Analysis of Feedback Contents and Estimation of Subjective Scores in Social Skills Training

Takeshi Saga[1] and Hiroki Tanaka[1] and Yasuhiro Matuda[2,3] and Tsubasa Morimoto[2]
and Mitsuhiro Uratani[2] and Kosuke Okazaki[2] and Yuichiro Fujimoto[1] and Satoshi Nakamura[1]

***Abstract*— This paper introduces our analysis results on the feedback contents of Social Skills Training and the consequences of automated score estimation of users' social skills with computational multimodal features. Although previous work showed the possibility of a computerized SST system as a clinical tool, its feedback strategies have not been well-investigated. We focused on the feedback content given by experienced SST trainers in human-human SST sessions to overcome this limitation. We analyzed the points mentioned by experienced SST trainers to determine where they focused during social skills evaluation. We calculated multimodal computational features from video and audio recordings inspired by the results and trained machine learning models for social skills evaluation using these features as input. We trained social skill score prediction models with the highest scores of 0.53 for correlation coefficient and 0.26 for R^2.**

***Clinical relevance*— We described our automated social skills evaluation method with machine learning models toward a computerized SST system, which can be an additional option to boost the effect of SST by experienced trainers in the future.**

I. INTRODUCTION

Social Skills Training (SST) is a rehabilitation program that has been used for more than 40 years to help people who struggle with social skills. Since social skills are essential for functioning in workplaces, schools, and even at home with a family, communication is difficult without those skills. Although we frequently use many types of social skills, Bellack et al. defined the following fundamental ones: listening, expressing positive or negative feelings, declining requests, and asking for favors [5] (We call them TELL, LISTEN, ASK, DECLINE in the rest of this paper). They proposed SST to effectively improve them. Basic training flow of SST is following. First, the trainer and the participant select a target skill from among various examples. Then, the trainer explains why that skill is essential and when it is needed. Once the participant has grasped why she must practice, the trainer shows both a good example and a bad example of that skill so that she can imagine its usage. After that, she engages in a role-play, followed by an evaluation and feedback from the trainer. This role-play and feedback process is repeated several times depending on the trainer's discretion. After the session, the trainer assigns homework to the participant to generalize the trained skill into daily communication situations: *"Thank somebody more than five separate times after someone has done something nice for you."*, for example. Although SST is a well-known rehabilitation program in clinical populations, it remains relatively inaccessible for everyone at anytime. One reason is the long training involved in becoming an SST trainer. To effectively conduct SST sessions, a professional must master how SST works and learn how to give feedback to maximize the improvement of trainees, which is time-consuming. One training program requires 8 to 12 weeks to complete, which is difficult to be enrolled in the training while working in the clinic in the daytime [6]. Therefore, researchers have studied automated SST systems for many years [1], [15].

For automated SST systems, the feedback generation process is critical. However, feedback is challenging since it is based on thorough evaluations by trainers. Moreover, the complexity of the possible combinations of social skills further complicates automatic evaluations. Various types of social skills depend on different situations. This paper focuses on social skills in interactive human-human daily communication. Social skills are vast and include gaze activities, gestures, voice intonation, speech contents, etc. [5]. The types of SST feedback are also varied. However, its broad feedback types cause data sparseness, which hampers machine learning processes. We utilized subjective scores from experienced trainers as objective values for machine learning (Section III for more detail). We chose them as metrics for the feedback generation of social skills since they are scored across different evaluation axes, each of which corresponds to an essential component of necessary social skills. We explain our analysis of the feedback contents and describe our experimental results in the rest of this paper.

II. RELATED WORK

Hoque et al. collected job interview data to develop an automated virtual agent called MACH [8], which asks questions with such interactive responses as nodding or mirroring smiles followed by feedback. Throughout their research, they showed the effectiveness of their interactive virtual agent and its feedback in job interviews. Their team

*Funding was provided by the Core Research for the Evolutional Science and Technology (Grant No. JPMJCR19A5).

*All data collection processes were approved by the ethical committees of Nara Medical University and the Nara Institute of Science and Technology. At the beginning of the recording, we explained the procedure to the participants and got informed consent.

[1]Takeshi Saga, Hiroki Tanaka, Yuichiro Fujimoto, and Satoshi Nakamura are with Nara Institute of Science and Technology, 8916-5 Takayama-Cho, Ikoma, Nara 630-0192, Japan saga.takeshi.sn0@is.naist.jp

[2]Yasuhiro Matsuda, Tsubasa Morimoto, Mitsuhiro Uratani, and Kosuke Okazaki, and are in the department of psychiatry, Nara Medical University, 840 Shijo-Cho, Kashihara, Nara 634-8521, Japan

[3]Yasuhiro Matsuda is in Osaka Psychiatric Medical Center, 3-16-21 Miyanosaka, Hirakata, Osaka 573-0022, Japan

applied their method into an SST context for older people and young people in their later works [2]. Voleti et al. collected a dataset of three role-playing scenes [16]. They collected 87 clinical subjects (44 bipolar-i disorder, 43 schizophrenia or schizoaffective disorders) and 22 healthy controls that participated in the SSPA tasks described by Patterson et al. [11]. They achieved ROC score of 0.960 just using text features. However, non-verbal skills should be addressed for better evaluations. Compared to the above research by Hoque and Voleti, we developed estimation models and multiple evaluation metrics. We employed healthy control subjects and both mentally and developmentally disordered participants, such as those suffering from schizophrenia and autism spectrum disorder.

III. DATASET

We used the human-human SST dataset collected in our previous research [13]. It has 16 autism spectrum disorder (ten males and six females, average age 26.5, SD of age 5.67), 15 schizophrenia (seven males and eight females, average age 32.07, SD of age 8.82) and 19 control (ten males and nine females, average age 28.42, SD of age 3.95) adult participants. The recruitment of symptomatic group members followed DSM-5 [3]. We targeted four basic ones: expressing positive feelings (TELL), listening (LISTEN), asking for assistance (ASK), and declining a request (DECLINE). The data modalities include facial video, audio, text transcriptions, and separate files of Kinect body points for each participant and trainer. We separated the data into each role-play and feedback phase for each SST session. We recorded for each participant's four SST sessions regarding the essential social skills defined by Belack [5].

We placed transparent partitions in the middle to reduce the risk of the COVID-19 infection. We put two cameras diagonally to avoid light reflection from the partitions to capture facial expressions. We placed another camera in the center to capture the entire scene. We used the Kinects to capture the body movements with a pretrained depth-based pose estimator. Our annotator synchronized all files by a hand-clap sound at the beginning and the end of each role-play, and transcribed all the utterances.

In addition, this dataset includes seven different types of clinical assessment scores corresponding to each symptomatic group. It also has social-skills-related subjective scores on a five-point Likert scale for each task on the following seven evaluation components: eye contact, body direction and distance, facial expression, voice variation, clarity, fluency, and social appropriateness for each task. We used averages of two experienced trainers' annotation (Cohen's kappa was 0.84).

IV. FEEDBACK CONTENT ANALYSIS

To understand the frequency distribution of the feedback content types, we analyzed what kinds of feedback were used by experienced trainers. We divided the feedback types by checking every feedback transcription for each SST session. We checked every session, and annotated feedback types for each session. Since SST feedback by the trainers took summary feedback strategy, each feedback type had never been double-counted within each SST session. Therefore, the number of annotation for each feedback type didn't directly be affected by its frequency of social skills use (e.g. smiles might be observed frequently within one session, but they got the feedback on it only one time at the end of the session). We confirmed 169 pieces of feedback, 37 different types of feedback in this dataset. Table I shows the top 10th most frequent feedback types, where *Freq.* stands for the total amount of frequency, and *Pos.* and *Neg.* denote the amount of positive and negative feedback.

TABLE I
FEEDBACK TYPES AND THEIR FREQUENCY

Feedback types	Freq.	Pos.	Neg.
Appropriate facial expression	72	49	23
Seems contrite	35	21	14
Concrete speech content	34	21	14
Eye contact	32	30	2
Backchannel	25	17	8
Clearly refused	23	21	2
Checked whether they have time to talk	19	16	3
Gesture	16	13	3
Voice amplitude	5	5	0
Express gratitude	4	4	0
Nodding	4	4	0

The distribution of the frequencies was imbalanced even when the table only included the top 10 types. The rest of the types were obviously less frequent. We identified several frequently indicated skill components in SST from this result. Based on this finding, we experimentally annotated +1 and -1 for the positive and negative feedback, respectively, and 0 if the trainer did not mention the content in the feedback. Although we preliminary attempted to train a random forest model, it resulted in poor performance. Perhaps the annotation was too sparse (most annotations were 0) since the distribution was highly imbalanced. Thus, we decided to investigate automated subjective score estimation.

V. AUTOMATED SCORE ESTIMATION WITH MULTIMODAL FEATURES

To overcome the annotation sparseness, we changed the objective values for the machine learning from discrete labels to continuous values using subjective seven-component scores by experienced trainers. Although a variety of trainer's feedback is extremely wide, we can create essential feedback based on the models' predicted subjective scores of fundamental social skills. Note that since this model doesn't generate feedback sentences directly, additional feedback selection strategy is needed, which will be done in future (Fig. 1 shows our schematic image of the model usage).

A. Method

We trained linear regression, SVR, and random forest as prediction models since these models are frequently used for small datasets. As objective values, we used the averages of subjective scores of two experienced SST trainers. Inspired

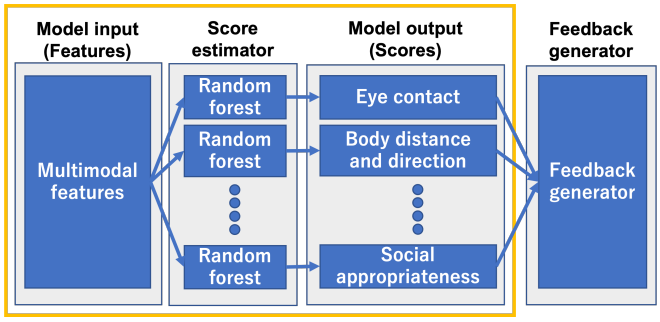

Fig. 1. **SST feedback system**. "Feedback generator" will be future work.

by observed feedback types in the previous analysis, we calculated 55 automated multimodal features to estimate those scores. Although we initially designed the feature-set to estimate the discrete feedback labels explained at the end of the previous section, we believed that the discrete feedback labels and continuous subjective scores had similar annotations for the social skills evaluation.

For the text modality, we calculated the number of content words, the words per minute, the number of backchannels, several types of sequential similarity, and binary flags for following types: appreciation words, apology words, precise refusal words, and conversational initializing phrases. We calculated the words per minute by dividing the number of words by the total duration based on the video data. We treated a spoken segment as a backchannel segment (separated with a 500-ms vocal pause) with less than five words. We measured the naturalness of consecutive words with sequential similarity [12], [14], which uses BERT, a neural network-based language model, to calculate the vectored word or sentence representations [7]. We obtained the sequential similarity by taking the average cosine similarity between consecutive words and calculated it at the sentence level and content word level for the participant utterances. We extended it to capture the correctness of the participant-trainer interactive conversations. Similar to those for participant utterances, we calculated the sentence level and content word level features for the concatenated time-aligned participant-trainer utterances.

For the audio modality, we calculated the average voice intensity and the coefficient of variation of the fundamental frequency. Although previous researchers used a wide variety of audio features [8], [10], [15], unfortunately we couldn't capture other features since the audio quality was poor for the following reasons. First, since the default microphones of video cameras recorded the audio, the distances between the microphone and the speaker's mouth were different in the recording. Second, the audio included loud environmental noises, such as sounds from air conditioners.

For the visual modality, we calculated the participant smile frequency, the mutual smile frequency, the average head angle, the coefficient of variation of the head angles, the total number of noddings, the average intensity and coefficient of variation of the facial action units (AUs) [4], and the coefficient of variation of the body points for the entire body, the upper body, and the arms. We equally treated the head angles of the yaw, roll, and pitch. Nodding was calculated based on Kawato's method [9]. We extracted the head angles and the action units with OpenFace [4].

We trained 28 models for each combination of the subjective scores and the SST tasks. We trained models independently since the evaluation must be different depending on each SST task. We used around 40 samples to train each model depending on each SST task. All input features were normalized before executing following training loop. We attempted the feature selection based on the trained random forest importance scores (n=5), which resulted in worse performance than the original one. Therefore, this paper report the results of the models without feature selection.

In the training loop, we employed a leave-one-segment-out cross-validation strategy to evaluate the effectiveness while keeping the training data as much as possible. First, we selected one segment as test data. Then, we deleted that test participant's segments from the training data to eliminate her/his personality from the training data. Next we trained the model with the training data to calculate the importance scores and selected five features based on those scores. Using the training data with selected features, we trained the model again. Finally, we predicted the subjective scores for the test segment with the trained model. We repeated this loop until we finished separately testing every segment.

B. Results and Discussion

Figure 2 shows the results of random forest models which showed the best result, where *CORREL* indicates the Pearson's correlation coefficient between the ground truth and predicted values. Bold values indicate significantly correlated predictions in the no-correlation test ($p<0.05$).

Except for the DECLINE task, at least two models showed significantly correlated predictions in three tasks. The facial expression model in the TELL task showed the highest scores of 0.53 for the correlation coefficient and 0.26 for R^2. Although we didn't observe any significant correlation in the DECLINE task, there were several weakly correlated models, such as *Eye contact* or *Social appropriateness*.

Interestingly, predictions for *Social appropriateness* showed generally higher correlations. This result indicates that although not every piece of our feature solely correlated to each evaluation axis, the combination of these features effectively estimated the user's overall social skills. Unfortunately, not all of intuitive features successfully measured each corresponding evaluation score directly, such as the smile frequency for the facial expression score. Although not all correlations and RMSEs showed synchronized results with each other, we confirmed that they showed a strong negative correlation with -0.81 by normalization with the standard deviation of ground-truth values.

VI. CONCLUSION

This paper explored the potential of our automated system to estimate social skills scores. We first analyzed the frequency of feedback by experienced trainers. After that, we

TASK	LABEL	R2	RMSE	CORREL	TASK	LABEL	R2	RMSE	CORREL
TELL	Eye contact	-0.20	1.07	0.06	ASK	Eye contact	-0.87	1.74	-0.22
	Body direct. and dist.	-0.18	1.12	0.16		Body direct. and dist.	-0.15	0.76	0.20
	Facial expression	0.26	1.17	**0.53**		Facial expression	0.09	1.41	**0.42**
	Voice variation	-0.05	1.40	0.24		Voice variation	-0.03	1.60	0.28
	Clarity	-0.41	2.25	-0.14		Clarity	0.03	1.55	**0.37**
	Fluency	-0.14	1.81	0.19		Fluency	-0.42	1.92	-0.06
	Social appropriateness	0.14	1.39	**0.40**		Social appropriateness	0.18	1.25	**0.47**
LISTEN	Eye contact	0.16	0.71	**0.46**	DECLINE	Eye contact	-0.11	1.34	0.28
	Body direct. and dist.	-0.15	0.91	0.17		Body direct. and dist.	-0.09	1.03	0.26
	Facial expression	-0.04	1.40	0.23		Facial expression	-0.19	1.76	0.14
	Voice variation	-0.18	1.70	0.14		Voice variation	-0.24	2.40	0.09
	Clarity	-0.10	1.44	0.14		Clarity	-0.20	2.26	0.08
	Fluency	-0.20	1.68	0.05		Fluency	-0.17	2.01	0.01
	Social appropriateness	0.13	1.11	**0.40**		Social appropriateness	-0.05	1.82	0.29

Fig. 2. **Prediction results**. Darker color indicates higher correlation coefficient

tried to predict subjective scores toward automated feedback generation. By using multimodal features, we exhibited the possibility of automated estimation. Yet, the room remains to further accuracy improvement of SST-feedback generation. A possible future direction is to improve features by adding more interactive ones. Analyzing the relationship between subjective scores and the actual feedback in SST might also be helpful for automated SST-feedback generation. In addition, we didn't analyze the distribution of social-skill-related subjective scores. Possibly, its detailed analysis will give us explanations for cases of score prediction failures.

ACKNOWLEDGMENT

We appreciate the generous help of an experienced clinician, Hidemi Iwasaka, including experiment designs.

REFERENCES

[1] Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K. Schubert, and Ehsan Hoque. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K. Schubert, and Mohammad Ehsan Hoque. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA, 2020. Association for Computing Machinery.

[3] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: Dsm-5*. Amer Psychiatric Pub Inc, May 2013.

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[5] Alan S. Bellack, Kim T. Mueser, Susan Gingerich, and Julie Agresta. *Social Skills Training for Schizophrenia: A Step-by-Step Guide*. Guilford Press, 2 edition, 2004.

[6] Social Skills Co. Front web page of social skills co. (Accessed September 17, 2020).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. Mach: My automated conversation coach. In *UbiComp '13: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, UbiComp '13, page 697–706, New York, NY, USA, 2013. Association for Computing Machinery.

[9] Shinjiro Kawato and Jun Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". In *Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000*, Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000, pages 40–45. IEEE Computer Society, January 2000.

[10] Iftekhar Naim, Md Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, PP, 04 2015.

[11] T. L. Patterson, S. Moscona, C. L. McKibbin, K. Davidson, and D. V. Jeste. Social skills performance assessment among older patients with schizophrenia. *Schizophrenia Research*, 48(2–3):351–360, 3 2001.

[12] Takeshi Saga, Hiroki Tanaka, Hidemi Iwasaka, and Satoshi Nakamura. Objective prediction of social skills level for automated social skills training using audio and text information. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 467–471, New York, NY, USA, 2020. Association for Computing Machinery.

[13] Takeshi Saga, Hiroki Tanaka, Hidemi Iwasaka, and Satoshi Nakamura. Multimodal dataset of social skills training in natural conversational setting. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, ICMI '21 Companion, New York, NY, USA, 2021. Association for Computing Machinery.

[14] Takeshi SAGA, Hiroki TANAKA, Hidemi IWASAKA, and Satoshi NAKAMURA. Multimodal prediction of social responsiveness score with bert-based text features. *IEICE TRANSACTIONS on Information and Systems*, E105-D(3), 03 2022.

[15] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE*, 12(8):1–15, 08 2017.

[16] Rohit Voleti, Stephanie Woolridge, Julie Liss, Melissa Milanovic, Christopher Bowie, and Visar Berisha. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. In *Proceedings of Interspeech 2019*, pages 1433–1437. International Speech Communication Association, 09 2019.