Using Interlocutor-Modulated Attention BLSTM to Predict Personality Traits in Small Group Interaction

Yun-Shao Lin
Department of Electrical Engineering
National Tsing Hua University
MOST Joint Research Center for AI Technology and All Vista Healthcare
Taiwan
astanley18074@gmail.com

Chi-Chun Lee
Department of Electrical Engineering
National Tsing Hua University
MOST Joint Research Center for AI Technology and All Vista Healthcare
Taiwan
cclee@ee.nthu.edu.tw

ABSTRACT

Small group interaction occurs often in workplace and education settings. Its dynamic progression is an essential factor in dictating the final group performance outcomes. The personality of each individual within the group is reflected in his/her interpersonal behaviors with other members of the group as they engage in these task-oriented interactions. In this work, we propose an interlocutor-modulated attention BSLTM (IM-aBLSTM) architecture that models an individual's vocal behaviors during small group interactions in order to automatically infer his/her personality traits. The interlocutor-modulated attention mechanism jointly optimize the relevant interpersonal vocal behaviors of other members of group during interactions. In specifics, we evaluate our proposed IM-aBLSTM in one of the largest small group interaction database, the ELEA corpus. Our framework achieves a promising unweighted recall accuracy of 87.9% in ten different binary personality trait prediction tasks, which outperforms the best results previously reported on the same database by 10.4% absolute. Finally, by analyzing the interpersonal vocal behaviors in the region of high attention weights, we observe several distinct intra- and inter-personal vocal behavior patterns that vary as a function of personality traits.

KEYWORDS

small group interaction, personality traits, behavioral signal processing, attention mechanism, social signal processing

ACM Reference Format:
Yun-Shao Lin and Chi-Chun Lee. 2018. Using Interlocutor-Modulated Attention BLSTM to Predict Personality Traits in Small Group Interaction. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3242969.3243001

1 INTRODUCTION

Small group interaction is defined as an unit of interaction including three to six people engage in face-to-face interactions [11]. This particular interaction setting is a common daily life scenario, especially in school and workplace. In general, all observable interpersonal behaviors within an arbitrary period are referred to as an "interaction process" and are also considered as mediators between group-related factors, such as member, task and environment, and group performance [22]. Group scholars have long agreed that the dynamics of the interaction process significantly affect the outcomes of the group performances [15]. For example, group-based cooperative learning is a strategic learning technique emphasizing the interaction process between student members, and it has been shown that helping behaviors between students would create a positive interaction process leading to an improvement in learning efficiency and academic achievement [34].

Personal attribute is a major factor in forming the interaction process of a small group as it shapes the expressed verbal and nonverbal interpersonal behaviors of an individual [23]. In fact, the relationship between personality traits and behaviors expressed during interaction has been well documented. For example, personality traits are correlated with how the emotion expressive behaviors are displayed [8, 10]; in specifics, the person who has the ability to express anger through vocal and facial expressions are more likely to be the dominant person. Within a group interaction, an individual would also develop "social impressions" of personal traits, e.g., an emergent leader would evolve naturally taking on the role in contributing to the collective goals by providing group directions [7]. These perceived social impressions of personal traits can be identified by examining the observed non-verbal (i.e., non-linguistics) multimodal behaviors during the interaction [27, 30, 32].

In recent years, by designing approrpriate group tasks, a series of multimodal databases have been collected; some notable audio-video databases include AMI [4], VACE [6], Mission Survival [26], and ELEA [29]. Researchers have, hence, investigated a variety of technical frameworks to automatically compute personality traits and social impressions [5], especially on attributes of *leadership* [2] and *dominance* [18], using audio-video recordings. In fact, these research effort has resulted in an interdisciplinery field of social signal processing (SSP) [33]. Among these databases, the ELEA database is one of the largest databases of small group interactions. Past works have examined computational frameworks in this database mostly to identify emergent leadership. For example, Aran and Gatica-Perez present one of the first works in predicting social impressions by deriving a series of handcrafted features on motion, speaking turn, vocal characteristic, visual activities and eye gaze for each target member. Okada et al. further improves the prediction

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243001

accuracy of personality trait by considering interactive behaviors between group members, i.e., the co-occurrent events between target speaker and other group members [25]. Fang et al. achieves the current highest prediction performance by considering both the intra-personal features, dyadic features and group features [9].

However, most of these works rely on engineering hand-crafted multimodal features in order to compute each target member's self behaviors and his/her interactive behaviors with other members in the group. In this work, we present a network architecture of *interlocutor-modulated attention* BLSTM (IM-aBLSTM) that captures both the target speaker's self vocal behaviors and his/her interactive behaviors with other group members jointly to improve the prediction of the target speaker's personality traits. The network involves a BLSTM structure that models the progression of the target speaker's vocal characteristic over an interaction. The personality-relevant information in the interactive vocal behaviors of the target speaker with other group members is embedded into the BLSTM using attention mechanism. Specifically, we jointly learn a *pair of attention weights* by computing attention of the current target speaker turn with the immediate preceding speaker turn and the follow-up speaker turn. These paired weighting are termed as the *interlocutor-modulated attention*, which are jointly optimized when learning the attention BLSTM vocal network.

We evaluate our proposed IM-aBLSTM in predicting personality traits in the ELEA corpus. Our IM-aBLSTM achieves an overall unweighted recall rates of 87.9% in tasks of high versus low categorization across ten different personality traits. Our framework outperforms the best reported results on the same dataset (Fang et al. [9]) by 10.4% absolute. The use of *interlocutor-modulated* attention mechanism shows an improvement of 7.5% relative over using target speaker-only attention mechanism, i.e., without considering interactive behaviors with other team members. This reinforces the importance of modeling the "interaction process" of the interlocutors jointly in improving the personality traits recognition. Lastly, by examining the high *interlocutor-modulated* attention region, we demonstrate that the prosodic differences between those target speakers with high score of personality traits and those with low score. Further analysis reveals the relative personality difference of the target speaker with his/her immediate surrounding (in terms of conversational turns) speakers, which the IM-aBLSTM leverages the most to infer the target speaker's personality traits.

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 conclusion and future work.

2 RESEARCH METHODOLOGY

2.1 The ELEA Corpus

The Emergent LEAder (ELEA) corpus [29] is one of the largest group interaction database and is originally collected for analyzing the emergent leadership in the newly formed groups without predefined roles. In the database, three or four people are asked to form a group to engage in interaction to complete the winter survival task. The group members are first told to imagine themselves as the survivors of an airplane crash, and they need to rank 12 objects according to the order of importance in order for the group to survive through winter. In the first five minutes, they are asked to form their own ranking list. Then, in the next 15 minutes, group members would engage in a discussion to identify a common shared list. The analysis data is derived from the audio-video recordings of these group discussion sessions.

	median	max	min
Agr	3.96	5.0	2.5
Con	3.92	5.0	2.38
Ext	3.75	4.88	2.0
Neu	2.38	4.5	1.13
Opo	3.5	4.58	2.17
RDom	2.5	4.0	1.0
PCom	3.81	4.88	2.67
PDom	2.92	4.83	1.5
PLead	3.29	4.75	1.5
PLike	4.38	5.0	2.17

Table 1: A summary on the range of the ten personality trait scores and their median values used in this work

In our work, we focus on analyzing group interactions with exactly four members only. This subset of the groups contains a total of 28 interactions including 112 unique persons. All of the audio data has been manually segmented into utterances and labelled with the speaker identity. We define a *turn* in this work as a complete speaker region before a speaker floor is changed. The target personality traits used in this work include two major categories: the self-assessed questionnaires on the Big Five personality traits [13] and the Perceived Interaction Scores. There are five traits of self-assessed personality scores: Extraversion (Ext), Agreeableness (Agr), Conscientiousness (Con), Neurotic (Neu), and Openness to Experience (Ope). The Perceived Interaction Scores are obtained with perceptual assessment by other group members in the team. It includes five additional personality traits scoring: Perceived Leadership (PLead), Perceived Dominance (PDom), Perceived Competence (PCom), Perceived Liking (PLike) and Dominance Ranking (RDom). Table 1 lists the range of ten scores for each of these traits in this dataset, and the median scores (also reported in Table 1) are used to define the binary level for each personality prediction task to provide the same experimental setup to the previous work [9].

2.2 Interlocutor-Modulated Attention BLSTM

Figure 1 shows our complete proposed interlocutor modulated attention BLSTM (IM-aBLSTM) architecture. This BLSTM is learned from the input of the Target Speaker (TS)'s vocal features. The time step is defined as every *turn* of the TS. Furthermore, for every k-th turn of the TS, we define a "Forward Turn (FT)" and a "Backward Turn (BT)". A FT is refers to the *turn* immediately preceding that k-th turn of the TS, and a BT refers to the *turn* immediately following that k-th turn of the TS. The FT and the BT will be used to derive our interlocutor-modulated attention weights, which capture the contextual interaction behaviors of the TS speaker with other team members. In the following sections, we will first describe the extraction of *turn*-level acoustic inputs for the BLSTM and further detail our proposed interlocutor-modulated attention mechanism.

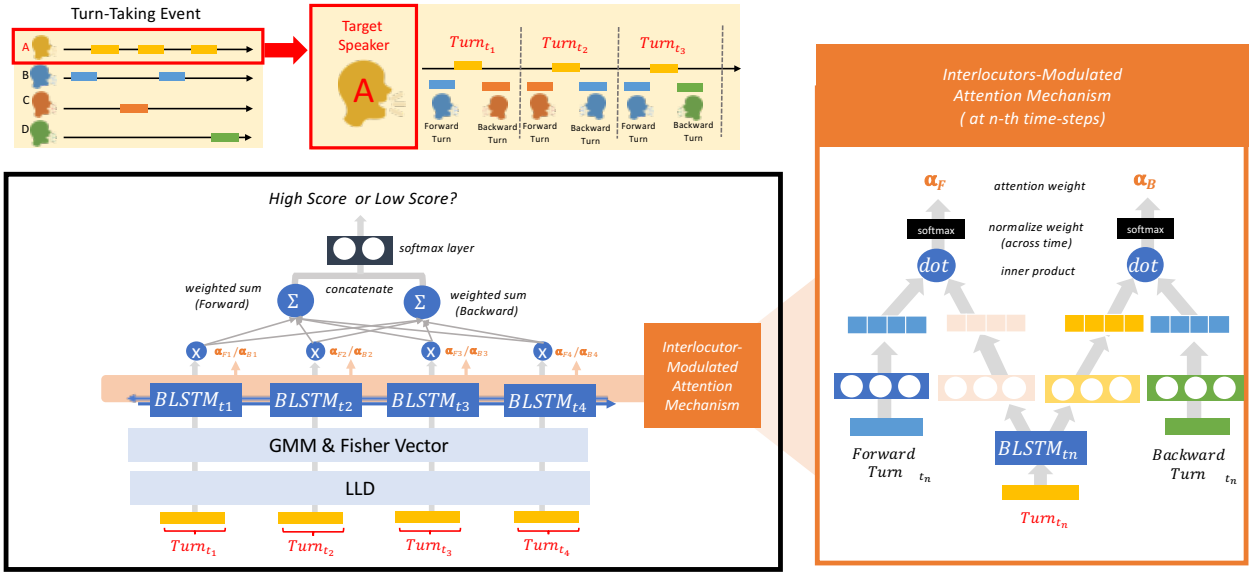

Figure 1: Detail Structure of Interlocutor-Modulated Attentional BLSTM: Our IM-aBLSTM introduces an Interlocutor-Modulated Attention Mechanism to emphasize the important turn-feature during small group face-to-face interaction in the ELEA corpus. The turn-level feature is a fixed high-dimensional acoustic feature encoded using GMM-based Fisher scoring. We model the progress of the turn-level features using BLSTM with the Interlocutor-Modulated Attention Mechanism. Finally, the learnable weight pair α_F and α_B with LSTM can be used differentiate the score of personality score.

2.2.1 Turn-level Acoustic Features. The turn-level acoustic features are computed on the speaking portions of each subject. First, we extract *frame*-level acoustic low-level descriptors (LLDs) on each of the target speaker, S_k, over their corresponding speaking portions during the interaction. We compute a total of 45 LLDs including pitch, intensity, MFCC, and their delta and delta-delta by using the Praat toolkit [3]. All LLD features are extracted at a frame-rate of 10ms and are z-normalized according to the speaker.

Then, we learn a single fixed dimensional vector of turn-level acoustic features as input to our BLSTM network. Each sequence of these LLDs are encoded into a high-dimensional acoustic space by using a method of Gaussian Mixture Model (GMM, λ) based Fisher scoring [28]. Briefly speaking, this method is operated by encoding a varying length sequence of LLDs for each data sample, $\bar{x}$, into a fixed-length feature vector through computing the gradient log-likelihood function, i.e., Fisher scoring (indicating the direction of λ to better fit $\bar{x}$), with respect to the first and second order statistics of the background GMM:

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

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

,with total frame number T and the posterior probability $r_t(c)$ given the observation x_t produced by the c-th Gaussian with mean μ_c and standard deviation σ_c. This encoded vector of $[g^X_{\mu_c}\ g^X_{\sigma_c}]$ is our turn-level acoustic features. Recently, this particular encoding method has been shown to be useful in speech-related recognition tasks, including detection of emotion [21] and paralinguistic attribute [19], and evaluation of impromptu speech [17].

2.2.2 Interlocutor-Modulated Attention Mechanism. We utilize Bi-directional Long Short Term Memory (BLSTM) neural network [14] to recognize each subject's ten personality traits by modeling the temporal progression of the target speaker's turn-level vocal features of during the interaction. BLSTM is an improved version of the LSTM [16] by combining both the forward LSTM and the backward LSTM to ensure the temporal gradient can be equally and fully transmitted.

For each meeting session, we input the speaker $S'_k s$ turn-level feature sequence x to obtain a corresponding output sequence of BLSTM's hidden states h_k, which is a concatenation of h_f obtained from a forward LSTM and h_b from a backward LSTM.

$$\{h_{f1}, \ \ldots \ , h_{fT}\} = LSTM_{forward}(\{x_{k1}, \ \ldots \ , x_{kT}\}) \quad (3)$$

$$\{h_{b1}, \ \ldots \ , h_{bT}\} = LSTM_{backward}(\{x_{k1}, \ \ldots \ , x_{kT}\}) \quad (4)$$

$$h_{kt} = [h_{ft}; h_{bt}] \quad (5)$$

Furthermore, the use of attention mechanism [1] in neural network sequence modeling has brought significant improvements with explainable results across a variety of recognition tasks, e.g., motion recognition [31], emotion recognition [24], addiction counseling [12], etc. The attention mechanism can be thought as additional neural network structures that help emphasize the important parts of the sequence for discriminative tasks. It is achieved operationally by having learnable attention weights applied on the sequence models, e.g., LSTM or BLSTM. In our work, while the

BLSTM concentrates on modeling the TS's self vocal cues, the interactive vocal behaviors of TS with other members of the group is further incorporated into the BLSTM using our proposed *interlocutor-modulated* attention weights.

Our proposed *interlocutor-modulated* attention weights capture the time-dependent interactive relationship of each TS's turn with his/her preceding speaker's turn (FT) and follow-up speaker's turn (BT); the architecture of our attention mechanism is shown in Figure 1 (right). In specifics, we propose to learn two *interlocutor-modulated* attention weights; one of them is a Forward Attention Weight, α_{Ft}, and another one is a Backward Attention Weight, α_{Bt}. For each time step, we first place an additional fully-connected hidden layer each to transform the original hidden state h_{kt} to h_{Ft} and h_{Bt} separately. The transformed hidden state sequence are denoted as:

$$h_{Ft} = tanh\left(W_F h_{kt} + b_F\right) \tag{6}$$

$$h_{Bt} = tanh\left(W_B h_{kt} + b_B\right) \tag{7}$$

We also transform FT't turn-level feature f_t and BT 's turn-level feature b_t to g_{Ft} and g_{Bt} by introducing also a fully connected hidden layer each separately with the same dimensions used for h_{kt}. The transformed hidden state sequence are denoted as:

$$g_{Ft} = tanh\left(W_{Ff} f_t\right) \tag{8}$$

$$g_{Bt} = tanh\left(W_{Bb} b_k\right) \tag{9}$$

We then compute our "Forward Attention Weight" u_{Ft} and "Backward Attention Weight" u_{Bt} for the "t-th" time-step for TS using:

$$u_{Ft} = \langle g_{Ft}, h_{Ft} \rangle \tag{10}$$

$$u_{Bt} = \langle g_{Bt}, h_{Bt} \rangle \tag{11}$$

Next, we obtain the time-normalized attention weight α:

$$\alpha_{Ft} = \frac{exp\left(u_{Ft}\right)}{\sum_t^T exp\left(u_{Ft}\right)} \tag{12}$$

$$\alpha_{Bt} = \frac{exp\left(u_{Bt}\right)}{\sum_t^T exp\left(u_{Bt}\right)} \tag{13}$$

These *interlocutor-modulated* bi-directional attention weights are combined to the TS's BLSTM's hidden vectors h_{kt} using the following equation:

$$s_F = \sum_t \alpha_{Ft} h_{kt} \tag{14}$$

$$s_B = \sum_t \alpha_{Bt} h_{kt} \tag{15}$$

We concatenate the representation of s_F and s_B to obtain the final representation S:

$$S = [s_F; s_B] \tag{16}$$

Finally, the binary prediction (high versus low score) can be achieved by inputing the representation S through a softmax function:

$$y = softmax(S) \tag{17}$$

If our target is to regress to the actual numerical scores, it can be achieved by input the representation S to the layer with "relu" activation function and the final layer without activation function in the following form:

$$S_{fc_0} = relu(W_{fc_0} S + b_{fc_0}) \tag{18}$$

$$y = W_{fc_1} S_{fc_0} + b_{fc_1} \tag{19}$$

3 EXPERIMENTAL SETUP AND RESULTS

3.1 Experimental Setup

In this work, we perform two different types of prediction tasks on the ten personality traits (mentioned in section 2.1). The first type is a classification task, where the the original score of personality traits are converted to binary values (high or low) by thresholding using median value. This experiment is carried out to provide the same setting as previous results on the same database [9]. We also additionally perform regression task to predict the original score values of these ten personality traits. The experiment is carried out using leave-one-group-out cross validation with the metric of unweighted average recall (UAR) and accuracy for the classification task and Pearson and Spearman correlation for the regression task.

3.1.1 Models Comparison. The baseline model is provided by Fang et al. [9], which is the best reported results in classifying these ten personality traits in the ELEA corpus. We further compare our proposed framework with the Target Speaker-Only Attention Bidirectional LSTM (TS-aBLSTM), Backward Speaker Attentional BLSTM (BS-aBLSTM), Forward Speaker Attentional BLSTM (FS-aBLSTM) models in both classification and regression tasks. The main differences between the models is that the TS-aBLSTM only use the target speaker's turn level features to derive the attention weight α, FS-aBLSTM use forward contextual attention weight only to derive $\alpha_{forward}$, BS-aBLSTM use backward contextual attention weight only to derive $\alpha_{backward}$, unlike the IM-aBLSTM, where the attention weights are computed by integrating contextual interactive behaviors of the TS with other group members.

3.1.2 Other Experimental Parameters. The BLSTM is trained with a fixed length (152 time-steps), which is the maximum number of turns that a participant have in the ELEA corpus. We zero-pad those speakers with less than 152 turns. The number of hidden nodes in the BLSTM is eight, i.e., each direction of LSTM has four units. The two different dense layers used in the interlocutor-modulated mechanism has eight units, which corresponds to the output size of layer with WF, W_B and the layer with WFf, W_{Bb}. During the training stage, we choose our batch size 45, learning rate 0.01 with ADAM optimizer [20]. Cross entropy is used as the loss function for classification task, and mean square error is used as our loss function for the regression task. Both the tasks are trained with 5 epochs for our proposed network structure.

3.2 Experimental Results and Analyses

3.2.1 Analysis on Model Performance. Table 2 summarizes our complete classification results. Among 4 different comparison models, our proposed IM-aBLSTM achieves the best overall classification accuracy on the 10 personality traits task (an average of 87.9% UAR calculated across 10 different personality traits). This method significantly outperforms baseline results obtained in the previous work [9] by 10.4% absolute and results also imply that the usage

	Agr	Con	Ext	Neu	Ope	RDom	PCom	PDom	PLead	PLike
Baseline [9]										
Accuracy	0.775	0.794	0.775	0.794	0.735	0.814	0.725	0.765	0.804	0.775
Target Speaker-Only Attentional Bidirectional LSTM (TS-aBLSTM)										
Accuracy	0.768	0.866	0.804	0.813	0.750	0.804	0.866	0.795	0.750	0.857
UAR	0.768	**0.866**	0.789	0.814	0.740	0.794	**0.866**	0.791	0.750	0.859
Backward Speaker Attentional Bidirectional LSTM (BS-aBLSTM)										
Accuracy	0.830	0.821	0.875	0.830	0.875	0.839	0.848	0.839	0.839	0.866
UAR	0.831	0.822	0.864	0.830	0.874	0.821	0.848	0.842	0.837	0.860
Forward Speaker Attentional Bidirectional LSTM (FS-aBLSTM)										
Accuracy	0.830	0.884	0.848	0.848	0.866	0.893	0.804	0.866	0.866	0.848
UAR	0.831	0.881	0.845	0.846	0.863	**0.887**	0.802	0.865	0.867	0.849
Interlocutor-Modulated Attentional Bidirectional LSTM (IM-aBLSTM)										
Accuracy	0.893	0.830	0.902	0.875	0.884	0.848	0.866	0.875	0.902	0.920
UAR	**0.893**	0.828	**0.899**	**0.876**	**0.890**	0.836	**0.866**	**0.877**	**0.900**	**0.922**

Table 2: Classification Result on Ten Binary Personality Trait Level: Classification experiments by threshold the scores using median value. Metrics used are unweighted average recall (UAR) and weighted accuracy. Comparison models include: a baseline model (best reported accuracy on the same database [9], the Target Speaker-Only Attention BLSTM (TS-aBLSTM), Backward Speaker Attentional BLSTM (BS-aBLSTM), Forward Speaker Attentional BLSTM (FS-aBLSTM) and Interlocutor-Modulated Attention BLSTM (IM-aBLSTM)

	Agr	Con	Ext	Neu	Ope	RDom	PCom	PDom	PLead	PLike
Target Speaker-Only Attentional BLSTM (TS-aBLSTM)										
Pearson	0.034	0.097	0.226	0.219	0.068	0.054	0.118	0.228	0.179	0.224
Spearman	0.053	0.094	0.255	0.180	0.033	0.072	0.116	0.167	0.147	0.244
Backward Speaker Attentional Bidirectional LSTM (BS-aBLSTM)										
Pearson	0.174	0.378	0.238	0.595	0.235	0.680	0.074	0.482	0.337	0.349
Spearman	0.173	0.283	0.213	0.610	0.202	**0.676**	0.048	0.478	0.377	0.264
Forward Speaker Attentional Bidirectional LSTM (FS-aBLSTM)										
Pearson	0.197	0.278	0.241	0.613	0.408	0.587	0.181	0.425	0.272	0.330
Spearman	0.211	0.259	0.242	**0.600**	**0.403**	0.593	0.225	0.457	0.317	0.265
Interlocutor-Modulated Attentional BLSTM (IM-aBLSTM)										
Pearson	0.205	0.429	0.400	0.477	0.370	0.574	0.247	0.485	0.414	0.446
Spearman	**0.150**	**0.373**	**0.379**	0.468	0.363	0.542	**0.236**	**0.469**	**0.402**	**0.341**

Table 3: Regression Result on the Ten Personality Traits: Comparison of prediction accuracy using the TS-aBLSTM, BS-aBLSTM, FS-aBLSTM and IM-aBLSTM. Metrics used are Pearson and Spearman Correlations

of both forward and backward attention shows the better result than only use either one. The use of attention BLSTM in time-series modeling by itself already provides improved discriminatory power over hand-crafted features by comparing TS-aBLSTM to baseline model (80.4% vs. 77.5% average UAR). However, in attributes such as RDom and PLead, we observe that baseline model is still competitive to TS-aBLSTM due to the fact that TS-aBLSTM does not explicitly model the interactive behaviors where the scores of these two attributes intuitively are related to the relative behaviors of the TS and other members. The proposed IM-aBLSTM, which integrates interlocutors behavior information, provides an overall improved classification accuracy (87.9% vs. 80.4%), and it outperforms TS-aBLSTM for most of the 10 personality traits (except for the attribute of Conscientiousness (Con)).

Table 3 lists the correlations obtained by using TS-aBLSTM and IM-aBLSTM to regress the actual ten personality scores. The TS-aBLSTM generally does not perform well in this more complex learning scenario (0.145 average Pearson correlation), where our proposed IM-aBLSTM obtains an overall 0.405 of Pearson correlation. As the classification task, the results also imply that the usage of both forward and backward attention shows the better result than only use either one. After separately examining the regression performance on each personality trait, we observe that our proposed IM-aBLSTM improves over TS-aBLSTM from insignificant correlations to moderate correlations on personality attributes of Conscientiousness (Con), Openness (Ope), and Dominance Ranking (RDom). In fact, the Spearman correlation obtained for RDom improves from 0.072 to 0.542. Almost all of the personality attributes

improve in the regression tasks, and specifically, we obtain encouraging Spearman correlations of 0.468, 0.542, 0.569, and 0.402 on attributes of Neurotic (Neu), Dominance Ranking (RDom), Perceived Dominance (PDom), and Perceived Leadership (PLead), respectively. Many of these attributes are related to perceived leadership and dominance, which has been shown to related closely to the interactive behaviors during the small group interaction process (e.g., [9], [25]).

3.2.2 Analyses of High Attention Regions of Interaction. We further provide two additional analyses by examining the highest attention weights regions of a TS given by our model (IM-aBLSTM).

- **Prosodic Analysis**: Vocal prosodic features have been indicated to be important descriptors of the personality traits. We first extract the *high attention* turns for all of TS speakers within their corresponded interaction, and then we compute mean and standard deviation of the pitch and intensity values on these turns. We perform the Student's two sample T-test to examine the difference of these prosodic indicators between high and low score of each personality trait.
- **Patterns of Interlocutors**: Since our IM-aBLSTM learns a pair of attention weights, where one of them indicates the importance of the immediate preceding speaker turn and the other one indicates the importance of the follow-up speaker turn. We further provide an analysis on these important high attention weights by examining whether there exist patterns between interlocutors types and target speaker on these segments. In specifics, we compute the absolute differences of the personality scores between the preceding speaker and the TS speaker over these high attention region, and similarly we carry out the same procedure for the follow-up speaker.

We present our prosodic analysis results in Table 4. We list the descriptors that show a statistically significant difference between high and low personality scores ($\alpha \leq 0.05$) with their associated directions. Several notable observations can be made: the person with a higher level of Perceived Competence (PCom) appears to have a statistically lower mean intensity values. For attributes related to dominance and leadership, we also observe distinct prosodic patterns. In specifics, the standard deviation of pitch is higher for speakers with higher level of Perceived Dominance (PDom), and the mean pitch is statistically higher in speakers with higher Perceived Leadership (PLead). Speakers of different self-assessed personality traits, such as Openness (Ope) and Agreeableness (Agr), have also shown to exhibit distinct prosodic manifestations.

Furthermore, the results on patterns of interlocutors type analysis are presented in Table 5. We examine the three perceived personality attributes of RDom, PDom, and PLead in this analysis. The first column of Table 5 shows that the average absolute score differences of that particular attribute in the complete ELEA database without examining the high attention turns. "Forward" column indicates the average absolute score difference of that particular attribute computed between the *preceding* speaker and the target speaker on the high attention turns, and "Backward" indicates the same average score differences with respect to the *follow-up* speaker. For the RDom, we observe a relative larger differences in this dominance ranking rating between the preceding speaker and the target

Forward Turn			
feature	label	feature value(label=1)	t-test
mean_intensity	Ope	low*	p=0.017
	PCom	low**	p=0.001
std_intensity	Agr	low*	p=0.017
	PCom	low*	p=0.013
mean_pitch	PLead	high*	p=0.038
std_pitch	Agr	high**	p=0.003
	PCom	high**	p<0.001
	PDom	high*	p=0.008
Backward Turn			
feature	label	feature value(label=1)	t-test
mean_intensity	PCom	low**	p=0.004
mean_ptich	PLead	high*	p=0.038
std_pitch	Ope	high*	p=0.017
	PDom	high*	p=0.033

Table 4: Prosodic Analysis on the Highest Attention Turn: We first extract the high attention turns using our model. We then compute mean and standard deviation of the pitch and intensity values on these turns. We report results ($\alpha \leq 0.05$) after performing the Student's two sample T-test to examine the difference of these prosodic indicators between high and low value of personality trait. * indicates $p \leq 0.05$, ** indicates $p \leq 0.01$

	Average	Forward	Backward
Rdom	1.244	1.302	1.215
PDom	0.762	0.712	0.796
PLead	0.649	0.592	0.67

Table 5: Pattern of Interlocutors Analysis on the Highest Attention Turn: the first column of table shows that the average absolute score differences in the ELEA database without examining the high attention turns. "Forward" column indicates the average absolute score difference computed between the *preceding* speaker and the target speaker on high attention turns, and "Backward" indicates the same differences with respect to the *follow-up* speaker

speaker, and a relatively more similar rating in RDom (smaller average absolute differences) between the follow-up speaker and the target speaker (comparing RDom-Backward with RDom-Average or RDom-Forward). The trend is reversed for PLead and PDom, where the preceding speaker has a relatively more similar perceived score rating in PLead and PDom with the target speaker, and the follow-up speaker tends to be more different in these ratings than the target speaker.

In summary, we present two different analyses. By examining the high attention turns, we observe several important prosodic differences existed in the target speaker that vary as a function of his/her personality values. Also, it is interesting to identify the pattern of interlocutor types that our IM-aBLSTM focuses on when making a prediction. We see that the algorithm relies more heavily on computing interactive vocal behavior on turns where the preceding and the follow-up speaker have a more similar (or disparate)

perceived personality rating as compared to the database as a whole in order to make accurate predictions.

4 CONCLUSIONS AND FUTURE WORKS

Small group interactions is most commonly found in workplace and school situations. Personal attributes of personality traits affect both individual's behaviors and his/her interaction process with other members of the group. The dynamics of the interaction process is key in dictating the final group performance outcomes. Due to the importance in objectively understanding the intricate behavior dynamics in small group interactions, there has been an increasing interest on research that build computational frameworks to automatically assess and recognize individual's personality traits. In this work, we propose an IM-aBLSTM framework that models the vocal behaviors of both the target speaker and his contextual interlocutors to improve the prediction performance on the score of ten different personality traits in the ELEA corpus.

Our method achieves a promising UAR of 87.9% over ten personality traits in high vs. low classification task, and it obtains an average Pearson correlation of 0.405 for regression task. IM-aBLSTM framework outperforms the current state-of-art personality recognition accuracy on this corpus. Furthermore, our analyses on the high interlocutor-modulated attention regions demonstrate that prosodic variations indeed vary according to each individual's personality trait. Also, the proposed IM-aBLSTM makes its improved recognition by concentrating on the interactive vocal behaviors during the conversational segments of speakers (preceding - target - follow-up) where there is more changes in the personality traits.

This work presents a preliminary personality prediction result by modeling vocal behaviors between interlocutors via embedding interaction-based attention mechanism in a BLSTM. Our future work will focus on leveraging multiple behavior modalities to advance our algorithm in modeling the relationship between target speaker and interlocutors to improve regression correlations across these ten personality traits. Additionally, we would also investigate algorithmic frameworks that jointly models individual behaviors at the group level to predict the final group performance outcome as they engage in a variety of small group interaction contexts.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2018. Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features. *IEEE Transactions on Multimedia* 20, 2 (2018), 441–456.

[3] P Boersma and D Weenink. 2003. Praat-A system for doing phonetics by computer [Computer Software]. *The Netherlands: Institute of Phonetic Sciences, University of Amsterdam* (2003).

[4] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 28–39.

[5] Oya Celiktutan and Hatice Gunes. 2017. Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. *IEEE Transactions on Affective Computing* 8, 1 (2017), 29–42.

[6] Lei Chen, R Travis Rose, Ying Qiao, Irene Kimbara, Fey Parrill, Haleema Welji, Tony Xu Han, Jilin Tu, Zhongqiang Huang, Mary Harper, et al. 2005. VACE multimodal meeting corpus. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 40–51.

[7] Walter H Crockett. 1955. Emergent leadership in small, decision-making groups. *The Journal of Abnormal and Social Psychology* 51, 3 (1955), 378.

[8] Michael R Cunningham. 1977. Personality and the structure of the nonverbal communication of emotion. *Journal of Personality* 45, 4 (1977), 564–584.

[9] Sheng Fang, Catherine Achard, and Séverine Dubuisson. 2016. Personality classification and behaviour interpretation: An approach based on feature categories. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 225–232.

[10] Howard S Friedman, M Robin DiMatteo, and Angelo Taranta. 1980. A study of the relationship between individual differences in nonverbal expressiveness and factors of personality and social interaction. *Journal of Research in Personality* 14, 3 (1980), 351–364.

[11] Daniel Gatica-Perez, Oya Aran, and Dinesh Jayagopi. 2017. *Analysis of Small Groups*. Cambridge University Press, 349âĂŞ367. https://doi.org/10.1017/9781316676202.025

[12] James Gibson, Dogan Can, Panayiotis Georgiou, David C Atkins, and Shrikanth S Narayanan. 2017. Attention networks for modeling behaviors in addiction counseling. In *Proc. Interspeech*.

[13] Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology* 59, 6 (1990), 1216.

[14] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

[15] J Richard Hackman and Charles G Morris. 1975. Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration1. In *Advances in experimental social psychology*. Vol. 8. Elsevier, 45–99.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Shan-Wen Hsiao, Hung-Ching Sun, Ming-Chuan Hsieh, Ming-Hsueh Tsai, Yu Tsao, and Chi-Chun Lee. 2017. Toward Automating Oral Presentation Scoring during Principal Certification Program using Audio-video Low-level Behavior Profiles. *IEEE Transactions on Affective Computing* (2017).

[18] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2009. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 3 (2009), 501–513.

[19] Heysem Kaya, Alexey A Karpov, and Albert Ali Salah. 2015. Fisher vectors with cascaded normalization for paralinguistic analysis. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[20] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Yun-Shao Lin and Chi-Chun Lee. 2017. Deriving Dyad-Level Interaction Representation using Interlocutors Structural and Expressive Multimodal Behavior Features. *Proc. Interspeech 2017* (2017), 2366–2370.

[22] Joseph Edward McGrath. 1964. *Social psychology: A brief introduction*. Holt, Rinehart and Winston.

[23] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.

[24] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2227–2231.

[25] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 15–22.

[26] Fabio Pianesi, Massimo Zancanaro, Bruno Lepri, and Alessandro Cappelletti. 2007. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation* 41, 3-4 (2007), 409–429.

[27] Jurgen Ruesch, Weldon Kees, Robert Goodloe Harper, Robert G Harper, Arthur N Wiens, and Joseph D Matarazzo. 1978. *Nonverbal Communication: The State of the Art*. Vol. 65. Univ of California Press.

[28] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. 2013. Image classification with the fisher vector: Theory and practice. *International journal of computer vision* 105, 3 (2013), 222–245.

[29] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.

[30] Barry Schwartz, Abraham Tesser, and Evan Powell. 1982. Dominance cues in nonverbal behavior. *Social Psychology Quarterly* (1982), 114–120.

[31] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* (2015).

[32] R Timothy Stein. 1975. Identifying emergent leaders from verbal and nonverbal communications. *Journal of Personality and Social Psychology* 32, 1 (1975), 125.

[33] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[34] Noreen M Webb. 1982. Student interaction and learning in small groups. *Review of Educational research* 52, 3 (1982), 421–445.