An Interaction Process Guided Framework for Small-Group Performance Prediction

YUN-SHAO LIN,
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
YI-CHING LIU,
[2]Department of Business Administration, National Taiwan University, Taiwan
CHI-CHUN LEE,
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

A small group is a fundamental interaction unit for achieving a shared goal. Group performance can be automatically predicted using computational methods to analyze members' verbal behavior in task-oriented interactions, as has been proven in several recent works. Most of the prior works focus on lower-level verbal behaviors, such as acoustics and turn-taking patterns, using either hand-crafted features or even advanced end-to-end methods. However, higher-level group-based communicative functions used between group members during conversations have not yet been considered. In this work, we propose a two-stage training framework that effectively integrates the communication function, as defined using Bales' interaction process analysis (IPA) coding system, with the embedding learned from the low-level features in order to improve the group performance prediction. Our result shows a significant improvement compared to the state-of-the-art methods (4.241 MSE and 0.341 Pearson's correlation on NTUBA-task1 and 3.794 MSE and 0.291 Pearson's correlation on NTUBA-task2) on the NTUBA (National Taiwan University Business Administration) small-group interaction database. Furthermore, based on the design of IPA, our computational framework can provide a time-grained analysis of the group communication process and interpret the beneficial communicative behaviors for achieving better group performance.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; • **Computing methodologies → Modeling methodologies**.

Additional Key Words and Phrases: small group interaction, Supervised Auto-encoder, communicative functions, multimodal behaviors

1 INTRODUCTION

Small-group interaction is a pervasive face-to-face interaction form in our daily life. The unique mutual communicative process involving knowledge and responsibility-sharing between multiple parties provides a natural

Authors' addresses: Yun-Shao Lin
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan, yunshaolin@gmail.com; Yi-Ching Liu
, [2]Department of Business Administration, National Taiwan University, Taiwan; Chi-Chun Lee
[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan, cclee@ee.nthu.edu.tw.

1551-6857/2022/8-ART111 $15.00
https://doi.org/10.1145/3558768

working mechanism for a group to achieve a shared goal. The consensus view from small-group research is that this mutual interaction process plays an intermediary role in how a group performs [15, 21]. As a result, there is a growing interest in the study of small-group interaction directly from audio–video recordings of members behaviors using computational methods [1, 27, 36]. Computationally modeling small-group behaviors can help develop technological solutions toward automated task management and enhanced communication effectiveness [11, 37]. For example, providing a means for direct intervention by using a virtual assistant to encourage the beneficial communicative behaviors, which are often associated with good task performance and to prevent mal-behaviors, which are related to poor performance, can help in effective group performance management [29, 47].

In general, communication is the process of exchanging information, facts, opinions, or feelings. Previous studies have identified several important factors for studying group communication and group performance. The group structure can influence its performance because it establishes the predefined framing to restrict the information communicated between members [43]. A computer-mediated environment could degrade the quality of communication and result in a longer time to achieve consensus compared to face-to-face interaction [22, 40]. Although these environmental factors can impact the communication process and the final outcome, modeling the group performance merely using static factors is not enough. The dynamic interaction process is shaped by the unique verbal and nonverbal behavioral exchanges shown by group members while discussing their concerns, working towards plans and executing the solutions. While modeling the content of the communication process is intuitively appealing for automatic prediction of the group outcome, the interaction behaviors between members are a complex and intricate process results in only few studies working on time-grained analysis of the communication process for the prediction task. Therefore, we argue that it is important to design a proper computational tool to automatically assess the behaviors during communication process for time-grained analysis.

Communication is an essential means for group members to collectively understand a task and determine the corresponding shared goal. During the process, the communicative functions are important for each member to convey information during the conversations. For instance, they could express the similar information through different communicative channels such as gestures, facial expressions, speech and verbal acts depending on their different intentions. Therefore, the analysis of communicative functions often requires a short-termed and time-grained analysis, such as at the speaker's sentence level, in order to capture a speaker's intention toward others at every point in a conversation. As the communication function possesses rich information needed to advance the modeling of members' conversational behaviors, it provides an important view in studying small-group interaction. For example, the computational studies of communicative functions in small group scenarios have used conventional dialogue acts schemes to analyze members' communication skills [39], empathy skill level [23], and interpersonal reactivity scores [24]. However, to the best of our knowledge, there have not been any works on modeling the communicative function for outcome prediction. Furthermore, the usage of the dialogue acts annotation system in previous works lacks a supporting reference in small group research and underestimates the social aspect information involved in group interaction.

Bales' interaction process analysis (IPA) [2] is a classic coding scheme of communicative functions for group scholars to analyze the interaction pattern in small-group conversations because of its capability to adequately capture the intention of each member's utterance from an overall group perspective [21]. As a task group is formed, during the interaction, members in the group need to cooperate with each other to determine their collaboration pattern and working strategy to accomplish the assigned task. From Bales' perspective, solely performing the task-related action without social aspect actions would harm the interrelationships among members and make the group fall apart. In order to maintain group cohesion, socio-emotional action often appears between task-oriented actions to keep both socio-emotional needs as well as task fulfillment. Therefore, unlike many existing dialogue act coding frameworks such as the ISO 24617-2 standard [8], the IPA provides categories comprising socio-emotional perspectives such as "show antagonism", "show solidarity", "show tension", and "show tension released". [17, 19].

As a result, IPA has been the standard method for analyzing group interaction throughout the decades of the late 20th century [20, 52].

Based on the nature of the communication process and the definition of a communication function, the analysis results in the previous studies have empirically shown that the task-related communication actions are related to the group performance [13, 14, 18]. Therefore, we further assume that learning the group interaction outcome from communication functions like IPA is the key to solving the current bottlenecks in the performance of the small group performance prediction task. However, building such a learning framework can suffer from multiple problems. In the training phase, there is no public corpus annotated with the IPA labels, and the IPA annotation process of IPA would be time-consuming and involve considerable manual effort. In the testing phase, using the communication function for group score prediction also requires the testing data to be annotated with the corresponding communication function, but having the testing data manually labeled using IPA is not feasible in a real-world scenario. In order to overcome these issues, we first collect the Mandarin NTUBA dataset, which is specifically designed for studying how a group performs, and then every sentence from the member is annotated with Bales' IPA labels by trained coders. Using the NTUBA corpus, which is one of the largest scaled corpora (totally 151 groups) for annotating the largest number of IPA (totally sixty-thousands utterances), we design a two-stage learning framework that can predict the group score in the absence of IPA annotation in the testing data.

Specifically, we propose an Interaction Process Guided Framework, i.e., a two-stage process to integrate high-level IPA tags into low-level behavior features for group performance score prediction. The framework is based on first training a Supervised Auto-encoder (SAE) for automatic IPA prediction and then aggregating the embedding from this pre-trained network for group performance prediction. As our proposed supervised interaction process Auto-encoder (SIPA) is trained using reconstruction loss as well as classification loss, its learned output preserves the behavior information while embedding the IPA related information. As a result, unlike directly using low-level behavior features as input, our framework can leverage the IPA information and consequently perform better on the group score prediction task. We evaluate our framework on the two separated subsets in the NTUBA corpus, i.e., NTUBA-task1 and NTUBA-task2. The experiment result demonstrates the robustness of our proposed framework for the group performance prediction task by achieving a promising regression performance on both two subsets, i.e., 4.241 MSE in NTUBA-task1 and 3.794 MSE in NTUBA-task2 (averaged across 10 random seeds).

In order to comprehensively examine the performance of our proposed frameworks, we evaluate the proposed approach from two perspectives. First, we examine the performance of using the behavior feature from various behavior modalities in our framework. Body movement, facial expression, speech, and language are considered as the behavioral forms of the communication function. Second, we directly compare the proposed two-stage framework to a single-stage learning method. Our experimental result shows that our proposed two-stage framework outperforms the empirically best-performing model such as RandomForests and the SOTA end-to-end deep model such as graph-convolutional neural networks. Like a previous study [30], our results shows that our SIPA network can perform the early fusion between facial expression and language features for better capturing the communication process and thus achieving a better group performance prediction result in NTUBA-task1. The evaluation of the parameters for the SIPA network is further examined to identify the important working mechanism when learning the communicative behavior representation for the group outcome score.

Besides the improvement of the model performance, compared to previous computational methods [32, 36], our framework can provide a time-grained interpretation of the dynamic communication process by analyzing the relationship between sentence-level interaction behavior and group performance. Similar to the analysis result in a previous work [13, 14, 18], our analysis result also indicates that an effective communication process requires members to convey more actions about tasks and fewer actions related to social anxiety. Therefore, our model not only demonstrates a robust prediction performance but also provides a grounded interpretation on the communication process. Finally, we summarize the main contributions of our framework in this study:

- **Achieving SOTA Group Score Prediction** Compared to other SOTA methods, our framework demonstrates robust performance on group score prediction task across two subsets in the NTUBA corpus. In our analysis, although two subsets manifest with diverse interaction patterns, our framework can achieve the best performance on both of the conditions.
- **Introducing the Two-Stage Framework for Integrating Communicative Function** We are the first to introduce a two-stage framework, which effectively leverages the information of the communication function, for group score prediction. Furthermore, the proposed framework can perform the group score prediction in the absence of IPA annotation for testing data in the real-world scenarios.
- **Time-Grained Analysis on Group Communication Process** With the usage of IPA predicting network, we could leverage the advantage of IPA on the interpretation of interaction behavior without additional effort. This interpretability provides an opportunity to manage the team interaction or build a direct intervention to members' behavior

In Section 2, we discuss related works on two topics, including the computational method for predicting group task performance and multimodal group interaction corpus. In Section 3, we describe how we design our corpus and process our label. We also describe the feature extractor used in our experiment and how our two-stage framework works. Section 4.1 explains the experimental setup, model parameters, and other compared modeling methods. Experiment results of our proposed method and the comparison between the baseline result and STOA result are presented in Section 4.3. The analysis of different parameters are shown in Section 4.4 and the robustness of our framework are discussed in Section 4.5. Further analyses on the interpretability of our proposed method are presented in Section 4.6. We finally conclude our work by discussing the limitations and future directions in Section 5.

2 RELATED WORK

2.1 Computational Modeling of Group Task Performance

Several recent works have demonstrated the feasibility of predicting group performance by computing the members' verbal and nonverbal behavior features during an interaction. By defining a variety of hand-crafted features, including personality traits from questionnaires, individual performance on the ranking task, visual gaze, and individual and group speaking cues, Avci and Aran [1] were the first to build a computational framework for group performance prediction using the ELEA (Emergent LEader Anal-ysis) corpus. By emphasizing the usage of the language feature, Murray and Oertel [36] compared the performance of various models with diverse verbal and nonverbal feature sets. With the emergence of a series of multimodal corpora [27], Kubasova et al. were the first to study the usage of verbal and nonverbal features on both the ELEA and GAP (Group Affect and Performance) corpus. They further proposed an effective graph-based feature for characterizing the conversation structure between members [26] on the ELEA, GAP and UGI (Unobtrusive Group Interaction) corpus. However, using hand-crafted features is often limited to the domain of the task, using contemporary deep neural network-based learning methods is a new approach for predicting the group performance [32, 55]. Lin and Lee used the graph convolutional network with the inter-speaker conversational dependency for group performance prediction [32]. Zhong et al. integrated the group-composite personality traits with the attention mechanism for improving the modeling on members' behavior and thus improved the task score prediction [55]. However, all of these works emphasize only the lower-level in-conversation behaviors, e.g., acoustics, turn-taking, and duration of speaking length. During such a complex interaction between group members, we believe higher-level constructs such as communicative function play a key role in shaping the group dynamics because they fundamentally captures the

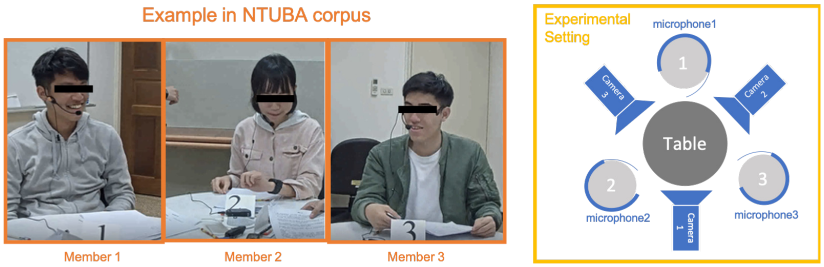

Fig. 1. NTUBA corpus: NTUBA is a large-scale multimodal small group interaction corpus. As the left part of figure shown, there are 3 persons in each group and they are recorded by 3 separated camera, headset microphone and physiology sensor on their wrist during the interaction. The right figure clearly illustrate the environmental setting in our dataset.

	NTUBA-Task1		NTUBA-Task2	
	$Score_{raw}$	$Number_{group}$	$Score_{raw}$	$Number_{group}$
task-oriented reflect.	35.13±13.03	24	29.84±9.73	22
emotion-oriented reflect.	32.5±14.61	31	32.21±6.95	29
mixed reflect.	38.55±13.09	23	34.68±4.33	22
All Dataset	35.09±13.75	78	32.24±7.45	73

Table 1. Reflection Condition and Corresponding Group Score in NTUBA corpus

information about every action used in this back-and-forth communication patterns. In fact, previous studies have indicated the correlation relationship between these actions and group performances [13, 14, 18].

2.2 Multimodal Corpora with Group Task Performance

A series of multimodal corpora have recently emerged for studying cause and effect between the members' behaviors during small-group interactions and group performance from the interaction outcome, e.g., Mission Survival Corpus [41], ELEA [44], GAP [7] and UGI [5]. However, most of the existing public data for studying the group performance have been collected under the protocol of survival games, which is an *intellective task* as per McGrath's Task Circumplex [35]. Given the fact that the small-group interaction process is highly influenced by conditions such as the member's backgrounds and the tasks they perform, a survival task (intellective task) is often dominated by a single member with strong background knowledge about the task [6]. Unlike the aforementioned multimodal database, our NTUBA have three advantages. First, it is the largest-scaled multimodal Mandarin small-group interaction corpus to date. NTUBA, which includes two NTUBA-task1 and a repeated task (NTUBA-task2), also has the largest number of groups, i.e., 78 groups, which is more than the total sum of the ELEA, GAP, and UGI corpus. Second, unlike the conventionally used survival task scheme, tasks in NTUBA belongs to the *choose quadrant* as per McGrath Task Circumplex. In our task, the members need to communicate to understand each other's needs and limitations for coordinating and designing the best plan to achieve the best group score.

3 METHODOLOGY

3.1 The NTUBA Database

The NTUBA database[1] is a Mandarin multimodal corpus, which includes audio, video, and physiology signals, collected at the College of Management of the National Taiwan University (NTU). A task-oriented reflection

[1] Approved by IRB: REC-201901HS021

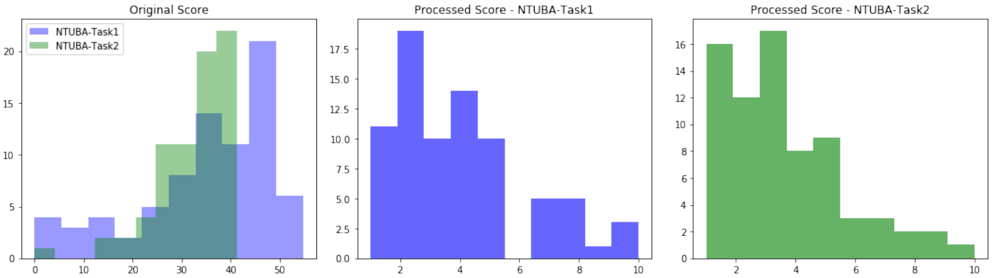

Fig. 2. Distribution of original score ($Score_{raw}$) and processed score ($Score_{group}$) in NTUBA dataset. The blue bar represent the distribution of NTUBA-task1 and the green bar represent the distribution of the the NTUBA-task2.

process for members to discuss how they achieved the shared goal in the previous session has been suggested as an effective intervention to improve the group task performance, as group members can learn from the experience and fix any problems they encounter using the corresponding strategy[45, 51, 53]. In addition to the task-oriented problems, a previous study also points out that emotion-oriented problems also have a critical impact on group performance [9, 38]. In order to determine whether emotion-oriented reflection would affect small-group interaction dynamics and task performance, we design the NTUBA with three diverse types of reflection processes, including task-oriented reflection and the other two experimental groups includes the emotion-oriented reflection and mixed reflection that is a combination of emotion-oriented and task-oriented reflection.

In the database, subjects are assigned a shopping-trip planning task [54]. The shopping-trip task requires participants to buy the best-quality groceries at the lowest prices in the least amount of time by discussing with each other. The task is time-limited with budget constraints, and groups are requested to shop at multiple locations, such as the supermarket, the bakery, and so on, each at varying distances from their house. As the nature of the task belongs to the “choose" quadrant as per the McGrath Task Circumplex, it requires members to understand each other’s needs (e,g., items to buy) and limitations (e.g., time available to shop together) in order to coordinate and design the best plan. The process can be divided into six parts for each group, including Rule explanation, Task1 (pre-intervention), Mid-Survey, Reflection (intervention), Task2 (post-intervention), and End-Survey. The experimenter first explains the rules. Then, the first shopping task begins with a time limit of 30 minutes, followed by a session where everyone complete a questionnaire about their emotions during the process and team-related aspects such as communication, efficacy, and cohesion. Next, in the Reflection part, the groups reflect on how to improve their scores. The second shopping task starts later with a time limit of 20 minutes, followed by the end-point questionnaire in which they are asked to evaluate their reflection process, emotions during the second task, and team-related aspects such as efficacy and cohesion, again.

In order to fully capture members’ behavior during the interaction process, every speaker is recorded using the Bluetooth-headset-microphone and fixed-camera to capture a clear frontal image of their face. Every utterance is manually segmented, labeled using an IPA annotation, and transcribed to the text. In our work, we use speech, video and text both separately and jointly for our framework. Task1 includes a total of 25.75 hours of audio recordings of 78 groups. Task2 includes a total of 18.18 hours of audio recordings of 73 groups. The difference in group numbers between task1 and task2 is because of unexpected failure in the recording process (such as the problem of running out of the battery, the recorder being shut down by the participant accidentally, or the microphone being attached in the wrong direction).

3.1.1 Group performance score. The group performance is scored by two trained coders using an objective grading policy provided by the scholar who developed the task [54]. Each group’s score is independently calculated

	NTUBA-task1		NTUBA-task2	
	num	**ratio**	**num**	**ratio**
agree	3626	0.106	2518	0.097
ask for opinion	1344	0.039	1155	0.045
ask for orientation	4886	0.143	3468	0.134
give opinion	2792	0.081	2661	0.103
give orientation	17986	0.527	13606	0.525
give suggestion	3014	0.088	2151	0.083
other	83	0.002	49	0.002
show tension	272	0.008	162	0.006
show tension release	116	0.003	123	0.005
total	34119	1	25893	1

Table 2. Distribution of nine IPA annotations in NTUBA corpus: We group the rare occurred IPA classes, i.e., disagree, show antagonism, show solidarity and ask for suggestion, as *"other"*.

by two coders and then discussed to identify their discrepancies. We then use the reconciled scores in our analysis. In table1, we present the original group $Score_{raw}$ according to three experimental conditions, i.e., task-oriented condition, emotion-oriented condition and mixed condition. We list the original task scores under three conditions in Table 1. Given that team reflection at the mid-point was not our focus - rather, our focus was on building a computational framework to predict group score. We used z-normalization to neutralize the effect of the three different types of reflection on group scores. In other words, for each condition, we derive our $Score_{z-norm}$ by using $\frac{Score_{raw}-Mean(Score_{condition})}{Std(Score_{condition})}$. After processing the group score with z-norm within three condition groups, in order to evaluate different subsets with the same scale, the processed group score is re-scaled by using MinMaxScaler from 1 to 10 following a previous work [27] to obtain our final group score $Score_{group}$. Overall, the mean and std of the processed score are 3.99±2.2 for task1 and 3.67±2.03 for task2. The re-scaling procedure provides the opportunity to examine the prediction result among different corpora and different tasks. Similar to a previous study [27], the lower $Score_{group}$ implies better task performance. Figure 2 summaries the distribution on original score ($Score_{raw}$) and processed score ($Score_{group}$) in task1 and task2.

3.1.2 IPA annotation. To code the IPA labels [2], two trained raters watch the video of each interaction session and then annotate each utterance using one of the 12 IPA tags. The inter-rater consistency of preliminary annotation is 0.56 (Cohen's kappa), but raters discussed any discrepancies between their codes and obtained reconciliations. In total, there are 34119 utterances labeled with IPA tags in the NTUBA-task1 and 25893 utterances annotated with IPA tags in the NTUBA-task2. Based on the annotation result, we found that four of the IPA categories, i.e., "disagree", "show antagonism", "show solidarity" and "ask for suggestion", occurred rarely. In order to mitigate the issue of annotation imbalance, we group these four rare occurred IPA labels as "Other" and retain the other eight categories because they include a sufficient number of data samples (over 0.1%). Table 1 presents the distribution of the final nine IPA categories in two NTUBA subset used in our framework.

3.2 Interaction Process Guided Framework

Our proposed computational framework (fig.4) is a two-stage process for integrating the communication function as a robust interaction representation to perform group performance prediction. In training stage 1, depending on the members' speaking time, the sentence-level behavior feature x_{input} is first extracted from the raw signal

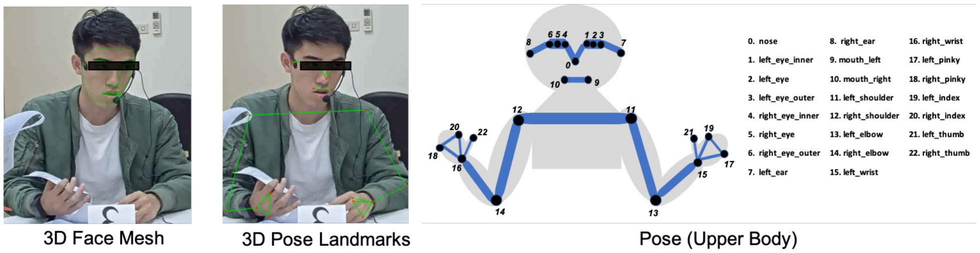

Fig. 3. Detection Result in our NTUBA corpus: the left figure shows the pose and face detection result by using the MediaPipe toolkit for the speaker 28_3; the right figure shows our selected 23 pose landmarks for upper body.

by using the pre-trained network or signal processing methods. We then train a Supervised Interaction Process Auto-encoder (SIPA) to classify the corresponding sentence-level IPA tags. In the training stage 2, we simply average the sentence-level *Emb* from the output of the SIPA in stage 1 as a group-level feature and feed it through a Ridge regressor to learn the final group task score.

3.2.1 Multimodal Features. In order to effectively capture how the group members communicate with each other during the interaction, we compute the language, speech, face and pose features from their expressive behaviors at every utterance and use them as the input for our model.

- **Language:** $\boldsymbol{Emb_{bert}}$
  Bidirectional Encoder Representations from Transformer (BERT) [10] is a self-supervised language representation model proposed by Google. It's unique multi-head self-attention mechanism enables it to attain the state-of-the-art results on various NLP tasks. In our work, the Emb_{bert} are extracted from the "bert-based-chinese"[2] released by the huggingface's Transformer toolkit. We extract the last layer from the pretrained BERT model and the pretrained BERT embedding is trained with 768 dimensions. As our sentence embeddings Emb_{bert} are derived by taking the average on all word embeddings in the same sentences, Emb_{bert} also has 768 dimensions.
- ***Speech*** : $\boldsymbol{IS10_{std-76}}$
  We use the same feature set as the previous work [27]. This acoustic feature set, including jitter, shimmer, mel-frequency cepstral coefficients (MFCCs), associated delta features, PCM loudness, F0 envelope, F0 contour, and voicing probability, are extracted by using the Opensmile[3] toolkit. This feature set corresponds to a commonly used configuration file that was developed for the INTERSPEECH 2010 Paralinguistic Challenge, called IS10. $IS10_{std-76}$ is a subset of IS10. As the previous work [27] mentioned, group studies often only include a small number of observations. In order to reduce the dimensions of a large number of overly redundant features, we select only the standard deviation features from the original set. This results in a final set of 76 speech features.
- **Face: 3D Face Mesh from MediaPipe**[4]
  The Face Mesh from Google's MediaPipe toolkit estimates 468 face landmarks for each face in a video frame. The method is composed of two models. First, a face detector (BlazeFace model[4]) is used on the full image and computes face locations. Second, depending on the face location, a 3D face landmark model [25] predicts the approximate surface geometry via regression. For each landmark, there are three predicted values, i.e., x, y, and z (depth), which results in 1404 dimensions in each frame from the original prediction score. In order to capture the member's facial expression during speaking, we first segment

[2] https://huggingface.co/bert-base-chinese/tree/main
[3] https://www.audeering.com/opensmile/
[4] https://google.github.io/mediapipe/

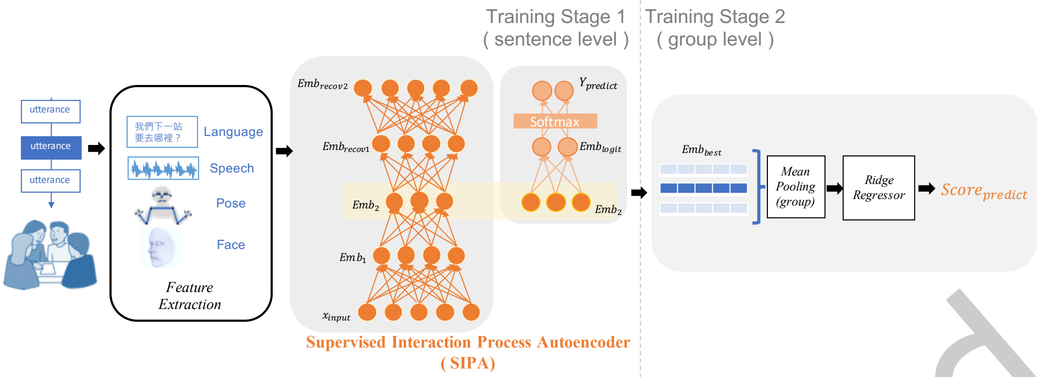

Fig. 4. Our Proposed Computational Framework with Supervised Interaction Process Auto-encoder: the proposed framework includes 2-staged training. In the first training stage, we train a sentence-level IPA prediction network with both reconstruction loss and classification loss. After having a well-trained model in stage 1, we aggregate the embedding vector from SIPA network for the final group performance prediction.

the video according to the start and end time of each utterance. After using MediaPipe's Face toolkit to extract the feature in each frame, we take the utterance-level mean and std as the representation of the facial expression. This results in 2808 dimension features, and figure 3 shows an example of a detection result in our dataset.

- **Pose: 3D Pose Landmarks from MediaPipe**
  Pose Landmarks is another function from MediaPipe released by Google. It infers 33 3D landmarks and a background segmentation mask on the entire body from RGB video frames by utilizing the BlazePose model [3]. For each landmark, there are three predicted values, i.e., the x, y, and z (depth), which results in 99 dimensions in each frame from the original prediction score. As our video only records member's upper bodies, we only select 23 of the original 33 landmarks, which results in 69 dimensions for each frame. With the same processing method as Face, we first segment the video according to the start and end time of the utterance. After using the toolkit to extract the feature in each frame, we take the mean and std on frames, which belong to the same utterances, as the representation of the body movement. This results in 138 dimension features per utterance. Figure 3 shows an example of a detection result and the details of the 23 upper-body landmarks in our dataset.

3.2.2 Stage 1: Supervised Interaction Process Auto-encoder (SIPA). An Auto-encoder is a common and effective non-linear method for compressing the data into a lower dimension while preserving the information from the original high-dimension space. Le et al.[28] have theoretically demonstrated that the Supervised Auto-encoder (SAE) network, which jointly trained by the supervised classification task and the reconstruction loss, is an efficient way to achieve model generalization. The network structure has been applied to different tasks, including the vowel classification, image classification task, and particle physics process prediction task, and the method empirically shows a good generalization performance [28].

In our work, as the nature of task computes action and socio-emotional functions that often involves a high level of ambiguity and overlap between multiple classes, we leverage the SAE as our model to learn the IPA enriched behavior representation with multi-views to avoid the model being overfitted on the data. In other words, by using SIPA as the first stage training, we expect our model not only to keep the information from the original behavior feature but also to enrich the embedding with IPA information. In general, a two-layer encoder-decoder structure with an additional single-layer classifier is trained with the reconstruction loss according to the input Emb_{bert} as well as the cross-entropy loss according to the annotated nine IPA classes. Similar to previous work on IPA prediction task [30], we also apply various class-weights (W_k) to mitigate the issue of imbalance class. Specifically, we build our SIPA model with the encoder-decoder structure as follows:

$$Emb_1 = ReLU(W_{encode1} * x_{input} + b_{encode1}) \tag{1}$$

$$Emb_2 = ReLU(W_{encode2} * Emb_1 + b_{encode2}) \tag{2}$$

$$Emb_{recov1} = ReLU(W_{decode1} * Emb_2 + b_{decode1}) \tag{3}$$

$$Emb_{recov2} = ReLU(W_{decode2} * Emb_{recov1} + b_{decode2}) \tag{4}$$

The output of the encoder, which is Emb_2 in (2), is also used for obtaining the unbounded logit value on nine IPA classes by passing through an additional layer,

$$Emb_{logit} = W_{classify} * Emb_2 + b_{classify} \tag{5}$$

Finally, we pass the logit value through softmax to obtain the predicted Y,

$$Y_{predict} = Softmax(Emb_{logit}) \tag{6}$$

During the training process, we apply a dropout layer in (1) and (3) to prevent overfitting. The classification loss and reconstruction loss are calculated in batch-wised with batch size N and class number $K = 9$:

$$Loss_{recons.} = \sum_{n}^{N} MSE(Emb_{recov2}, x_{input}) \tag{7}$$

$$Loss_{class.} = \sum_{n}^{N} \sum_{k}^{K} -Y_{true}^{k} \log(Y_{predict}^{k}) * W_k \tag{8}$$

The class weights W_k are derived by:

$$W_k = \frac{Num_{utt}}{K * Num_k}$$

, where Num_{utt} is the total number of utterance, IPA class number $K = 9$ and Num_k is the number of utterances for class k. Finally, we sum up both two losses jointly with $\lambda = 0.5$ to update our network:

$$Loss_{total} = (1 - \lambda) * Loss_{recons.} + \lambda * Loss_{class.}$$

3.2.3 Stage 2: Group Score Prediction . Average pooling is a prevalent and effective strategy to aggregate the sequential features. For example, the well-known acoustic feature extraction toolkit Opensmile heavily relies on the mean pooling to aggregate frame-level acoustic features and the experimental result demonstrates the efficacy of this mean-pooling strategy on the performance of many para-linguistic tasks. For the group score prediction task, Subburaj et. al. [46] also show that the straightforward equally weighted strategy is the most effective way to aggregate the behavior between different members compared to other weighted methods. Therefore, after stage 1, we take the average of the trained sentence-level embedding Emb from our nine-class SIPA network as our group-level features, which can be thought of as behavior representation enhanced with information of the communicative and socio-emotional functions. We train a Ridge regressor on the training set to perform the final group score prediction on the testing set. Given that we did not assume that any specific layer in SIPA model is the best-performing embedding Emb_{best}, our strategy is to decide the Emb_{best} based on the predicted performance on the validation set. We then apply Emb_{best} on the testing set.

$$Score_{predict} = Ridge(Mean_{group}(Emb_{best})) \tag{9}$$

4 EXPERIMENT SETUP AND RESULT

4.1 Experimental Setup

4.1.1 Experiment. NTUBA is a multimodal corpus that includes fruitful expressive behaviors of group members, we aim to study the efficacy of using behavior features from multi-modalities for the group scores prediction. Our evaluation comprises of two separate parts, i.e., "Baseline" and "Proposed" methods. For the "Proposed" part, it is the proposed framework mentioned in Section 3.2 (the two-stage communication function enhanced learning method). For the "Baseline" part, we use the empirically best-performing RandomForest [27] as the single-stage learning method with group level behavior features. The group level features in single-stage method are different from those in our proposed method. Specifically, unlike two-staged methods, they directly takes the average on all sentence level features (mentioned in Section 3.2.1) as group level features and learns the group score from it.

Both "Baseline" and "Proposed" are composed of "Unimodal" and "Multimodal" parts. In the "Unimodal" part, four different modalities, i.e., language, speech, face and pose, are used separately as input features. For the "Multimodal" part, we perform the early fusion to jointly integrate the information from multiple modalities for the group performance prediction task. Among four different modalities, using language as features captures the communication process best as compared to other modalities, and it achieves the best results compared to previous works [30]. Therefore, we design three different combinations by using language with three other modalities, i.e., "language + speech", "language + face", and "language + pose".

In addition to the three multimodal combination pairs, in the "Baseline" part, we further compare with the previous study [27] by reproducing the best-performing feature proposed by Kubasova et al. [27]: (Emb_{bert} + $IS10{-}_{std-76}$ + $Ling$). In that study [27], the authors includes the linguistic feature from multiple aspects: spaCy's [5] Dependency Parse Features (bag-of-relations (type-token ratios), branch factor, and maximum branching factor), Bag of spaCy's Part-of-Speech Tags (type-token ratio), count of agreement words including "好(ok)" and "對(right)", count of hesitation words like "呃(hmm)", and "阿(huh)" and Sentence Length. These handcrafted linguistic features result in a total of 48 dimensions for each group. Concatenation of the aforementioned speech($IS10{-}_{std-76}$) and language(Emb_{bert}) features result in 892 dimensions.

4.1.2 Settings. In our work, two NTUBA subsets, NTUBA-task1 and NTUBA-task2, are examined as separated datasets. In order to validate our framework, we use five-fold cross-validation scheme in our experiment. In each fold, we split all groups in each dataset in the ratio of 3:1:1 to create the training set, validation set, and testing set. As our work is a two-stage framework, we make sure the two stages have consistent training, validation, and testing sets. In other words, each of the SIPA models is followed by its corresponding regression model. Therefore, the testing data is completely isolated from training data in each fold under this two-staged training.

In each fold, the training set is trained with the same model parameter mentioned in Section 4.1.4. For the SIPA model, in order to prevent overfitting on the training data, we stop the training process with the best-performed epochs based on the validation loss in each fold. The best-performed embedding (Emb_{best}) from the proposed SIPA model is selected based on the predicted performance on the validation set in the second stage. The final evaluation result is obtained by aggregating all the testing sets from all five folds. We evaluate our method using Mean Square Error (MSE) and Pearson's correlation (corr). We perform this experimental scheme with ten different random seeds and present both evaluation metrics with "mean ± std" on the testing data.

[5] https://spacy.io/models/zh

4.1.3 Model Parameters. We use PyTorch to implement our SIPA model. Different behavior feature sets with the corresponding input dimension X_{dim} are trained with similar network parameters. The Auto-encoder part of our SIPA model is built with a symmetric structure, where the encoder has $W_{encode1}$=[X_{dim}, 128], $W_{encode2}$=[128, 128] and $b_{encode1}$=[128] , $b_{encode2}$=[128], the decoder has $W_{decode1}$=[128, 128] , $W_{decode2}$=[128, X_{dim}] and $b_{decode1}$=[128] , $b_{decode2}$ = [768]. For the classification part, we have the parameters $W_{classify}$=[128, 9], $b_{classify}$=[9]. All parameters are trained with batch size=8, learning rate=0.01, dropout rate = 0.5 and epochs=40. We set the hyperparameter λ as 0.5 for training our model equally with both reconstruction loss and cross-entropy loss.

4.2 Experimental Result

Overall, our two-stage framework consistently achieves a robust predicting performance across two separate subsets in NTUBA. Specifically, our proposed method achieves the best performance with 4.241 MSE and 0.341 corr in NTUBA-task1 and 3.794 MSE and 0.273 corr in NTUBA-task2 (averaged across ten random seeds) for the group task score prediction task. From Table 3, we observe that our two-stage framework surpasses the performance of the conventional single-stage learning method for group score prediction task on both metrics, i.e., MSE and p-corr. Among different behavior modalities, we show that language is the most effective feature compare to other modalities, and our framework can jointly fuse the face and the language features to achieve the best performance for the NTUBA-task1.

According to Table 3, we can find that the result between task1 and task2 shows significant differences in their prediction performance, which are 4.241 MSE for task1 and 3.794 MSE for task2. We find it reasonable for the following two reasons. First, they are different in scenarios which designed with different collection protocols. For instance, instead of having the same execution time as task1, which is 30 minutes, the participants only have 20 minutes to finish task2. The differences are also evident in the group score. According to the distribution of group scores in figure 2 and the corresponding mean and std value (3.99±2.2 for task1 vs. 3.67±2.03 for task2). Second, since task2 is the activities followed by a behavior intervention (reflection stage) and the participants in task2 naturally are more acquainted with each other than in task1, it is reasonable to assume that they would behave distinctively differently between the tasks.

4.2.1 Unimodal part. In Table 3, with the proposed framework and baseline method (RandomForest), using the language feature (Emb_{bert}) leads consistently to the best performance among four different modalities in NTUBA-task1 and NTUBA-task2. The trend of the performance on four different modalities also displays a similar pattern, where the performance has the following order: *Language* > *Face* > *Pose* > *Speech*. Similar to the result in experiment 1, our framework generally performs better than the single-stage learning method regardless of the modalities. For example, in NTUBA-task1, our frameworks on average improves 0.53 MSE and 0.145 corr for language, 1.472 MSE for speech, 0.553 MSE for face, and 0.903 MSE for pose; in NTUBA-task2, our frameworks on average improves 0.226 MSE and 0.059 corr for language, 1.218 MSE for speech, 0.808 MSE for face, 1.309 MSE for pose. We find the result reasonable and intuitive because the language features directly contain task-oriented and interaction-oriented information during the interaction process [16, 34, 42]. For example, the number of task-related words provides direct evidence of group progress [26] and the linguistic style could even provide an indicator of the team's social situation [12].

4.2.2 Multimodal part. In Table 3, among the three different multimodal fusion pairs mentioned in Section 4.1.2, fusing the language with face feature demonstrates a better performance for both the baseline method and our proposed method in NTUBA-task1. Specifically, our experiment shows that "language + face" can achieve better results than using only "language" and can achieve the best result at 4.241 MSE and 0.341 corr for NTUBA-task1. The improvement based on the face modality is similar to the previous study [30], which suggests that the communication function is highly expressed by using language but can be slightly improved

		NTUBA-task1		**NTUBA-task2**	
	Modality	mse	corr	mse	corr
Baseline					
Unimodal	Language	4.824±0.318	0.18±0.071	**4.02±0.277**	**0.214±0.087**
	Speech	6.498±0.349	-0.239±0.08	5.438±0.495	-0.141±0.093
	Face	5.418±0.335	0.106±0.075	4.984±0.417	0.015±0.098
	Pose	5.679±0.387	0.071±0.082	5.506±0.479	-0.081±0.098
Kubasova et. al.	Lang.+Speech	4.898±0.322	0.166±0.07	4.239±0.277	0.135±0.091
Multimodal	Lang.+Speech	4.864±0.331	0.175±0.073	4.285±0.223	0.13±0.076
	Lang.+Pose	4.762±0.38	0.198±0.082	4.063±0.226	0.196±0.08
	Lang.+Face	**4.691±0.433**	**0.221±0.101**	4.049±0.326	0.201±0.102
Proposed					
Unimodal	Language	4.294±0.219	0.325±0.062	**3.794±0.199**	**0.273±0.074**
	Speech	5.026±0.106	-0.061±0.052	4.222±0.128	-0.128±0.082
	Face	4.865±0.09	-0.081±0.11	4.176±0.104	-0.11±0.08
	Pose	4.776±0.093	0.076±0.087	4.197±0.105	-0.133±0.076
Multimodal	Lang.+Speech	4.414±0.169	0.28±0.057	4.138±0.214	0.101±0.092
	Lang.+Pose	4.313±0.294	0.317±0.087	3.835±0.177	0.251±0.071
	Lang.+Face	**4.241±0.165**	**0.341±0.059**	4.075±0.124	0.121±0.066

Table 3. Result of Experiment: Comparison between our proposed 2-staged methods and the 1-staged RandomForest model as baseline methods.

	NTUBA-task1		**NTUBA-task2**	
	mse	corr	mse	corr
Mean Guess	4.879±0.079	-0.117±0.093	4.177±0.104	-0.113±0.08
RandomForest	4.824±0.318	0.18±0.071	4.02±0.277	0.214±0.087
Ridge	5.519±0.839	0.271±0.109	4.571±0.381	**0.291±0.049**
DNN	4.934±0.111	-0.104±0.075	4.169±0.109	-0.045±0.126
GCN	4.957±0.15	-0.087±0.119	4.145±0.248	0.098±0.114
Transformer	4.986±0.195	-0.062±0.114	4.284±0.281	-0.062±0.146
proposed	**4.294±0.219**	**0.325±0.062**	**3.794±0.199**	0.273±0.074

Table 4. Result of SOTA learning methods: With language feature (Emb_{bert}) as input, we compare our proposed framework with different learning methods, including MEAN GUESS, RandomForest and other SOTA end-to-end networks

by additionally considering facial expression. Although the multimodal fusion could gather the information from different behavior facets, our experiment result points out that it can not always successfully improve performance, because we did not observe a similar improvement in NTUBA-task2.

4.3 Comparison with SOTA learning Methods

4.3.1 SOTA methods. Our results in Section 4.2 imply that language features works as a more robust feature for achieving the best results. We further use the language feature (Emb_{bert}) as a benchmark feature and compare our proposed two-stage method with other SOTA methods, including tree based and other end-to-end learning methods. Specifically, we compare our method with,

- **MEAN GUESS**
  As the distribution of the group scores is similar to Gaussian distribution, conventionally, using the mean of the labels in the training set as a value to predict the group score in the testing set is considered as a chance baseline. This method does not need the x_{input} as a feature to predict.
- **Ridge Regressor**
  Ridge regressor is a simple linear regresssion model with l2 regularization. As we use Ridge regression in our framework, we also compare the same regression model with Emb_{bert} as input features. Ridge Regressor is trained with regularized parameters $alpha = 0.01$.
- **RandomForest**
  Tree-based regressor is an immediate and effective way to achieve good prediction performance and is commonly used in previous works [27, 36] with various hand-crafted conversational features. We set the number of estimators as 20.
- **DNN**
  A three-layer DNN with single-stage training for group performance score prediction task is compared with our proposed framework. The person-level Emb_{bert} are first averaged to obtain the group-level features as input for the network.
- **GCN [32]**
  We reproduce a recent state-of-the-art end-to-end network for group performance prediction, i.e., two-layer Conversational Graph Convolutional Network with person-level Emb_{bert} as input features. This method models inter-speaker dependency during the conversation with a graph network structure.
- **Transformer [50]**
  Transformer has achieved outstanding performance on multiple sequential modeling. In order to compare to this SOTA neural architecture's performance on this task, we use a two-layer Transformer encoder with the person-level Emb_{bert} as input features for group score prediction. Specifically, each timestep in the Transformer model is one of the members in a group. Unlike the GCN [32], the Transformer automatically learns inter-speaker dependency.

4.3.2 Comparison Result. In Table 4, the overall result indicates that the proposed framework surpasses other SOTA methods when tested on the same feature (Emb_{bert}) as input. Our method demonstrates a substantial improvement over the RandomForest. Specifically, the proposed method outperforms RandomForest with 0.53 MSE in NTUBA-task1 and 0.226 MSE in NTUBA-task2. Although our SIPA model is trained in deep learning fashion like DNN, GCN and Transformer, our framework outperforms those one-stage end-to-end training methods without considering IPA in the network.

As all SOTA learning methods are single-stage methods that directly learn the group score without considering the communication function, they perform poorly as demonstrated in our experiments. For example, even the result of the best performing RandomForest relatively improves by 1% in NTUBA task1 and relatively by 3.5% in NTUBA task2 compare to the mean guessing baseline. These results suggests that directly learning the group score from low-level features is not effective. Although deep-learning-based methods have been shown to outperform most conventional learning methods in multiple domains like computer vision and speech recognition, limited by the number of data samples in group score prediction task, the empirically best performing RandomForest still outperforms the deep learning methods like DNN, GCN and Transformer in our NTUBA dataset. Instead of using deep learning method directly for group score, we believe using it to capture the communication function in each utterance (that includes a more sufficient amount of data samples) would better capture the information of the communication process.

		x_{input}	Emb_1	Emb_2	Emb_{logit}	$Y_{predict}$
NTUBA-task1	valid	5.404±0.891	4.927±0.846	4.626±0.722	4.398±0.76	**4.32±0.795**
	test	5.746±0.937	5.497±0.895	5.081±0.624	4.577±0.479	**4.241±0.165**
NTUBA-task2	valid	4.56±0.773	8.053±1.958	6.206±0.702	4.473±0.85	**3.858±0.598**
	test	4.571±0.381	7.407±1.34	5.84±0.695	4.715±0.512	**3.794±0.199**

Table 5. Analysis the strategy of choosing Emb_{best}: we compare group score prediction performance in stage 2 with different *Emb* from SIPA network in stage 1

	NTUBA-task1			**NTUBA-task2**		
Encoder	DNN	SIPA	AE	DNN	SIPA	AE
λ	1.0	0.5 (proposed)	0.0	1.0	0.5 (proposed)	0.0
Stage 1: IPA classification result						
F1	0.538±0.033	0.533±0.033	0.141±0.08	0.608±0.009	0.617±0.019	0.124±0.045
ACC	0.514±0.034	0.509±0.033	0.127±0.07	0.594±0.01	0.605±0.017	0.106±0.028
Stage 2: Group Score Prediction						
MSE	4.282±0.219	**4.241±0.165**	4.873±0.083	3.805±0.206	**3.794±0.199**	4.173±0.104

Table 6. Result of different encoder: We show the comparison result between the performance of different λ value ($\lambda = 0.0, 0.5, 1.0$) corresponding to 3 different models (Auto-encoder(AE), SIPA (Supervised Interaction Process Auto-encoder), and DNN). The weighted F1 (F1) and Accuracy (ACC) are used to evaluate the performance in stage 1 (IPA classification task) and mean square error(MSE) are used to evaluate performance in stage 2 (group score prediction task).

4.4 Analysis of Model Parameters

4.4.1 Strategy for choosing Emb_{best}. We further examined strategy of selecting the Emb_{best} in stage 1 for both NTUBA-task1 and NTUBA-task2. Under the five-fold cross-validation scheme, we evaluate the performance using four embeddings extracted from different layers of the SIPA network, denoted as Emb_1 in (1), Emb_2 in (2), Emb_{logit} in (5), $Y_{predict}$ in (6), and compare the result with the original input feature x_{input} that is not passed through our network. We present our result in Table 5. x_{input} represents the concatenation of "Language + Face" features for NTUBA-task1 and "Language' features for NTUBA-task2. From table5, we can first observe that our strategy for identifying the best-performing embedding according to the validation result works well for both NTUBA subsets as the validation set and the testing set show a consistent pattern on the prediction performance. In addition, based on the analysis, we summarize the results into two important points. First, the embedding in our SIPA gradually performs better when they pass through different layers, i.e., the embeddings like Emb_{logit} and $Y_{predict}$ extracted from the classification part of the SIPA model are better than the embeddings like Emb_1 and Emb_2 extracted from the encoder part. Second, although both the Emb_{logit} and $Y_{predict}$ already outperform the result of using the x_{input} as feature, using $Y_{predict}$ empirically demonstrates consistently better results in our framework.

4.4.2 Encoder in Stage 1. As the encoder acts as an important role in our two-stage framework, it is crucial to compare the proposed SIPA with other deep learning-based methods. Here, we further compare our proposed method with the other two candidate models, including DNN and Auto-encoder (AE), as the utterance-level encoder. In other words, according to the equation of $Loss_{total}$, we further examine how reconstruction loss and classification loss contribute to our final prediction result individually. For DNN, it corresponds to set the hyper-parameter λ as 1. For AE, it corresponds to set the hyper-parameter λ as 0. The prediction results are summarized in Table 6 with the original weighted F1 and Accuracy score of the performance in nine IPA classification tasks.

	NTUBA-task1		NTUBA-task2	
Regressor	mse	p-corr	mse	p-corr
Ridge (proposed)	4.294±0.219	0.325±0.062	3.794±0.199	0.273±0.074
RandomForest	5.077±0.547	0.194±0.092	4.288±0.352	0.205±0.095
DNN	4.79±0.366	0.141±0.138	4.201±0.249	0.088±0.109
GRU	4.953±0.089	-0.147±0.059	4.613±0.432	-0.072±0.123
Transformer	5.004±0.212	-0.076±0.102	4.289±0.241	-0.02±0.132

Table 7. Result of different regressor: we compare the prediction performance of using the input $Y_{predict}$ with different regressors, including Ridge, DNN, and Transformer.

In general, we can see that our proposed method with two equally contributing losses result in the best-performing model compared to the other two models. Furthermore, we know that two different losses can improve the performance in their own way compared to using original x_{input} feature. For $\lambda = 1$, it achieves an average of 4.282 MSE for task1 and 3.805 MSE for task2. The result outperforms using Emb_{bert} for the same Ridge regressor by 1.464 MSE on task1 and 0.776 MSE on task2. For $\lambda = 0$, it achieves an average of 4.873 MSE for task1 and 4.173 for task2. The result outperforms using Emb_{bert} for the same Ridge regressor by 0.873 on task1 and 0.398 MSE on task2.

Based on the classification performance (F1 and ACC) of IPA prediction task, which is 0.509 ACC and 0.533 F1 in the task1 and 0.605 ACC and 0.617 F1 scores in the task2, it shows that both accuracy rate and F1 score are not directly correlated to the final performance of the group score prediction task. Furthermore, although the embedding learned from a simple Auto-encoder ($\lambda = 0$) did not include the information of IPA and performed worse on the IPA prediction task in stage 1, it also effectively prunes the noisy information by compressing the dimension from the original X_{dim} to 128. Finally, according to the experimental result, we know that classification loss contributes more to the final performance, and it shows that using IPA is crucial for obtaining a more discriminative group score prediction.

Regarding the other encoding methods such as LSTM, we consider it not suitable for this task for the following two reasons. First, LSTM and DNN perform essentially the same on the IPA prediction accuracy in previous work[30]. In addition, as the analysis result in our Table 6 shows that similar IPA prediction accuracy in stage 1 didn't correlate much to the group score performance in stage 2, LSTM version of IPA recognizer is likely to have very minimal effect in the overall framework.

4.4.3 Regressor in Stage 2. In our proposed two stage framework, it is important to understand the effectiveness of using different regression methods in the second stage. Therefore, we compare the prediction performance of 4 different models, including Ridge regressor, Randomforest, DNN, and Transformer. In Table 7, the results clearly show that Ridge regressor is the best performed model with the feature $Emb_{best} = Y_{predict}$. In our opinion, we think the relationship between group score and predicted communication acts ($Y_{predict}$) is linear because $Y_{predict}$ is a simple 9-dim vector and part of the dimensions are significantly correlated with group score according to our analysis in 4.6.3. In addition, similar to previous work [36], the limited number of group data often make the deep learning based SOTA method perform poorly with under-fitting issue. Although using the method like GRU and Transformer can capture more sequential dependency between IPA, these models are too heavy for our scenario. Therefore, we believe capturing overall interaction pattern by time-series model is another topic required with more sophisticated experimental design, e.g., the augmentation methods and sub-sequence segmentation method. As a result, we choose Ridge regression as a simple but effective regressor in our framework.

		NTUBA-task1		NTUBA-task2	
	Modality	mse	corr	mse	corr
Low-Level	Language	5.519±0.839	0.271±0.109	4.571±0.381	0.291±0.049
High-Level	$Count_{ipa}$	5.859±0.548	0.093±0.068	5.672±1.295	0.105±0.13
	$Ratio_{ipa}$	4.935±0.295	0.132±0.07	3.82±0.256	0.271±0.08
Proposed	Language	**4.294±0.219**	**0.325±0.062**	**3.794±0.199**	**0.273±0.074**

Table 8. Result of Experiment: Comparison between our proposed 2-staged methods and the 1-staged RandomForest model as baseline methods.

4.4.4 Effectiveness of high-level features. As our framework heavily leverages the IPA as important information to learn the embedding for predicting the group score, we further conduct the following experiment to directly demonstrate the effectiveness of the IPA label itself. In our experiment, we compare the prediction performance between low-level features, our proposed methods, and high-level features, which includes both "Count" and "Ratio" of the manually annotated IPA classes in each group. In specific, for $\boldsymbol{Count_{ipa}}$, it is a 9 dimension vector and each dimension is the number of the IPA tags in each group. $\boldsymbol{Ratio_{ipa}}$ is derived by dividing the value of $\boldsymbol{Count_{ipa}}$ by the total number of utterances in the corresponding group, so it is also a 9-dimensional vector, and the value of each dimension summed up to 1 in each group. This kind of feature is similar to unigram in language modeling and is also used in previous works for studying the relationship between dialogue act and communication skill [39]. To be consistent, we use the Ridge regressor with the same parameter (regularized parameters $alpha = 0.01$) to perform the experiment. By analyzing the differences between the 3 different methods in Table 8, we can find that $Ratio_{ipa}$ outperforms the result of using low-level features. Therefore, our results clearly indicated the fact that the lower-level verbal behavior is not sufficient to perform group score prediction tasks. However, by using our proposed framework, our method could leverage both the information in low-level features and high-level features to achieve a better prediction performance.

4.5 Robustness of the framework

Generally, the "Robust" means that the model perform stably against the noise from the real-world scenario. In this discussion, based on the result in the sec 4.3 and 4.4, we address the robustness of our framework on the following two aspects. First, according to the result in Table 8, since our proposed framework, which aggregates both the low-level feature and high-level information as the embeddings Emb_{best}, outperforms the prediction performance of only using high-level features itself, it demonstrates that our framework is robust enough against the noise in the annotation process of IPA. In other words, our framework not only predicts the label itself but also includes the information from low-level behavior features, so it won't be limited by the noise within the label. Second, based on the result in Table 3, we can find that our framework is robust across two different subsets. Based on our understanding, two different subsets, including NTUBA-task1 and NTUBA-task2, have different level of task difficulties, and the members are manifested with different interaction patterns. However, since our framework consistently outperforms the baseline methods, it validates the robustness of the proposed framework. As a result, we believe our framework is robust for applying in real-world scenarios.

4.6 Analysis of Communication Process

Based on our experimental result in Section 4.2 and Section 4.3, communication function plays an important role for automatic prediction of group performance score. In order to understand how the predicted IPA contributes to the final prediction of group score in our two-staged framework, we design the following 3 analyses.

- Analysis 1: Analysis on predicted IPA distribution

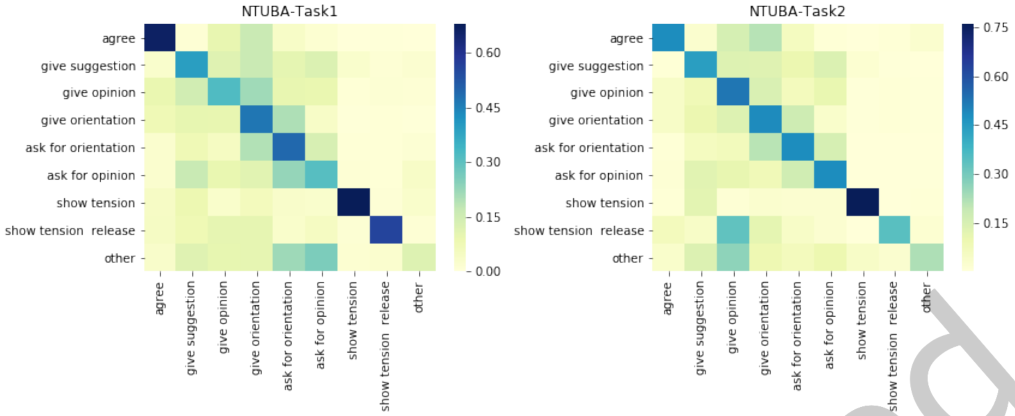

Fig. 5. Analysis between manually annotated IPA class and predicted IPA distribution: In the matrix, the y-axis represents the manually annotated IPA class (k) and the x-axis denote the averaged and predicted IPA value $Mean(Y_{predict})$ corresponds to class k.

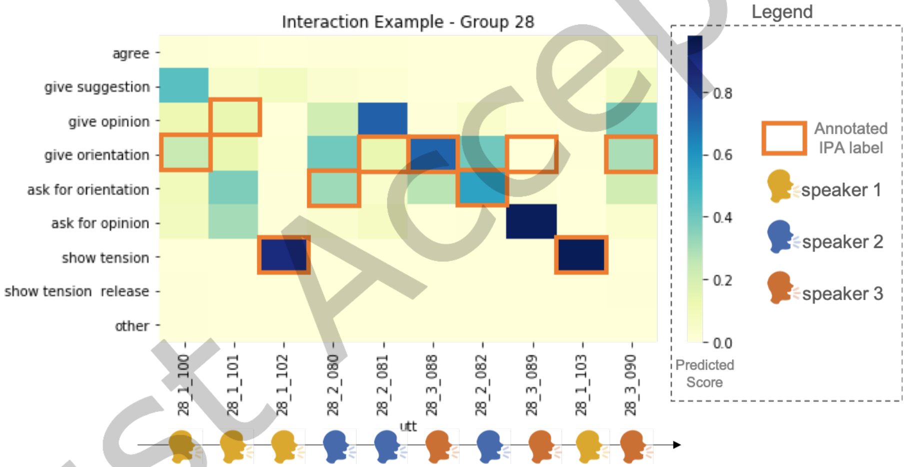

Fig. 6. Interaction Example from Group 28: We shows the predicted IPA distribution on 10 consecutive utterances from 3 speakers in group 28. In the matrix, the y-axis represents predicted IPA value ($Y_{predict}$) for the utterances and the x-axis represent the utterances name

- Analysis 2: Analysis of the examples in NTUBA
- Analysis 3: Analysis between predicted IPA tags and group score

For analysis 1, we provide a general view on the relationship between manually annotated IPA label and automatically predicted IPA label. In analysis 2, we provide an actual interaction segment from group 28 as an example to demonstrate how our proposed method works on the unique communication process. In analysis 3, we analyze the relationship between the predicted IPA and the group score.

4.6.1 Analysis on predicted IPA distribution. According to the equation (9) and the analysis result in Table 5, the predicted results of IPA ($Y_{predict}$) on every utterance is the best performed feature for the final group score

utt	start	end	IPA	Transcript
28_1_100	21:01.9	21:06.1	give orient.	那我就八點二十一去農夫市集 So I will go to the market at 8:21
28_1_101	21:08.1	21:10.5	give opin.	欸其實我可以同時去嗎 I can actually go with you at the same time?
28_1_102	21:10.9	21:11.8	show tens.	幫我看一下 Help me check for it
28_2_080	21:11.2	21:14.2	ask for orient.	我們我們有規定就是一次只能出一個人 We have rule that only one person is allow to go
28_2_081	21:14.6	21:14.8	give orient.	應該沒有 I don't think we have
28_3_088	21:14.2	21:17.5	give orient.	只有一台車 We only have one car
28_2_082	21:17.9	21:18.7	ask for orient.	喔真的喔只有一台車 Oh really… we only have one car
28_3_089	21:18.1	21:19.7	give orient.	是這樣吧 that's how it look like
28_1_103	21:19.1	21:20.7	show tens.	我看一下 Let me check
28_3_090	21:20.1	21:24.7	give orient.	對啊就是應該說明天要搭車然後只有一台車 Yes, it says we should take a ride tomorrow and there is only one car

Table 9. Transcript for Interaction Example from Group 28: this is the transcript corresponding to figure 6. This table include the information of utterance index with *group_speaker_index* as **utt**; start time and end time are presented with *minute*:*second*, transcript are presented with original Mandarin text and translated English text.

prediction in stage two. Since the $Y_{predict}$ are learned for capturing the behavior feature and communication function by using both reconstruction loss and classification loss in SIPA network, we first want to analyze the relationship between $Y_{predict}$ and manually annotated Y_{true}. In figure 5, we introduce two comparison matrices and each row in the matrices are the average predicted value of $Y_{predict}$ according to the annotated class k. In other words, we take the $Mean(Y_{predict})$ on the utterances for every class k in both database and the results can reveal the information of how SIPA network captures the communication function in two different datasets.

Overall, from the analysis result, we find that the matrices from NTUBA-task1 and NTUBA-task2 are similar but not exactly the same. Intuitively, we see that each of the nine predicted classes correspond the highest to its own intended annotated class. However, there are several interesting observations. Actions like 'agree' and 'show tension' overlap less with others. The other 7 classes often have a certain level of overlap with each other. For example, 'give/ask opinion' is similar to 'give/ask orientation'. From the semantic aspect, it implies that people often express their opinion with task information simultaneously. Similarly, people often reveal their opinions when they discuss the task. As a result, by using the $Y_{predict}$, our method leverages this additional information that is not well captured in the original annotated labels. Finally, we found that 'show tension release' and 'other' are the most different behaviors between the two tasks which indicates a situation that people in two different tasks engage in different communication processes specifically demonstrated in the use of these two communicative functions.

	NTUBA-task1				**NTUBA-task2**			
	$Mean_{g.}(Y_{predict})$		$Count_{ipa}$		$Mean_{g.}(Y_{predict})$		$Count_{ipa}$	
	corr	**p_val**	**corr**	**p_val**	**corr**	**p_val**	**corr**	**p_val**
agree	**0.312**	**0.005***	**0.285**	**0.011***	-0.091	0.444	-0.01	0.931
give suggestion	0.028	0.809	0.191	0.094	**0.257**	**0.028***	0.26	**0.026***
give opinion	0.112	0.33	0.096	0.405	-0.009	0.942	-0.085	0.477
give orientation	-0.188	0.1	0.037	0.749	-0.178	0.133	-0.027	0.819
ask for orientation	-0.21	0.064	-0.042	0.714	-0.103	0.386	0.086	0.469
ask for opinion	-0.15	0.19	0.039	0.736	0.147	0.215	-0.08	0.503
show tension	**0.223**	**0.049***	0.211	0.064	**0.26**	**0.026***	0.189	0.108
show tension release	**0.267**	**0.018***	0.106	0.355	0.072	0.545	0.099	0.403
other	0.361	0.001	0.19	0.096	0.185	0.116	0.164	0.165

Table 10. Analysis the correlation between different IPA and group performance score: We compare the learned $Mean_{group}(Y_{predict})$ and the manually annotated $Count_{ipa}$. We highlighted the correlation with significant value ($p < 0.05$) by using bold type and $*$.

4.6.2 Analysis of the examples in NTUBA. In this section, we present the result of predicted IPA tags on an actual example of an unique interaction slice in our dataset. As the ambiguity often exists between different communication functions on the real interaction data, we specifically analyze the sentences with different levels of confidence weights according to the predicted IPA score. Figure 6 shows the heat map of our $Y_{predict}$ with ten utterances from group 28. In Table 9, we present the corresponding transcript, time and IPA tags on the same ten utterances from figure 6.

From the example, we see that our SIPA network can predict well on those utterances that have little ambiguity on the semantic information. For example, for utterances 28_1_102, 28_2_082, 28_3_088 and 28_1_103, which are relatively short sentences contained with only a few words, our model can clearly identify the correct IPA categories without having other weights on other classes. However, as we mentioned in Section 4.3.1, there often exists an ambiguity between orientation and other classes like suggestion and opinion. For example, for the cases such as 28_1_100 28_2_080, and 28_3_090, although the model does not correctly predict to the originally annotated IPA class, they also have a relatively high confidential weight on it. In summary, although the ambiguity exists between different communication functions as our examples show, our framework could effectively aggregate the information from those ambiguous cases and therefore achieve better performance on the final score prediction. Furthermore, our method possesses an natural visualization approach to depict the patterns of communication function during an interaction, as shown in figure 6.

4.6.3 Analysis between predicted IPA tags and group score. For analysis 3, we compare the group score with the manually annotated IPA ($Count_{ipa}$) and mean pooling result of predicted IPA ($Mean_{group}(Y_{predict})$). This analysis can give us more insight about how to perform better during the interaction by directly providing suggestions on the group communication process. In other words, it provides us an opportunity to design proper interventions for eliciting positive communicative behavior and prevent more negative behavior.

In Table 10, we present an analysis of the relationship between the IPA-defined communicative behavior and group performance using Pearson's correlation at the group level. For $Count_{ipa}$, we present the correlation between the count of the manually annotated IPA class and group score. Besides analyzing the manually annotated IPA, we also analyze the correlation between the predicted and averaged value ($Mean_{group}(Y_{predict})$) on each IPA class and group task score, because $Mean_{group}(Y_{predict})$ is the input for our proposed framework in stage 2 and it was derived by averaging the $Y_{predict}$ from the SIPA network to get $Mean_{group}(Y_{predict})$. By comparing with the $Count_{ipa}$ and $Mean(Y_{predict})$, we can interpret how the proposed SIPA network effectively achieves a well-represented final embedding for the group performance prediction task.

Although we found that task1 and task2 do not have exactly the same trends, in general, our SIPA network increases the correlation between communicative behavior and group score. For example, in NTUBA-task1, we found that 'show tension', 'show tension release' are not significantly correlated with the group score in $Count_{ipa}$, but they became significantly correlated (p value < 0.05) after passing through the SIPA network. Similarly, we also found the same insight on "show tension" in NTUBA-task2. Not that the value of group score is lower if the group performs better (Section 2), the analysis result in Table 10 points out the interesting and reasonable fact that the group performs better when showing more action on "give orientation" and "ask for orientation" and showing less act on "show tension" during the interaction. This finding is similar to the analysis result in the previous study [13, 14] showing that more task-oriented information and less emotion-oriented information can generally achieve better group performance. By looking in details on two different scenarios, showing more "agree", "show tension release" is considered as negative-performing behavior in NTUBA-task1. In contrast, "give suggestion" is considered as negative-performing behavior in NTUBA-task2.

5 CONCLUSION AND FUTURE WORK

Recently, many computational advancements have been applied to the group performance prediction task, e.g., characterizing member's in-conversation behaviors for task performance prediction using hand-crafted features [27, 36] or graph-based conversation features [26] as well as contemporary deep network-based learning methods [31, 32, 55]. However, the use of handcrafted features often leads to underestimating the variability of the recorded behavior signals. The deep learning method also suffers from the problem of over-fitting with limited data size. In addition, most of the existing work often focuses on low-level behavior features only. Unlike previous works, the proposed methods tracks and models the "communication pattern" during the small group interaction is key to better predicting the group score computationally. Based on our experiment results, we have shown that our proposed SIPA network can better model low-level features and further summarize "the communication pattern" through the use of supervision with reconstruction (that simultaneously maintains data variability and embed meaningful communicative acts) as embedding for the group performance prediction. The mapping from low-level features to high-level information is a component of the framework, and a delicate design of such a mapping is essential to contribute to the problem of group performance using members' multimodal communicative cues. The method surpasses the previous works and achieves a promising result at 4.241 MSE and 0.341 Pearson's correlation on NTUBA-task1 and 3.794 MSE and 0.291 Pearson's correlation on NTUBA-task2. Our analysis further demonstrates the robustness and interpretability of the framework for the group communication process.

The limitation of our current work is that we only focus on the overall distribution of the interaction process label. The sequential pattern of how groups members use their behaviors resulting in a sequence of back-and-forth communicative functions should also be an important consideration. In fact, according to previous works ([48, 49]), more complete communication units, which include the orientation-planing-evaluation communication pattern, indicate a better-communicated flow and therefore have a better chance to become a better-performing group. Although direct modeling the IPA embedding with the sequential data is not effective in our preliminary result in Section 4.4.3 , we believe leveraging the communication cycle is crucial for improving our model further and the sequential modeling method is definitely important for achieving our goal. As a result, we plan to use advanced SOTA learning methods such as LSTM and Transformer for modeling the communication cycle in time series for future work.

In conclusion, our work uniquely contributes to demonstrating the idea of using higher-level features to summarize the overall conversation pattern by providing comprehensive experiments. Therefore, since the communication act fundamentally plays an important role during the multiparty conversation, our result also implicates the fact that the extracted embedding can also be beneficial in other scenarios such as empathy skill estimation ([23]), next speaker prediction ([33]). We believe using higher-level features should be an important step toward modeling human behavior, internal mental state, and complex interaction outcome.

REFERENCES

[1] Umut Avci and Oya Aran. 2016. Predicting the performance in decision-making tasks: From individual cues to group interaction. *IEEE Transactions on Multimedia* 18, 4 (2016), 643–658.

[2] Robert F Bales. 1950. Interaction process analysis; a method for the study of small groups. (1950).

[3] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. *arXiv preprint arXiv:2006.10204* (2020).

[4] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047* (2019).

[5] Indrani Bhattacharya, Michael Foley, Christine Ku, Ni Zhang, Tongtao Zhang, Cameron Mine, Manling Li, Heng Ji, Christoph Riedl, Brooke Foucault Welles, et al. 2019. The unobtrusive group interaction (UGI) corpus. In *Proceedings of the 10th ACM Multimedia Systems Conference*. ACM, 249–254.

[6] Preston C Bottger and Philip W Yetton. 1988. An integration of process and decision scheme explanations of group problem solving performance. *Organizational behavior and human decision processes* 42, 2 (1988), 234–249.

[7] McKenzie Braley and Gabriel Murray. 2018. The Group Affect and Performance (GAP) Corpus. In *Proceedings of the ICMI 2018 Workshop on Group Interaction Frontiers in Technology (GIFT)* (Boulder, CO).

[8] Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation.. In *LREC*. 430–437.

[9] Nicholas Clarke. 2010. Emotional intelligence and learning in teams. *Journal of Workplace Learning* (2010).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Joan Morris DiMicco, Katherine J Hollenbach, Anna Pandolfo, and Walter Bender. 2007. The impact of increased awareness while face-to-face. *Human–Computer Interaction* 22, 1-2 (2007), 47–96.

[12] Ute Fischer, Lori McDonnell, and Judith Orasanu. 2007. Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. *Aviation, space, and environmental medicine* 78, 5 (2007), B86–B95.

[13] H Clayton Foushee. 1984. Dyads and triads at 35,000 feet: Factors affecting group process and aircrew performance. *American Psychologist* 39, 8 (1984), 885.

[14] H Clayton Foushee and Karen L Manos. 1981. 5. INFORMATION TRANSFER WITHIN THE COCKPIT: PROBLEMS IN mTRACOCKPIT COMMUNICATIONS. *C. E. Billings: Ames Research Center. E. S. Cheaney: Battelle's Columbus Division, Mountain View, California.* (1981), 63.

[15] Marilyn E Gist, Edwin A Locke, and M Susan Taylor. 1987. Organizational behavior: Group structure, process, and effectiveness. *Journal of Management* 13, 2 (1987), 237–257.

[16] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 1 (2010), 3–19.

[17] Christopher A Gorse and Stephen Emmitt. 2007. Communication behaviour during management and design team meetings: a comparison of group interaction. *Construction management and economics* 25, 11 (2007), 1197–1213.

[18] Christopher A Gorse and Stephen Emmitt. 2009. Informal interaction in construction progress meetings. *Construction Management and Economics* 27, 10 (2009), 983–993.

[19] Christopher A Gorse, Stephen Emmitt, Mike Lowis, and Andrew Howarth. 2001. Project performance and management and design team communication. In *Proceedings of the Association of Researchers in Construction Management, 17th Annual Conference*. 5–7.

[20] J Richard Hackman and Nancy Katz. 2010. Group behavior and performance. (2010).

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

[22] Starr Roxanne Hiltz, Kenneth Johnson, and Murray Turoff. 1986. Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences. *Human communication research* 13, 2 (1986), 225–252.

[23] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2018. Analyzing gaze behavior and dialogue act during turn-taking for estimating empathy skill level. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction.*

31–39.

[24] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2019. Estimating Interpersonal Reactivity Scores Using Gaze Behavior and Dialogue Act During Turn-Changing. In *International Conference on Human-Computer Interaction*. Springer, 45–53.

[25] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. 2019. Real-time facial surface geometry from monocular video on mobile GPUs. *arXiv preprint arXiv:1907.06724* (2019).

[26] Uliyana Kubasova and Gabriel Murray. 2020. Group Performance Prediction with Limited Context. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 191–195.

[27] Uliyana Kubasova, Gabriel Murray, and McKenzie Braley. 2019. Analyzing Verbal and Nonverbal Features for Predicting Group Performance. *arXiv preprint arXiv:1907.01369* (2019).

[28] Lei Le, Andrew Patterson, and Martha White. 2018. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

[29] Gilly Leshed. 2009. *Automated language-based feedback for teamwork behaviors*. Cornell University.

[30] Sixia Li, Shogo Okada, and Jianwu Dang. 2019. Interaction Process Label Recognition in Group Discussion. In *2019 International Conference on Multimodal Interaction*. 426–434.

[31] Yun-Shao Lin and Chi-Chun Lee. 2018. Using Interlocutor-Modulated Attention BLSTM to Predict Personality Traits in Small Group Interaction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 163–169.

[32] Yun-Shao Lin and Chi-Chun Lee. 2020. Predicting Performance Outcome with a Conversational Graph Convolutional Network for Small Group Interactions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8044–8048.

[33] Usman Malik, Julien Saunier, Kotaro Funakoshi, and Alexandre Pauchet. 2020. Who Speaks Next? Turn Change and Next Speaker Prediction in Multimodal Multiparty Interaction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 349–354.

[34] Melanie J Martin and Peter W Foltz. 2004. *Automated team discourse annotation and performance prediction using LSA*. Technical Report. NEW MEXICO STATE UNIV LAS CRUCES.

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

[36] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 14–20.

[37] Michael Nowak, Juho Kim, Nam Wook Kim, and Clifford Nass. 2012. Social visualization and negotiation: effects of feedback configuration and status. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. 1081–1090.

[38] Jon Ohlsson. 2013. Team learning: Collective reflection processes in teacher teams. *Journal of Workplace Learning* (2013).

[39] Shogo Okada, Yoshihiko Ohtake, Yukiko I Nakano, Yuki Hayashi, Hung-Hsuan Huang, Yutaka Takase, and Katsumi Nitta. 2016. Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 169–176.

[40] Bolanle A Olaniran. 1994. Group performance in computer-mediated and face-to-face communication media. *Management Communication Quarterly* 7, 3 (1994), 256–281.

[41] 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.

[42] David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language* 76 (2014), 29–46.

[43] Thornton B Roby and John T Lanzetta. 1956. Work group structure, communication, and group performance. *Sociometry* 19, 2 (1956), 105–113.

[44] 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.

[45] Anit Somech. 2006. The effects of leadership style and team process on performance and innovation in functionally heterogeneous teams. *Journal of management* 32, 1 (2006), 132–157.

[46] Shree Krishna Subburaj, Angela EB Stewart, Arjun Ramesh Rao, and Sidney K D'Mello. 2020. Multimodal, multiparty modeling of collaborative problem solving performance. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 423–432.

[47] Yla R Tausczik and James W Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 459–468.

[48] Franziska Tschan. 1995. Communication enhances small group performance if it conforms to task requirements: The concept of ideal communication cycles. *Basic and Applied Social Psychology* 17, 3 (1995), 371–393.

[49] Franziska Tschan. 2002. Ideal cycles of communication (or cognitions) in triads, dyads, and individuals. *Small Group Research* 33, 6 (2002), 615–643.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[51] Michael A West. 2000. Reflexivity, revolution and innovation in work teams. In *Product development teams*. Jai Press, 1–29.

[52] Susan A Wheelan. 2005. *The handbook of group research and practice*. Sage.

[53] Annika Wiedow and Udo Konradt. 2011. Two-dimensional structure of team process improvement: Team reflection and team adaptation. *Small Group Research* 42, 1 (2011), 32–54.

[54] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

[55] Shun-Chang Zhong, Yun-Shao Lin, Chun-Min Chang, Yi-Ching Liu, and Chi-Chun Lee. 2019. Predicting Group Performances Using a Personality Composite-Network Architecture During Collaborative Task. *Proc. Interspeech 2019* (2019), 1676–1680.