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

Yun-Shao Lin1,3, Susan Shur-Fen Gau2, Chi-Chun Lee1,3

1Department of Electrical Engineering, National Tsing Hua University, Taiwan
2 Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taiwan
3MOST 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 part 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 means for finding out about 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 others’ 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-awareing 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 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 of 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
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 $\alpha_{tn}$ with LSTM can be used to differentiate between the three ASD subgroups.

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. 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
### Table 2: Comparison of model performance from M0 to M4. It shows the unweighted average recall (UAR). The overall result shows 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%.

<table>
<thead>
<tr>
<th>Models</th>
<th>Participant-only Methods</th>
<th>Interlocutor-Modulated Attention Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M0 (Baseline [20])</td>
<td>M1 (PO-LSTM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M2 (PO-aLSTM)</td>
</tr>
<tr>
<td></td>
<td>UAR</td>
<td>M3 (IB-aLSTM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M4 (IM-aLSTM)</td>
</tr>
<tr>
<td></td>
<td>0.525</td>
<td>0.576</td>
</tr>
<tr>
<td></td>
<td>AD</td>
<td>0.619</td>
</tr>
<tr>
<td></td>
<td>AS</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>HFA</td>
<td>0.455</td>
</tr>
</tbody>
</table>

Next, we obtain the time-normalized attention weight $\alpha_{it}$:

$$\alpha_{it} = \frac{exp(u_{it})}{\sum_t exp(u_{it})}$$  \hspace{1cm} (7)

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

$$s_t = \sum_i \alpha_{it} h_{it}$$  \hspace{1cm} (8)

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

$$y_i = \text{softmax}(s_i)$$  \hspace{1cm} (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.](image-url)
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.

Figure 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 obtains 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 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.

#### 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


