiences. It further strengthens the idea that the difference between the three subgroups may be related to the manifested behavior expressions exhibited during the interactive spoken interaction when 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 methodology. Section 3 summarizes our experimental results and discussions. Section 4 is a conclusion and future work.

II. Research Methodology

A. The ADOS Audio-Video Database

The ADOS audio-video database[1] was collected by SSG (one of the corresponding authors) at the Department of Psychiatry of the National Taiwan University Hospital (NTUH), Taipei, Taiwan. The ADOS is a semi-structured dyadic interview between a clinical investigator and an ASD participant. The procedure of ADOS usually lasts about 45 to 60 minutes. It includes a series of activities for evaluating different functions of the ASD participant, e.g., communication, social interaction, emotional experience, telling a story, etc. In this work, we utilize the *Emotion* part of the ADOS session as our analysis data. In order to include the samples with enough communicative ability, i.e., the ability to carry out a meaningful conversation, we include individual samples of ADOS administrated with Module 3 and Module 4 in our dataset; this set corresponds to subjects with relatively fluent expressive language levels and mature chronological age. Each *Emotion* part of the ADOS lasts around 5 to 7 minutes, and it includes a spontaneous conversation between the investigator and the participant. The investigator utilizes a series of semi-structured questions to ask the participant about their past emotional experiences, including four basic emotional experiences: happy, angry, fear, and sad, in their daily life. The semi-structured format of the *Emotion* part usually involves the investigator to engage the participant in a conversation as follows:

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

TABLE I
Detailed Demographics of the ASD Subgroups: The Table Includes the Information About Sample Numbers, Ages and Module of ASD Participants in Each Subgroup

Clinical Diagnosis on 3 ASD Subgroups			
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

Investigator: Do you feel the [emotion] sometimes?
Participant: [Yes, when I .....; No, I don't ....].
Investigator: What happens, when you are [emotion] ?
Participant: .....
Investigator: Can you describe the feeling of the [emotion]?
Participant: .....

The ADOS audio-video database includes audio recordings from two separate Bluetooth wireless lapel microphones (each microphone directs at an interlocutor) and video recordings using two fixed positioned high-definition cameras. Each video recording is collected with 30 fps, and the audio is collected at 44100 sampling rate per channel. Table I summarizes the detailed demographics of the participant's information in our database. We use a total of 60 ASD subjects (three different subgroups), and this dataset includes the same amount of data samples as done in the previous works [41], [43]. The diagnostic outcome is determined based on a combination of clinical diagnosis by senior child psychiatrist's clinical judgment and other relevant clinical interviews and assessments like ADOS and Autism Diagnosis Interview-Revised (ADIR) [44]. This ADOS Audio-Video Dataset is to our knowledge one of the largest clinically-valid audio-video corpuses for research.

B. Multimodal Interlocutor-Modulated Attentional BLSTM

The complete multimodal interlocutor-modulated attentional bi-directional long short term memory network (Multimodal IM-aBLSTM) structure is presented in Figure 1. The overall architecture is composed of three different sub-networks, including a speech-IM-aBLSTM, a motion-IM-aBLSTM, and a final fusion network, each designed with a different purpose. With the characteristics of the question-answer pattern in the ADOS *Emotion* part, the choice of a BLSTMs time step is at every *turn*. We define the turn boundary as a complete speech portion of a participant before the speaking floor changes to the investigator (and vice versa), and this particular definition of turn-taking events has also been used in the previous study on the same dataset.

Both speech-IM-aBLSTM and motion-IM-aBLSTM include *interlocutor-modulated* attention mechanism that learns to integrate the interlocutors' behavior information jointly as attention that reweights the turn-level behavior representations to perform the three subgroup classifications. For the speech part (speech-IM-aBLSTM), we input the vocal features from both the investigator and the participant to learn the attention weight and reweight the participants' vocal behavior to perform the recognition. For the motion part (motion-IM-aBLSTM), based on the gestural features derived from the tracked body joints

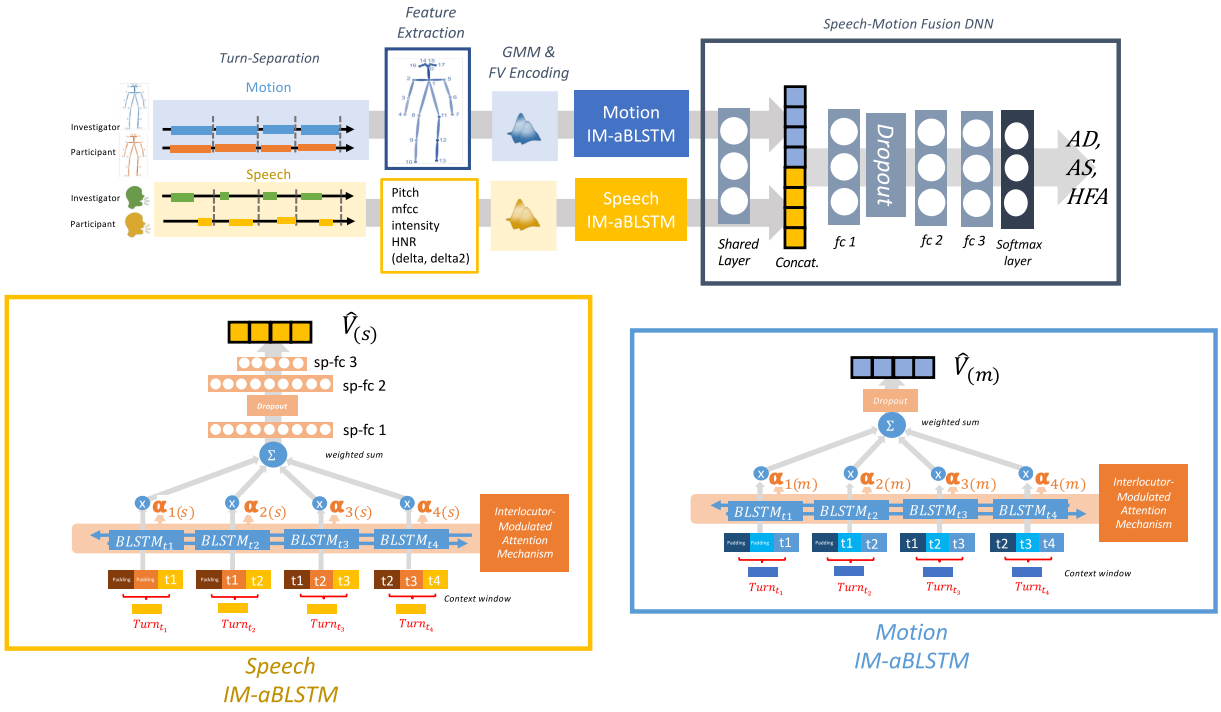

Fig. 1. The Framework of Multimodal Interlocutor-Modulated Attentional BLSTM (Multimodal IM-aBLSTM): The learnable weight pair α_F and α_B within the BLSTM are termed as the interlocutor-modulated attention (IM) that integrates dyad's information to improve the discriminability in differentiating between different subgroups.

of each frame, we input these motion features from both the interlocutors to learn the attention network and reweight the investigator's motion feature to perform the prediction. In order to leverage both discriminative information from both behavior modalities, we design a third fusion network to combine the information from the two different modalities to make the final classification. In the following sections, we will elaborate on the extraction of acoustic features, the extraction of motion features, turn level behavior representations, and the IM-attention mechanism used in the BLSTM.

1) Acoustic Features Extraction: The turn-taking event we defined in this paper is labeled manually. First, we segment the *Emotion* part into multiple turn-taking event regions. In order to segment a simple question-answer pattern, we disregard the back channels during the conversation. Each of the turn-taking events is made of 2 turns, a complete floor exchange in the form of "the investigator - the participant." In specific, the "i-th" turn-taking event starts from the start time of investigator's turn $t_{(i)inv\_start}$ to the ending time of participant's turn $t_{(i)part\_end}$. The definition ensures the value of turn-taking event with the following order, $t_{(i)inv\_start} < t_{(i)inv\_end} < t_{(i)part\_start} < t_{(i)part\_end}$.

For the speech turn-level feature, we only extract the speech feature in the non-silence part of the speaker, i.e., the investigator's features are extracted from the corresponding speaker's portion starting from $t_{(i)inv\_start}$ to $t_{(i)inv\_end}$. Similarly, the participant's feature is extracted from $t_{(i)part\_start}$ to $t_{(i)part\_end}$. 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 by using the Praat toolkit [45]. Totally, 48 dimensions of acoustic features, including pitch, intensity, MFCC and HNR, are all extracted at a framerate of 10 ms; these LLDs are z-normalized with respect to each speaker.

2) Motion Features Extraction: In the previous work [46], [47], the deficit motion perception during the social interaction has been considered as the general symptom of an ASD subject. Furthermore, the other study [48] also shows different levels of perception between ASD subgroups, like HFA and AS. Based on the understanding of the motion perception deficit, we further extract the interlocutor's body movement during the ADOS interaction. For motion feature, since we want to capture the complete expressive motion behavior within each of the turn-taking events, we consider the same interval from $t_{(i)inv\_start}$ to $t_{(i)par\_end}$ for both participant and investigator to extract the feature. For each speaker, we extract *frame*-level body joint angle as features by utilizing the state-of-the-art body joint extractor, the Openpose toolkit [49], [50]. It computes 2D body joint positions for every frame in the original 30 fps video sequence. We remove frames with low confidence detection accuracy. Furthermore, in order to eliminate the variability of body types for different people, we compute the body joints angle as the most basic unit of measurements. We have a total of 8 different angles of body joints according to the 10 different body joints tracked on the upper body. Figure 2 shows the location of the 10 different points. Besides computing the original 8 angles, which represents the static gestures of the body, we further extract delta and delta-delta to capture the dynamic of these body gestural changes. To sum up, we extract 24-dimensional

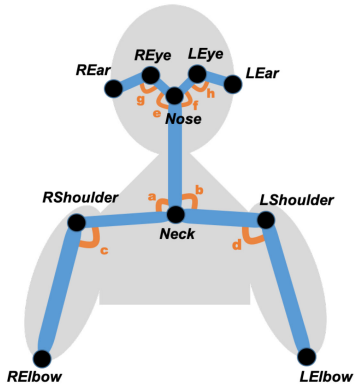

Fig. 2. 8 Angles of Body Joints: we have 8 different angles of body joints from a to h. Besides, we calculate the delta (first-order difference) and delta-delta (second-order difference) on each angle per frame. Totally, there are 24 different low-level motion features to measure the behavior of the upper body.

motion LLD features to capture the interlocutor's body language in each frame.

Since the real coordination of body joints should be a 3-dimensional vector, the body joint we extracted is a 2D projection vector of the real position without the depth information. Because the recording environment is set with the fixed position camera and fixed position desk, we neglect the slight difference between different samples and compute the angle between the body joint directly as our motion feature. Specifically, three different parts of the upper body movement are captured by these angles. Instead of considering as facial expression movement, the angle of e, f, g and h represent the head movement like looking up and down. The angles like a and b correlate to the movement of the neck, and angles c and d represent the movement of the arm.

3) Turn-Level Behavior Representation: In order to derive a vector characterizing each of the turns consisting of varying length LLD sequences, we encode the temporal sequence of behavior LLDs to a single fixed dimensional vector of turn-level representation using a method of Gaussian Mixture Model (GMM, λ) based Fisher encoding [51]. The method first constructs an unsupervised background GMM learned from the LLDs in the whole training set, and in order to encode a data sample consists of a sequence of LLDs, we compute the gradient log-likelihood function (fisher scoring function) to this GMM. This fisher scoring function (indicating the direction of λ to better fit $\bar{x}$) of the first and second-order statistics is derived below:

$$g^X_{\mu_c} = \frac{1}{T\sqrt{\pi_c}} \sum_{i=1}^{T} r_t(c) \left(\frac{x_t - \mu_c}{\sigma_c} \right) \tag{1}$$

$$g^X_{\sigma_c} = \frac{1}{T\sqrt{2\pi_c}} \sum_{i=1}^{T} r_t(c) \left(\frac{(x_t - \mu_c)^2}{\sigma_c^2} - 1 \right) \tag{2}$$

where T is the total frame number, and $r_t(c)$ is the posterior probability given the observation x_t produced by the c-th Gaussian with mean μ_c and standard deviation σ_c. This derived encoded vector of $[g^X_{\mu_c}\ g^X_{\sigma_c}]$ is our turn-level features.

The GMM-fisher encoding method was first introduced in the image classification task [51]. Recently, it also has shown to be a powerful representation for speech-related recognition tasks, e.g., detection of emotion [52], paralinguistic attributes [53], and evaluation of impromptu speech [54]. Due to its wide usage in providing a powerful representation on frame-level descriptors, we adopt this method for obtaining the turn-level representation as input for both the motion-IM-aBLSTM and the speech-IM-aBLSTM. We empirically set mixture numbers as four in the training of our background GMM for both speech and motion modality.

4) Interlocutor-Modulated Attention Mechanism: Here, for both the motion and speech modality, we utilize the bi-directional Long Short Term Memory (BLSTM) neural network [55] to model the time-dependent relation between the turn-level feature sequences. As an improved version of RNN, the inclusion of the gating mechanism in LSTM can learn the time-dependent sequential information using the transmitted temporal gradient mitigating gradient vanishing or gradient exploding problem. Moreover, BLSTM consist of 2 different directions of LSTMs [56], which are the forward LSTM and the backward LSTM to further improve the modeling capacity.

In order to capture a particular speaker's time-dependent behavior progress during the overall interaction, we build the investigator's (or participant's) BLSTM only with the investigator's (or participant's) turns. As our experiment results show, we found that speech IM-aBLSTM with participant's speech feature can achieve better results than using the participant's speech features. In contrast, we found that motion IM-aBLSTM with an investigator's motion feature can achieve a better result than using the investigator's speech features. Therefore, in speech IM-aBLSTM, we input the "i-th" participant's turn-level vocal feature sequence f(s) (Section II-B3) to obtain a corresponding output sequence of speech BLSTM's hidden states, h_s:

$$\{h_{1(s)}, \ldots, h_{T(s)}\} = BLSTM_{\text{part}}(\{f_{1(s)}, \ldots, f_{T(s)}\})$$

Similarly, in motion IM-aBLSTM, the turn-level motion feature sequence $\mathbf{f_{(m)}}$ from investigator can also be transformed to hidden states sequence $\mathbf{h_{(m)}}$ using a motion BLSTM:

$$\{h_{1(m)}, \ldots, h_{T(m)}\} = BLSTM_{\text{invt}}(\{f_{1(m)}, \ldots, f_{T(m)}\})$$

Moreover, we incorporate the use of attention mechanisms [57] into our BLSTM time series modeling. The attention mechanism has been considered as a general soft-selecting neural network structure that can automatically learn to increase the weighting on important parts and decrease the effect on non-important parts of the feature sequence during the network learning procedure. It is known to obtain improved performance in multiple recognition tasks, e.g., motion recognition [58], emotion recognition [59], prominent counselor and client behaviors during addiction counseling [60], etc. In our work, we also utilize the attention mechanism in order to leverage the controlling mechanism of attention to emphasize the interaction segment between dyads. Based on the idea that interlocutors would demonstrate synchronized entrainment and mutually dependent behaviors in spontaneous dialogs, we propose a novel

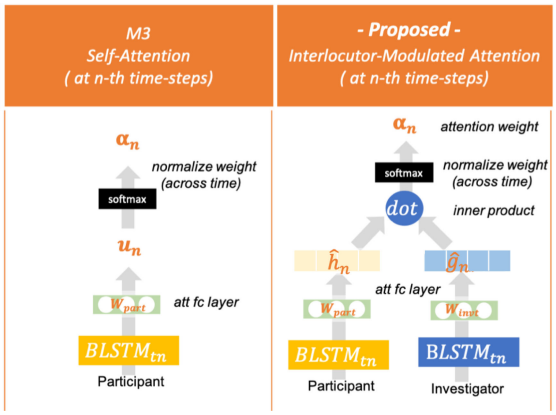

Fig. 3. Attention Network Architecture Differences between M3 and our proposed models: We improved the conventional attention mechanism in M3 by jointly considering the interlocutors information to construct the IM-attention mechanism in our proposed IM-aBLSTM.

interlocutor-modulated attention weights, α_i, as the extension and modification to the conventional attention weights.

The *interlocutor-modulated* attention weights are introduced with an intent to capture the time-dependent interactive relationship between the interlocutors (the architecture of this attention mechanism is shown in Figure 3). We first additionally train an investigator's speech BLSTM and a motion BLSTM using the investigator's turn-level acoustic features $\mathbf{f_{(m)}}$ and participant's turn-level motion features $\mathbf{f_{(m)}}$. Then, the hidden state sequences of $\mathbf{g_{(s)}}$ and $\mathbf{g_{(s)}}$ are derived for the investigator:

$$\{g_{1(s)}, \ldots, g_{T(s)}\} = BLSTM_{\text{invt}}(\{f_{1(s)}, \ldots, f_{T(s)}\})$$

$$\{g_{1(m)}, \ldots, g_{T(m)}\} = BLSTM_{\text{part}}(\{f_{1(m)}, \ldots, f_{T(m)}\})$$

In order to learn the non-linear relationship between these hidden state sequence of $\mathbf{g_{(s)}}$ and $\mathbf{g_{(s)}}$. We add a separated fully-connected (fc) layer to these two hidden states:

$$\hat{h}_{t(s)} = ReLU\left(W_{invt(s)} h_{t(s)}\right) \tag{3}$$

$$\hat{g}_{t(s)} = ReLU\left(W_{part(s)} g_{t(s)}\right) \tag{4}$$

The same procedure with the motion part is employed to obtain the sequence of $\hat{\mathbf{h}}_{\mathbf{(m)}}$ and $\hat{\mathbf{g}}_{\mathbf{(m)}}$ with parameters $W_{invt(m)}$ and $W_{part(m)}$. After passing the separated fc layer with the parameters W_{invt} and W_{part}, we then compute the similarity score as a dot product between the two interlocutors' hidden states as our attention weight u_t for the"t-th" time-step using:

$$u_{t(s)} = \left\langle \hat{h}_{t(s)}, \hat{g}_{t(s)} \right\rangle \tag{5}$$

Similar to the conventional attention weight, we normalize our IM-attention weights across time to obtain α_t for speech IM-aBLSTM:

$$\alpha_{t(s)} = \frac{exp\left(u_{t(s)}\right)}{\sum_t^T exp\left(u_{t(s)}\right)} \tag{6}$$

These *interlocutor-modulated* attention weights are combined to the participant's BLSTM's hidden vectors $h_{t(s)}$ using the following equation:

$$v_{(s)} = \sum_t^T \alpha_{t(s)} h_{t(s)} \tag{7}$$

For the final representation $\hat{V}_{(s)}$ in speech IM-aBLSTM, we pass the $v_{(s)}$ through 3 fully-connected (fc) layers and 1 dropout layers as Figure 1 shows.

Similar to the speech IM-attention mechanism, we carry out the equation (3,4) with parameter $W_{invt(m)}$ and W_{part} and following (5,6) to obtain the $\alpha_{t(m)}$ for motion IM-aBLSTM and apply the attention weights to the investigator's motion BLSTM's hidden vectors $h_{t(m)}$ using the equation (8).

$$v_{(m)} = \sum_t^T \alpha_{t(m)} h_{t(m)} \tag{8}$$

For the final representation $\hat{V}_{(m)}$ in motion IM-aBLSTM, we pass the $v_{(m)}$ through a dropout layer as Figure 1 shows.

5) Speech-Motion Fusion DNN: In order to merge the discriminative information from the two different modalites BLSTMs, we construct a Speech-Motion Fusion DNN structure. The architecture is presented in Figure 1. First, the final representation $\hat{V}_{(s)}$ and $\hat{V}_{(m)}$ for every dyad can be used to predict the probability of being in one of the ASD subgroups.

$$y_{(s)} = softmax(\hat{V}_{(s)}) \tag{9}$$

$$y_{(m)} = softmax(\hat{V}_{(m)}) \tag{10}$$

Therefore, we first pre-train the speech-IM-aBLSTM and the motion-IM-aBLSTM separately by using the cross-entropy loss $Loss_{(s)}$ and $Loss_{(m)}$.

$$Loss_{(s)} = \sum_n^N \sum_k^K -Y_{true}^k \log(y_{(s)}^k) \tag{11}$$

$$Loss_{(m)} = \sum_n^N \sum_k^K -Y_{true}^k \log(y_{(m)}^k) \tag{12}$$

After the pre-training, we freeze the two separated structures and take the $\hat{V}_{(s)}$ and $\hat{V}_{(m)}$ output from networks as the input for Speech-Motion Fusion DNN. The final recognition result of the three groups of ASD $Y_{(fusion)}$ can be derived after training the fusion network with $Loss_{(fusion)}$ listed below:

$$Y_{(fusion)} = FusionNet(\hat{V}_{(s)}, \hat{V}_{(m)}) \tag{13}$$

$$Loss_{(fusion)} = \sum_n^N \sum_k^K -Y_{true}^k \log(Y_{(fusion)}^k) \tag{14}$$

III. Experimental Setup and Results

A. Experimental Setup

1) Experiment 1. Models Comparison: We compare our proposed interlocutor-modulated attention mechanism with four different models in the task of differentiating between the three ASD subgroups: AD, AS, and HFA using either motion or speech modality.

TABLE II
EXP 1: MODEL COMPARISON (WE COMPARE THE PERFORMANCES OBTAINED BETWEEN M1, M2, M3 AND OUR PROPOSED IM-ABLSTM. WE ALSO EXAMINE THE MODEL WITH DIFFERENT CONTEXT WINDOWS SIZE FROM 0 TO 4 FOR BOTH SPEECH MODALITY AND MOTION MODALITY). THE RESULT WITH (.)* IS THE HIGHEST RESULT IN EACH MODALITY

	Speech							
	Investigator				Participant			
context_num	M1 (SVM)	M2 (mean-BLSTM)	M3 (self-aBLSM)	Proposed (IM-aBLTM)	M1 (SVM)	M2 (mean-BLSTM)	M3 (self-aBLSTM)	Proposed (IM-aBLSTM)
0	0.350	0.376	0.360	**0.418**	0.380	0.509	0.552	0.525
1		0.383	**0.414**	0.375		0.586	0.537	0.540
2		0.325	0.337	0.366		**0.592**	0.556	0.595
3		0.362	0.324	0.324		0.557	0.548	0.594
4		**0.418**	0.353	0.365		0.567	**0.588**	**0.633***

	Video							
	Investigator				Participant			
context_num	M1 (SVM)	M2 (mean-BLSTM)	M3 (self-aBLSM)	Proposed (IM-aBLTM)	M1 (SVM)	M2 (mean-BLSTM)	M3 (self-aBLSTM)	Proposed (IM-aBLSTM)
0	0.402	0.507	0.451	0.444	0.337	**0.406**	**0.411**	0.385
1		0.434	0.422	0.433		0.381	0.396	**0.424**
2		**0.496**	0.435	**0.534***		0.391	0.344	0.398
3		0.421	0.45	0.494		0.396	0.357	0.355
4		0.428	**0.466**	0.471		0.334	0.329	0.376

- *M1-SVM*: Instead of using BLSTM to model the progression of either one of the interlocutor's behavior, we use statistical functional method to characterize the overall progression and train a SVM model to differentiate between the three ASD subgroups. Here, we use 9 different statistical functional encodings, include mean, std, max, min, median, 25th-percentile, 75th-percentile, 1st-percentile and 99th-percentile.
- *M2-Mean-Pooling BLSTM*:
  Without integrating interlocutors' information, we use either one of the interlocutor's vocal/motion BLSTM with mean pooling to differentiate between the three ASD subgroups
- *M3-Self-Attentional BLSTM*:
  Without integrating interlocutors' information, we use the participant's vocal/motion BLSTM with self-attention learning mechanism to differentiate between the three ASD subgroups.
- *Interlocutor-Modulated Attentional BLSTM*:
  Considering both interlocutors' behaviors to emphasize the important segments, we use either one of the interlocutor's vocal/motion BLSTM with our proposed "interlocutor-modulated attentional mechanism" to differentiate between the three ASD subgroups.

We further consider the effect of using context windows in expanding our turn-level representations. We experiment context window of size from 0 to 4 in experiment 1. For example, the context window size "n" in Table II means, for the input at t-th timestep, we concatenate the original turn-level representation from the timestep t-n to timestep t. Figure 3 shows the detail attention mechanism architecture of M3 and our proposed models.

2) Experiment 2. Multimodal Fusion: We evaluate different multimodal fusion methods on the task of differentiating the three ASD subgroups.

- Baseline: The multimodal method previously proposed by Chen *et al.* [42] to perform recognition by computing dyadic low-level behavior descriptors on the same dataset
- M1-Decision Fusion: In experiment 1, both the Motion M1 and Speech M1 can output the decision score on each of the subgroup class based on the trained SVM model. Under the same experiment setting, we add the output decision score between the two models and perform fusion recognition. We consider different fusion pairs between the two modality, including "part(sp)+invt(m)," "part(sp)+part(m)", "invt(sp)+part(m)" under this experiment setting.
- Multimodal IM-aBLSTM (proposed):
  Our Multimodal IM-aBLSTM is composed of three different sub-nets, including a Speech-IM-aBLSTM, a Motion-IM-aBLSTM, and a Speech-Motion Fusion DNN. Here, we freeze the best performing Speech-IM-aBLSTM and Motion-IM-aBLSTM according to the result in experiment 1 and train the Speech-Motion Fusion DNN by inputting the 2nd layer prior to softmax from both model. We also consider all different fusion pairs between the two modality, including "part(sp)+invt(m)," "part(sp)+part(m)", "invt(sp)+part(m)" and "invt(sp)+invt(m)" under this experiment setting.

3) Other Experimental Parameters: In this work, we perform 5-fold cross-validation for both GMM Fisher Vector encoding and IM-aBLSTM on our dataset to evaluate the model's performance, i.e., there are 48 training samples and 12 testing sample for every cross-validation fold. The final testing result is derived by accumulating the prediction result on overall testing data in each of the folds. In other words, in order to prevent issue of overfitting, we look for a general well-performed parameter setting across 5 different cv folds. Under this setting, testing data would be always guaranteed to prevent from involving in the training process at any stage. The fusion result is also examined in this experimental setting. We freeze the model in each CV

TABLE III
EXP 2: MULTIMODAL FUSION RESULTS (WE COMPARE DIFFERENT FUSION MODEL BETWEEN M0, M1 AND OUR IM-ABLSTM. DIFFERENT PAIRS OF BEHAVIOR COMPOSITION ARE EXAMINED TO ACHIEVE THE HIGHEST PERFORMANCE)

multimodal fusion result		
fusion model	type	UAR
M0 baseline	multimodal	**0.540**
M1-SVM Decision Fusion	invt(sp)+invt(m)	**0.436**
	part(sp)+part(m)	0.339
	part(sp)+invt(m)	**0.436**
	invt(sp)+part(m)	0.339
Mutlimodal-IM-aBLSTM (proposed)	invt(sp)+invt(m)	0.431
	part(sp)+part(m)	0.600
	part(sp)+inv(m)	**0.668***
	invt(sp)+part(m)	0.436

TABLE IV
PERFORMANCE ON EACH OF CV FOLDS: UNDER THE 5-FOLD CROSS-VALIDATION SCHEME, WE REPORT THE RESULT OF THE UAR ON THE TESTING SET IN EACH FOLD. WE PRESENT THE SPEECH IM-ABLSTM(PART) WITH UAR 0.633, MOTION IM-ABLSTM(INVT) WITH UAR 0.534 AND MULTIMODAL FUSION RESULT WITH UAR 0.688

cv fold	Speech IM-aBLSTM	Motion IM-aBLSTM	Fusion
0	0.611	0.861	0.778
1	0.528	0.583	0.528
2	0.700	0.433	0.633
3	0.5	0.389	0.556
4	0.458	0.347	0.500
All	0.633	0.534	0.668

loop and redo the inner CV loop for the fusion model. Therefore, in this fashion, Table IV shows the testing result on 5 different folds by using 3 different models, and testing data in each fold would always be guaranteed to prevent from involving in the training process. We use the unweighted average recall (UAR) as our evaluation metric. Compared to conventional accuracy, the UAR metric is a better metric for evaluating the performance of the model on imbalanced data. For example, the baseline performance for random guessing to the majority class (AD) would lead to an UAR of 0.333.

The BLSTM is trained with a fixed length (51 time-steps), which is the maximum number of turn-takings that occurred between the investigator and the participant in our dataset. We zero-pad those sessions with fewer than 51 turn-takings events. For the Speech-IM-aBLSTM, the number of hidden nodes in the BLSTM is 128, 64 nodes for the forward LSTM and 64 nodes for the backward, the att-fc layers in IM-attention mechanism have 128 nodes and the sp-fc1, sp-fc2 and sp-fc3 have 64, 16,16 nodes. For the Speech-IM-aBLSTM, the number of hidden nodes in the BLSTM is 16, 8 nodes for the forward LSTM and 8 nodes for backward, the att-fc layers in IM-attention mechanism also have 16 nodes. In the training procedure, we set the dropout ratio to 0.1 in both the speech and the motion model. We choose batch size 5, learning rate 0.01 with Adamax optimizer [61], cross-entropy is used as our loss function with 30 epochs when learning our proposed network structure. For each model, we examine 20 different random seeds and present the highest result in Table II. With the same setting, we freeze the modality in each CV loop and redo the inner CV loop for the fusion model. We use Pytorch [62] toolkit to build our network.

As an extension to our previous work [43], we clarify the differences of experimental detail between these two papers. First, we use the 5-fold cross-validation in this paper instead of using the leave-one-out scheme in the previous one to adjust the model's parameters. Under the 5-fold cross-validation, we completely leave out the testing set and only use the training data to build GMM and fisher vector encoding in this paper. In other words, we modify the original encoding method [43], in which we perform the unsupervised GMM on the whole dataset. By using the nested cross-validation pipeline, our results are less prone to issue of overfitting. Third, we have adjusted IM-attention architecture and utilize the context window in this paper. Therefore, the performance in previous work can not be directly compared with the performance in this paper.

B. Experiment Result and Analysis

1) Analysis on Model Performance: Our proposed multimodal IM-aBLSTM achieves overall the best performance at 66.8% UAR on the three ASD subgroup classification task. It outperforms the previous multimodal method presented by Chen *et al.* by 14.3%. Furthermore, it also outperforms the simple M1-SVM statical functional encoding method by 22.2%. The fusion between different modality and jointly considering both interlocutor's behavior information are both needed to achieve improved performance in all experimental settings. Under the M1 experiment setting, the M1-fusion result in Table III shows that the fusion model can improve the performance by 3.6 % when compared to the results of using only the investigator's motion M1-SVM in table II, and it achieves an overall 43.6% UAR. Similarly, with our proposed experimental setting, the speech-motion fusion DNN in the proposed multimodal interlocutor also demonstrates improved classification rates by combining the two different modalities and improves 3.5% when compared to use only the speech-IM-aBLSTM.

There are three important findings based on the result in experiment 1. First, the participant's vocal behavior can provide more discriminative information than the investigator's vocal behaviors for speech modality. This result, however, is not the same between the two different modalities. In fact, we observe an opposite result in the motion modality that the investigator's gestural behaviors contain more discriminative information about the three subgroups when compared to the participant's motion behaviors. Specifically, the participant's speech IM-aBLSATM can achieve overall the best performance at 63.3 % UAR and outperforms the investigator's speech IM-aBLSTM by 21.5 %. The investigator's motion IM-aBLSATM has the highest performance at 53.4 % UAR and outperforms the participant's motion IM-aBLSTM by 11%.

Second, we generally observe that the information in context windows is required for both speech and motion modality to achieve higher performance. In specific, the participant's speech IM-aBLSATM needs a longer context window size of 4, and investigator's motion IM-aBLSTM uses a shorter context window size of 3.

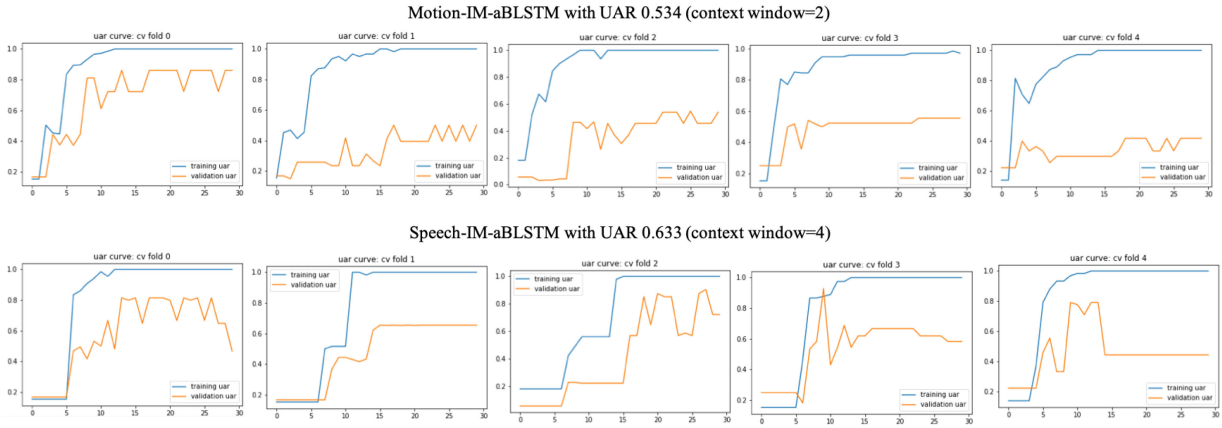

Fig. 4. Learning Curve on Each of CV Folds: Under the 5-fold cross-validation scheme, we present the UAR curve during the training procedure (30 epochs) in each fold. The two proposed models are presented: the Speech IM-aBLSTM(part) with UAR 0.633 and the Motion IM-aBLSTM(invt) with UAR 0.534.

Third, when comparing M2, M3 and our model, the result of ours shows that integrating interaction relationship between dyad using our proposed IM-attention mechanism, which reweights the overall time progression, is crucial for improving the overall classification performance in both speech and motion modality. However, it is interesting to see that M2 mean-pooling BLSTM model sometimes performs better than M3 self-attentional BLSTM model. It suggests that it is the *interactive dependent* phenomenon between interlocutors that is important in deriving the attention weights, and our proposed use of the IM-attention mechanism in our structure is demonstrated to be effective in learning the appropriate attention weight to improve the performance comparing to the conventional attention mechanism. In specific, when using the participant's vocal behavior as input in speech modality, our model outperforms the M2 by 4.1%, and it also outperforms the M3 by 4.5%. For using the investigator's motion behavior features as input in motion modality, our model outperforms the M2 by 2.7%, and it also outperforms the M3 by 6.8%.

2) Analysis of Model Training: In this part, we evaluate the model training detail of our proposed Speech IM-aBLSTM(part) and the proposed Motion IM-aBLSTM(invt). In Figure 4, we can see that both of the models generally converge around 20 epochs, where the training data can approach an UAR close to 1.0 with the testing data achieves around $0.5 \sim 0.6$. In Table IV, we present the corresponding testing result in different cv folds. We can observe that the cv folds 3 and 4 have the worst performance among all the training folds for both modalities. However, both of the cv folds can be improved with multimodal fusion. Although the Motion-aBLSTM only achieve UAR at 0.534, we also observe its complementary nature that helps with the improvement for cv fold 0, 3 and 4.

Due to the fact that the performance of the neural network model would change with different initialization and different batch-wised update, we test with 20 different random seeds. In these experiments, our Speech IM-aBLSTM results in a std. of 0.068, the Motion IM-aBLSTM has a std. of 0.059 and the Multimodal IM-aBLSTM's std. is 0.079. While being relatively stable, we think that the stability can be further improved in our future work by using the data augmentation method like dropping the utterance with a certain probability and using the domain adaptation method on other interaction part in ADOS.

3) Analysis of IM-Attention Mechanism: In Figure 4, we show some examples of our learned IM-attention weight on selected data samples. Based on the attention weights on different time steps, we can observe that attention weights in most of the samples are highly concentrated. Since the interaction process is not easy to discern manually, our attention mechanism provides a possible interpretation by highlighting the important part with high discriminative information. Therefore, we further analyze whether the speech and the motion LLD features on these highest attention segments within the *Emotion* part of the ADOS would behave differently among different ASD subgroups. By zooming in on these learned high IM-attention regions, we can safely assume that the LLD features in the selected turns would contain enough discriminative information between subgroups.

We calculate the functional value, including avg, std, max and min, on frame-level LLD features in order to provide an intuitive insight for understanding the behavior during the selected high attention turns. We examine the value and investigate the distribution differences between groups. Specifically, two-sample t-tests are used to investigate whether there are any significant differences in the functional values between three different pairs of subgroups, i.e., AD-AS, AS-HFA and AD-HFA. We list the features resulting in statistically significant differences (p-value below 0.05) between the pair of the groups in Table IV. For speech modality, we observe that the participant's "average" and "median" value of pitch on the selected turn shows significantly different between HFA-AD and HFA-AS. In other words, the subjects of HFA tend to speak with a higher value of pitch compared to either AS or AD.

In terms of motion features, we find that different angles of body joints correlate to the differences between ASD subgroups. Most of the head orientation-based LLD features, including b-delta, b-delta2, e and h, all demonstrate significant differences between the AD-HFA pair. In specifics, the wider range and faster head movement from the investigator would indicate that he/she is interacting with subjects of the AD. Furthermore, we

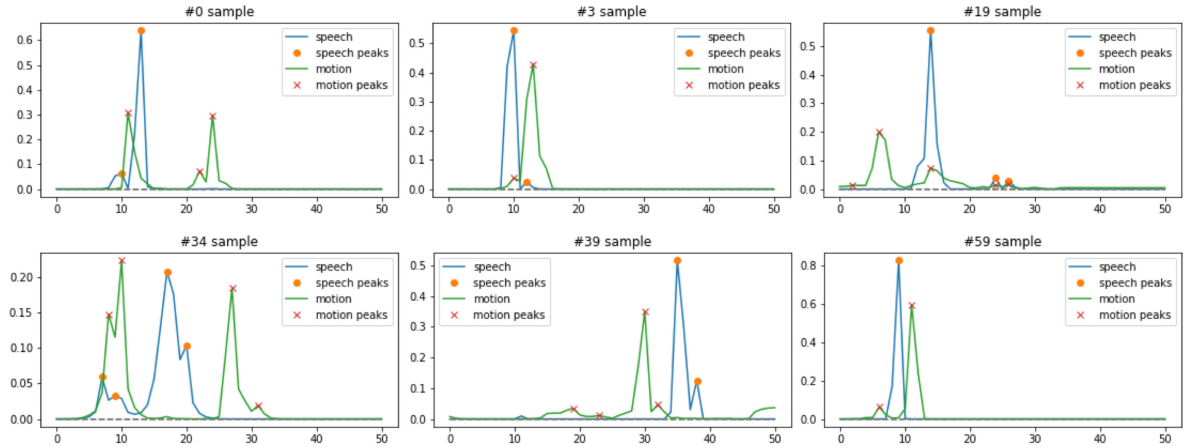

Fig. 5. Attention Weights Visualization on Selected Test Samples: We plot 2 different aspects of attention weight distribution in the same session. The marks of the peaks indicate the local maximum of high similarity behavior measured by our automatically learned network. Based on the figure, our attention weight shows a peaky characteristic. We further examine the behavior feature at the highest attention (the peak) between different subgroups in Section III.B.3 and explore important emotion topic using WCI metric introduced in section III.B.4. Besides, we also observe the coherence between 2 different networks in the figure. A further complete analysis is done by using the confusion matrix in Figure 6.

also observe that the arm movement related LLD features like d, d-delta, d-delta2 show significant differences when comparing HFA with the other two subgroups. This result can be interpreted as a situation that the investigator uses hand as body language very often during his/her conversation or making notes during the assessment. The analysis results about the motion features are quite interesting, it indicates that while the vocal characteristics of the ASD subjects intuitively show the types of ASD symptoms, it is the investigator (the one who interacts with the ASD subject) would display gestural differences indicating the particular ASD subgroup - further underscoring the importance in analyzing the social back-and-forth aspect of an ASD subject with his/her interacting partners.

Furthermore, we analyze the attention weight distribution between the two modalities. We first identify the peaks of the learned attention weight of each modality for each sample. We would like to examine whether each of the behavior modalities is maximum within the same conversation segment. We define a conversation segment according to the 4 emotion topics that the investigator asks the participant to talk about. The nearest peaks pair can be analyzed with totally 16 pairs on the topic matrix, i.e., if the nearest speech attention peak is at the emotion topic "Angry" and the nearest motion attention peak is at "Sad," then the "Angry-Sad" is the nearest pair for the data sample. We accumulate the sample of all pairs and present the result as a confusion matrix shown in Figure 5. From the result, we observe that most of the data samples have the nearest peaks between speech and motion occurring in the same topic, i.e., indicating that these two modalities' specific "hot spots" tend to come from the *same* conversation segments.

4) Analysis on the Emotion With Attention Centroid: In order to identify the important segments within the *Emotion* part of the ADOS that are useful in differentiating the subgroups, we calculate the attention centroid by weighting the turn index with the attention weight and identify which emotion topic is important according to this weighted centroid index(WCI)

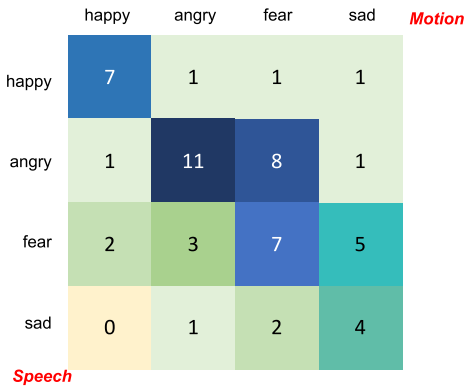

Fig. 6. Emotion Topic Analysis: The Position of the Nearest Pair of Peaks between Two Modalities (We display the speech-motion emotion topic pair with a confusion matrix. There are 52.7% of the data lying on the diagonal of the topic matrix showing that most of the nearest peaks occur in the same topic).

for each sample. By introducing the WCI, we define it as an index to highlight the important emotion subpart during the interaction. In specific, with the total timesteps $T = 51$ and the stored attention value α_t, our WCI can be derived from

$$WCI = \sum_{t=1}^{T} \alpha_t * t$$

Since $\sum_{t=1}^{T} \alpha_t = 1$, the value of WCI would be guaranteed to be 1 to 51. The highlighted emotion part can then be identified from the position of WCI with the recorded time. We present corresponding highlighted emotion in Table V and Table VI by using speech WCI and motion WCI separately. In each table, we report the accumulated number of samples for each emotion topic. Based on the result of speech-IM-aBLSTM, we observe that the less important part for participant's vocal behavior is "happy," and the most important part is "angry" in differentiating the three subgroups. For the motion-IM-aBLSTM, the model

TABLE V
STATISTICAL SIGNIFICANCE LLD FEATURES BETWEEN SUBGROUPS: WE PERFORM TWO-SAMPLE T-TESTS ON THE STATISTICAL FUNCTIONAL FEATURE BETWEEN PAIRS OF SUBGROUPS. THIS TABLE LISTS THE LLD FEATURES WITH $p \leq 0.05$

Investigator's Motion LLD - turn with max attention				
feature	functional	subgroups	p-value	large
b-delta	avg	AD-HFA	0.029	AD
	std	AD-HFA	0.023	AD
	max	AD-HFA	0.016	AD
	median	AD-HFA	0.046	AD
b-delta2	avg	AD-HFA	0.022	AD
	std	AD-HFA	0.019	AD
	max	AD-HFA	0.023	AD
	median	AD-HFA	0.024	AD
c	max	AD-AS	0.031	AS
		AS-HFA	0.050	HFA
c-delta	median	AD-HFA	0.028	AD
c-delta2	median	AD-HFA	0.016	AD
d	std	AS-HFA	0.032	HFA
d-delta	avg	AS-HFA	0.017	HFA
		AD-HFA	0.040	HFA
	std	AS-HFA	0.016	HFA
		AD-HFA	0.048	HFA
	max	AS-HFA	0.010	HFA
		AD-HFA	0.015	HFA
	median	AD-HFA	0.049	AD
d-delta2	avg	AS-HFA	0.029	HFA
	std	AS-HFA	0.020	HFA
	max	AS-HFA	0.011	HFA
		AD-HFA	0.015	HFA
e	std	AD-HFA	0.035	AD
h	std	AD-HFA	0.025	AD
Participant's Speech LLD - turn with max attention				
feature	functional	subgroups	p-value	large
pitch	avg	AS-HFA	0.013	HFA
		AD-HFA	0.009	HFA
	median	AS-HFA	0.034	HFA
		AD-HFA	0.034	HFA

TABLE VI
EMOTION TOPIC ANALYSIS: WCI IN SPEECH-IM-ABLSTM

	happy	scared	angry	sad
AD	3	**11**	8	6
AS	0	5	6	**8**
HFA	2	2	**8**	0
Total	5	18	**22**	14

focuses on the "scared" as the most important part, and the less important parts are "happy" and "sad". It is intriguing to see that both the modalities consider the "happy" part as the least important segment. The result implies that participant's (ASD subject) self-disclosure on their negative emotion experiences would result in subtly different behavior manifestation between the three subgroups; this behavior differences not only come from the participant (the ASD subject) but also are evident in the reciprocal gestural behaviors of the clinical investigators.

Our finding suggests that self-disclosing the topic of negative emotion is not only suitable for observing the differences between TD and ASD [26] but also is an important method in eliciting the subtle behavior differences between ASD subgroups. We hypothesize that this result may be connected to the ASD subject's life experience. The frequency for ASD individuals on experiencing negative emotion has shown to be more often than

TABLE VII
EMOTION TOPIC ANALYSIS: WCI IN MOTION-IM-ABLSTM

	happy	scared	angry	sad
AD	2	9	**10**	6
AS	6	**7**	3	3
HFA	2	**5**	4	1
Total	10	**21**	17	10

experiencing positive emotion [63], [64]. Moreover, appropriately regulating the emotion experiences of anger and anxiety are considered as one of the key deficits in ASD individuals [23]; the cause of the emotion disturbance is often related to their maladaptive emotional regulation strategy [23], [27]. Although more detailed clinical study is needed to further understand the relationship between negative emotion episodes and subtle behavior differences when disclosing these life experiences among different subtypes of ASD, our research has added to a consistent finding that the differential diagnosis of ASD may depend on the utilization of discussing personal *negative* emotion experiences.

IV. CONCLUSION

Developing objective methods to elicit and measure the differences between ASD subgroups remains a major challenge in performing differential diagnoses and advancing targeted intervention. In this work, we propose a Multimodal IM-aBSLTM to model the vocal behaviors and body movements in ADOS interview *Emotion* part for differentiating the three ASD subgroups (AS, AD, HFA). The Multimodal IM-aBLSTM embeds the interlocutors' behavior coordination using the interlocutor-modulation attention mechanism, where it automatically learns to emphasize the important segments during the interaction progression by jointly considering the dyad together. We use two different networks, a Speech-IM-aBLSTM and a Motion-IM-aBLSTM, to model the speech and gestural movement separately and a fusion network is used to combine the information to perform the final classification. Our method achieves an overall performance of 66.8 % UAR for the classification task. We further analyze and understand the attention mechanism from those highly weighted segments. Based on the ADOS procedure, our attention weights suggest that the participant's self-disclosure vocal behavior on the *anger* and the investigator's body movement in the *fear* part shows the differences between subgroups. We speculate there exists a connection between the behavior manifestation with their emotion experience and internal emotional regulation strategy.

In order to understand the exact question-answer content during these self-disclosing emotion episodes, our immediate future work is to extend the framework on studying the lexical content of the dialog. The word usage can reveal more intimate emotion and high-level complex attitude toward negative emotion e.g., hesitation, insecure and unwilling attitude. Aside from considering different frameworks, this work also provides a possible direction for redesigning a more targeted experiment to study the subgroup differences from the perspectives of sensitive emotion experience and issues of emotion regulation. We hope to continue advancing various signal processing and machine

learning frameworks to quantitatively investigate issues around human behaviors and mental health applications [65], [66].

REFERENCES

[1] C. Antaki, R. Barnes, and I. Leudar, "Self-disclosure as a situated interactional practice," *Brit. J. Social Psychol.*, vol. 44, no. 2, pp. 181–199, 2005.

[2] V. J. Derlega and J. H. Berg, *Self-Disclosure: Theory, Research, and Therapy.* Plenum Press, 1987.

[3] P. C. Cozby, "Self-disclosure: A literature review." *Psychological Bull.*, vol. 79, no. 2, 1973, Art. no. 73.

[4] S. M. Jourard and P. Lasakow, "Some factors in self-disclosure." *J. Abnormal Social Psychol.*, vol. 56, no. 1, 1958, Art. no. 91.

[5] J.-P. Laurenceau, L. F. Barrett, and P. R. Pietromonaco, "Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges." *J. Personality Social Psychol.*, vol. 74, no. 5, 1998, Art. no. 1238.

[6] K. DINDIA, M. Fitzpatrick, and D. Kenny, "Self-disclosure in spouse and stranger interaction a social relations analysis," *Human Commun. Res.-HUM COMMUN RES*, vol. 23, pp. 388–412, 1997.

[7] N. L. Collins and L. C. Miller, "Self-disclosure and liking: A meta-analytic review," *Psychological Bull.*, vol. 116, no. 3, 1994, Art. no. 457.

[8] J. G. Shapiro, H. H. Krauss, and C. B. Truax, "Therapeutic conditions and disclosure beyond the therapeutic encounter." *J. Counseling Psychol.*, vol. 16, no. 4, 1969, Art. no. 290.

[9] B. A. Farber, "Patient self-disclosure: A review of the research," *J. Clin. Psychol.*, vol. 59, no. 5, pp. 589–600, 2003.

[10] A. E. Kelly, "Helping construct desirable identities: A self-presentational view of psychotherapy." *Psychological Bull.*, vol. 126, pp. 475–494, 2000.

[11] M. Worthy, A. L. Gary, and G. M. Kahn, "Self-disclosure as an exchange process." *J. Personality Social Psychol.*, vol. 13, no. 1, 1969, Art. no. 59.

[12] A. W. Gouldner, "The norm of reciprocity: A preliminary statement," *Amer. Sociol. Rev.*, vol. 25, no. 2, pp. 161–178, 1960.

[13] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[14] S. L. Koole and W. Tschacher, "Synchrony in psychotherapy: A review and an integrative framework for the therapeutic alliance," *Frontiers Psychol.*, vol. 7, 2016, Art. no. 862.

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

[16] R. E. McEvoy, S. J. Rogers, and B. F. Pennington, "Executive function and social communication deficits in young autistic children," *J. Child Psychol. Psychiatry*, vol. 34, no. 4, pp. 563–578, 1993.

[17] M. Turner, "Annotation: Repetitive behaviour in autism: A review of psychological research," *J. Child Psychol. Psychiatry Allied Disciplines*, vol. 40, no. 6, pp. 839–849, 1999.

[18] K. A. Pelphrey, J. P. Morris, and G. McCarthy, "Neural basis of eye gaze processing deficits in autism," *Brain*, vol. 128, no. 5, pp. 1038–1048, 2005.

[19] H. Tager-Flusberg *et al.*, "Language and communication in autism," in *Handbook of Autism and Pervasive Developmental Disorders*, 3rd ed., Hoboken, NJ: Wiley, vol. 1, 2005, pp. 312–334.

[20] C. Lord *et al.*, "Austism diagnostic observation schedule: A standardized observation of communicative and social behavior," *J. Autism Developmental Disorders*, vol. 19, no. 2, pp. 185–212, 1989.

[21] 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," *Res. Autism Spectrum Disorders*, vol. 5, no. 1, pp. 455–462, 2011.

[22] L.-H. Quek, K. Sofronoff, J. Sheffield, A. White, and A. Kelly, "Co-occurring anger in young people with asperger's syndrome," *J. Clin. Psychol.*, vol. 68, no. 10, pp. 1142–1148, 2012.

[23] A. C. Samson, W. M. Wells, J. M. Phillips, A. Y. Hardan, and J. J. Gross, "Emotion regulation in autism spectrum disorder: Evidence from parent interviews and children's daily diaries," *J. Child Psychol. Psychiatry*, vol. 56, no. 8, pp. 903–913, 2015.

[24] B. P. Ho, J. Stephenson, and M. Carter, "Anger in children with autism spectrum disorder: Parent's perspective." *Int. J. Special Edu.*, vol. 27, no. 2, pp. 14–32, 2012.

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

[26] C. Rieffe, M. M. Terwogt, and K. Kotronopoulou, "Awareness of single and multiple emotions in high-functioning children with autism," *J. Autism Developmental Disorders*, vol. 37, no. 3, pp. 455–465, 2007.

[27] A. C. Laurent and E. Rubin, "Challenges in emotional regulation in asperger syndrome and high-functioning autism," *Topics Lang. Disorders*, vol. 24, no. 4, pp. 286–297, 2004.

[28] C. Rieffe, P. Oosterveld, M. M. Terwogt, S. Mootz, E. Van Leeuwen, and L. Stockmann, "Emotion regulation and internalizing symptoms in children with autism spectrum disorders," *Autism*, vol. 15, no. 6, pp. 655–670, 2011.

[29] A. C. Samson, A. Y. Hardan, I. A. Lee, J. M. Phillips, and J. J. Gross, "Maladaptive behavior in autism spectrum disorder: The role of emotion experience and emotion regulation," *J. Autism Developmental Disorders*, vol. 45, no. 11, pp. 3424–3432, 2015.

[30] D. Bone *et al.*, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *J. Speech, Lang., Hearing Res.*, vol. 57, no. 4, pp. 1162–1177, 2014.

[31] 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.

[32] 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," *J. Child Psychol. Psychiatry*, vol. 43, no. 3, pp. 327–342, 2002.

[33] 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," *J. Autism Developmental Disorders*, vol. 42, no. 6, pp. 920–928, 2012.

[34] C. Lord *et al.*, "A multisite study of the clinical diagnosis of different autism spectrum disorders," *Arch. General Psychiatry*, vol. 69, no. 3, pp. 306–313, 2012.

[35] A. P. Association, *et al.*, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Virginia, USA: American Psychiatric, 2013.

[36] U. Frith, "Emanuel miller lecture: Confusions and controversies about asperger syndrome," *J. Child Psychol. Psychiatry*, vol. 45, no. 4, pp. 672–686, 2004.

[37] T. Bennett *et al.*, "Differentiating autism and asperger syndrome on the basis of language delay or impairment," *J. Autism Developmental Disorders*, vol. 38, no. 4, pp. 616–625, 2008.

[38] C. Ecker, W. Spooren, and D. Murphy, "Developing new pharmacotherapies for autism," *J. Internal Med.*, vol. 274, no. 4, pp. 308–320, 2013.

[39] 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," *Behav. Modification*, vol. 36, no. 3, pp. 270–297, 2012.

[40] L. Schreibman, "Intensive behavioral/psychoeducational treatments for autism: Research needs and future directions," *J. Autism Developmental Disorders*, vol. 30, no. 5, pp. 373–378, 2000.

[41] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Toward differential diagnosis of autism spectrum disorder using multimodal behavior descriptors and executive functions," *Comput. Speech Lang.*, vol. 56, pp. 17–35, 2019.

[42] 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, pp. 2361–2365.

[43] Y.-S. Lin, S. S.-F. Gau, and C.-C. Lee, "An interlocutor-modulated attentional lstm for differentiating between subgroups of autism spectrum disorder," in *Proc. Interspeech*, 2018, pp. 2329–2333.

[44] 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," *J. Autism Developmental Disorders*, vol. 24, no. 5, pp. 659–685, 1994.

[45] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.0.37, Mar. 14, 2018. [Online]. Available: http://www.praat.org/.

[46] L. Centelles, C. Assaiante, K. Etchegoyhen, M. Bouvard, and C. Schmitz, "Understanding social interaction in children with autism spectrum disorders: does whole-body motion mean anything to them?" *L'Encephale*, vol. 38, no. 3, pp. 232–240, 2012.

[47] E. Milne, J. Swettenham, and R. Campbell, "Motion perception and autistic spectrum disorder: A review," *Current Psychol. Cognition*, vol. 23, no. 1/2, 2005, Art. no. 3.

[48] S. Tsermentseli, J. M. OBrien, and J. V. Spencer, "Comparison of form and motion coherence processing in autistic spectrum disorders and dyslexia," *J. Autism Developmental Disorders*, vol. 38, no. 7, pp. 1201–1210, 2008.

[49] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7291–7299.
[50] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4724–4732.
[51] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.
[52] Y.-S. Lin and C.-C. Lee, "Deriving dyad-level interaction representation using interlocutors structural and expressive multimodal behavior features," in *Proc. Interspeech*, 2017, pp. 2366–2370.
[53] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 909–913.
[54] 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 Trans. Affect. Comput.*, vol. 10, no. 4, pp. 552–567, Oct.–Dec. 2017.
[55] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5-6, pp. 602–610, 2005.
[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[57] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[58] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," 2015, *arXiv:1511.04119*.
[59] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2227–2231.
[60] 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.
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[62] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017.
[63] D. B. Shalom *et al.*, "Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism," *J. Autism Developmental Disorders*, vol. 36, no. 3, pp. 395–400, 2006.
[64] A. C. Samson, O. Huber, and J. J. Gross, "Emotion regulation in asperger's syndrome and high-functioning autism." *Emotion*, vol. 12, no. 4, 2012, Art. no. 659.
[65] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.
[66] 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 Process. Mag.*, vol. 34, no. 5, pp. 196–195, 2017.

Yun-Shao Lin (Student Member, IEEE) received the B.S. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2016. He is currently working toward the Ph.D. degree with the Electrical Engineering Department, NTHU, Hsinchu Taiwan. His research interests are in the human-centered behavioral signal processing, machine learning, and multimodal multiparty interaction. He was the recipient of Merry Electroacoustics Thesis Award, NTHU President's Scholarship, NOVATEK Scholarship, and FUJI Xerox Research Award. He is also a Student Member of the IEEE Signal Processing Society.

Susan Shur-Fen Gau received the M.D. degree from Chun-Shan Medical University, Taichung, Taiwan, in 1988 and the Ph.D. degree from Yale University, New Haven, CT, USA. She is a Professor of Psychiatry, Psychology, Epidemiology, Brain and Mind Sciences, Clinical Medicine, and Occupational Therapy with National Taiwan University, Taipei, Taiwan. She was the Director of Department of Psychiatry, National Taiwan University Hospital (NTUH) and College of Medicine (2009–2015), the Director of Department of Medical Genetics (2015–2018) in NTUHl, President of Taiwanese Society of Child and Adolescent Psychiatry (2014–2018), Vice-President of International Association of Child and Adolescent Psychiatry, and Allied Professionals (IAACAP, 2014–2018), and General Secretary of International Federation of Psychiatric Epidemiology (2012–2019). She has been an Editor of several monographs of IACAPAP since 2014. She has authored more than 250 SCI/SSCI articles since 2001, of which she is the first Author and Corresponding Author for more than 190 SCI/SSCI papers. Her main research interests include psychiatric, genetic, and pharmacological epidemiology of the three main mental disorders: sleep disorders, attention-deficit hyperactivity disorder (ADHD), and autism spectrum disorders (ASD). She has co-developed and prepared several Chinese versions of instruments for ADHD and ASD, conducted several studies on pharmacotherapy for ADHD, and been conducting the follow-up, family, neuropsychological, neuroimaging, neurophysiological, microbiomics, and genetic studies on ADHD and ASD. Her collaborative research also covers animal (mice & flies) and cellular (iPSC) models. She got outstanding research awards from the National Science Council (2012), National Taiwan University (2013), National Health Research Institute (2014), and NTUH (2016), and Lifetime Academic Achievement Award from Taiwanese Society of Psychiatry (2019), Taiwan. Her ADHD consortium got NTUH Outstanding Research Award contributing to medical research in 2019. She was awarded as the Best Clinical Professor by National Taiwan University College of Medicine Alumni Association in North America (2018). She is the keynote speaker/state-of-the-art/plenary speaker at more than 20 international congresses. She and her team have presented their work in peer-reviewed Congress on more than 700 occasions.

Chi-Chun Lee (Member, IEEE) received the B.S. and Ph.D. degrees both in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2007 and 2012, respectively. He is an Associate Professor with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan. His research interests are in the behavior computing, affective multimedia, and health analytics. He is an Associate Editor for the IEEE TRANSACTION ON MULTIMEDIA (2019–2020), and a TPC member for APSIPA, IVM. and ML. He serves as an Area Chair for Interspeech 2016, 2018, and 2019, senior program committee for ACII 2017, 2019, Publicity Chair for ACM ICMI 2018, sponsorship and special session Chair for ISCSLP 2018, 2020, and a Guest Editor in *Journal of Computer Speech and Language* on special issue of *Speech and Language Processing for Behavioral and Mental Health*. He led a team to the 1st place in Emotion Challenge in Interspeech 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a Co-Author on the Best Paper Award/finalist in Interspeech 2008 (top 12), Interspeech 2010 (top 3), IEEE EMBC 2018 (top 15), Interspeech 2018 (top 12), IEEE EMBC 2019 (Asia Pacific), APSIPA ASC 2019, and the most cited paper published in 2013 in *Journal of Speech Communication on Automatic Modeling of Couples' Behaviors* during therapeutic sessions. He was the recipient of the Taiwan MOST 2018 and 2019 Futuretek Breakthrough Award, MediaTek Genius for home Special Prize Award, and the USC Annenberg Fellowship. He is a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also a member of ISCA. He is currently involved in multiple granted interdisciplinary research projects with a focus on modeling human multimodal information using signal processing and machine learning.