IEICE TRANS. INF. & SYST., VOL.E105–D, NO.3 MARCH 2022

PAPER Special Section on Human Communication IV

Multimodal Prediction of Social Responsiveness Score with BERT-Based Text Features

Takeshi SAGA[†a)], **Hiroki TANAKA**[†b)], ***Members***, **Hidemi IWASAKA**[††c)], ***Nonmember***, ***and*** **Satoshi NAKAMURA**[†d)], ***Member***

SUMMARY Social Skills Training (SST) has been used for years to improve individuals' social skills toward building a better daily life. In SST carried out by humans, the social skills level is usually evaluated through a verbal interview conducted by the trainer. Although this evaluation is based on psychiatric knowledge and professional experience, its quality depends on the trainer's capabilities. Therefore, to standardize such evaluations, quantifiable metrics are required. To meet this need, the second edition of the Social Responsiveness Scale (SRS-2) offers a viable solution because it has been extensively tested and standardized by empirical research works. This paper describes the development of an automated method to evaluate a person's social skills level based on SRS-2. We use multimodal features, including BERT-based features, and perform score estimation with a 0.76 Pearson correlation coefficient while using feature selection. In addition, we examine the linguistic aspects of BERT-based features through subjective evaluations. Consequently, the BERT-based features show a strong negative correlation with human subjective scores of fluency, appropriate word choice, and understandable speech structure.

key words: *social skills training, social communication, social responsiveness scale, linear regression, feature selection*

1. Introduction

Social Skills are essential for communicating with others. According to Bellack et al., the components of social skills can mainly be divided into three groups: expressive behavior, such as eye contact and posture; receptive behavior, which includes attention to and interpretation of the relevant cues for emotion recognition; and interactive behavior, such as response timing and turn-taking [1]. Although most people possess these abilities, some lack them due to mental disabilities such as autism spectrum disorder (ASD) or schizophrenia [2]. Daily life is quite tricky without social skills since most normal situations require us to integrate and apply several different social skills.

One solution to this difficulty is a process called Social Skills Training (SST). According to Bellack et al., it is based on conditioned reflex therapy and psychotherapy by reciprocal inhibition and social learning theory [3]–[6]. Since the introduction of SST, it has been extensively used in a wide range of areas [1]. The minimum setup for SST is one trainer and one trainee [1]. It can also be done in a group setting with several trainees. In that setting, an assistant can be designated from among the trainees to help the trainer. Figure 1 shows the basic SST procedure. First, the trainer and the trainees determine the required objective skill with a specific situation and goal of the SST. Second, the trainer demonstrates an exemplary model of the skill/goal by acting in the situation himself/herself. Third, an individual trainee imitates the trainer's example. Fourth, positive and negative feedback is given to the trainee by the trainer. Fifth, based on the feedback, the trainee repeats the performance while trying to improve it. Sometimes, homework is assigned to trainees for further improvement.

Although SST is a well-known method, in practice, access to it is difficult for several reasons. First, there is a social stigma attached to those with mental illnesses [7], [8]. Many such people are not entirely accepted in certain situations. Patients sometimes fear that their family, including themselves, might be abused or shunned by neighbors who discover their illness. Another reason is the difficult training involved in becoming an SST trainer. To conduct SST effectively, a professional must master how it works and learn how to give feedback to maximize the trainees' improvement, which is time-consuming. One training program requires 8 to 12 weeks to complete [9]. Therefore, automated SST systems have been studied for many years [10]–[13].

Our series of research works focused on the first two steps of automated SST (steps within the red box in Fig. 1): measuring the user's skills level and identifying those skills required in the SST training. In previous research, we tried

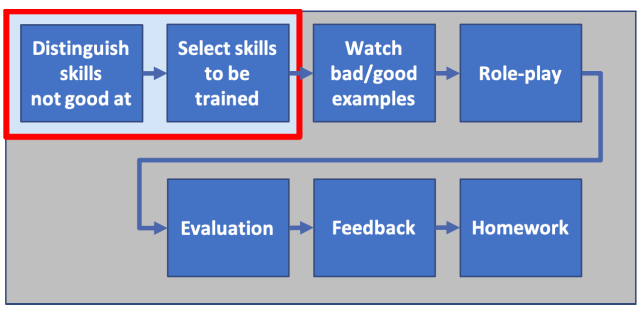

Fig. 1 **Flowchart of SST**. Our research target includes steps in the red box.

Manuscript received April 16, 2021.
Manuscript revised September 11, 2021.
Manuscript publicized November 2, 2021.

[†]The authors are with Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[†]The author is with Department of Psychiatry, Nara Medical University, Kashihara-shi, 634–8521 Japan.
a) E-mail: saga.takeshi.sn0@is.naist.jp
b) E-mail: hiroki-tan@is.naist.jp
c) E-mail: iwasaka@heartland.or.jp
d) E-mail: s-nakamura@is.naist.jp
DOI: 10.1587/transinf.2021HCP0009

to estimate the user's severity of social impairment by using linear regression with text and audio features [14]. To minimize the subjectivity of ground-truth values, we applied the Social Responsiveness Scale Second Edition (SRS-2) to obtain standardized ground-truth values for the regression model. Although we demonstrated the potential of our approach, the prediction accuracy was not high enough because the number of features was limited. Moreover, there was need for further investigation into the kinds of linguistic information that could be captured by seq-similarity (called BERT_word in this paper).

In this paper, to extend the previous research, we used a feature-selection method based on Pearson's correlation coefficient with a regression model. We also revised multimodal feature sets to include visual features and other sophisticated features. In particular, we added some variations of *embedding-based* textual features to investigate the effects of BERT-based features. *Embedding* was initially proposed in the natural language processing field. It is a type of vector used to express a word's meaning in a distributional space. Furthermore, we examined subjective evaluations to find the relationships between BERT-based features and human cognition. It should be noted that, since SRS-2 has been proven effective not only for socially disabled people but also for non-disabled subjects, this research evaluated its effectiveness only for non-disabled subjects in this preliminary feasibility study. The novelty of this research can be summarized as follows:

- SRS-2 score estimation model was developed with a Pearson's correlation coefficient of 0.76
- BERT-based features were applied to evaluate the quality of a one-minute talk
- Effectiveness of the BERT-based features was confirmed by subjective evaluation

Section 2 introduces several related works in this area, including therapy systems based on an automated agent. Next, in Sect. 3, we describe the *2th edition of the Social Responsiveness Scale* as a source of ground-truth values in this paper. In Sect. 4, we show details of the dataset used in this paper. Then, in Sect. 5, we show our method of estimating social skills level and its capabilities. In Sect. 6, we present experiments using subjective evaluation to investigate the details of linguistic features. Finally, in Sect. 7, we conclude the paper with a discussion on several research limitations.

2. Related Work

This section introduces several previous research efforts related to automated behavior scoring or training systems to help the reader better understand the rest of this paper.

SimSensei has been acknowledged as one of the epic projects in this field. It was a virtual trainer that conducted therapy sessions for PTSD patients [15]. The authors integrated many different libraries to capture multimodal features of the user. They reported that its quality approached that of face-to-face sessions.

Naim et al. attempted to estimate job interview performance by using multimodal features [16]. They set average ratings of 16 interview traits as ground-truth labels. The several regression models they used achieved the best prediction performance by SVR with a correlation coefficient of 0.70 for overall performance and hiring recommendation traits. They also examined the relative weights of individual features in their regression models. Their results show that prosodic features have relatively larger weights for predicting engagement and excitement, which is intuitively reasonable.

In the SST context specifically, Tanaka et al. automated the SST process with a computer system using a multimodal dialogue system having an embodied conversational agent [12]. Their system goal was to improve the user's speaking skills. They used a subjective score, the speaking skill score, as an evaluation metric to measure a user's speaking level. The scores were annotated by several experienced trainers and averaged as ground-truth labels for each bit of data. Although the scores were partially based on professional knowledge, they still lacked objectivity because they depended on the trainer's skill level. More critically, since most psychiatric symptoms are not physical phenomena like blood pressure level, in principle, psychiatric physicians cannot rely on objective measures as other physicians routinely do [17], [18]. Therefore, there is need for an objective evaluation method based on a quality-guaranteed metric.

Although psychiatrists commonly accept the importance of speech contents, few works using them have been conducted due to the difficulty of capturing speech contents computationally. Aramaki et al. investigated the relationships between the SRS-2 score (details in Sect. 4) and the usage of difficult words. Their results indicate that people with a high SRS-2 score tend to use less complicated words in conversation.

3. Dataset

We used the monologue video dataset gathered by Tanaka et al. [13]. This dataset includes 27 participants who were non-disabled adults with a mean age of 25.1. Participants were asked to talk for one minute to a virtual agent on a screen about positive events they recently experienced. Four fundamental social skills were defined by Bellack: expressing positive behavior, listening to others, asking favors, and refusing requests. The task "talking about a recent positive topic" is related to one of the four skills, expressing positive behavior [1].

Figure 2 shows an image of the adopted recording system in use. The agent blinks its eyes once every three seconds and nods a few seconds after recognizing an utterance. The authors reported that the agent's effectiveness was comparable to or even better than an unfamiliar partner [11]. Furthermore, in subjective self-evaluation of the improvement in the trainee's talking ability after the session, SST

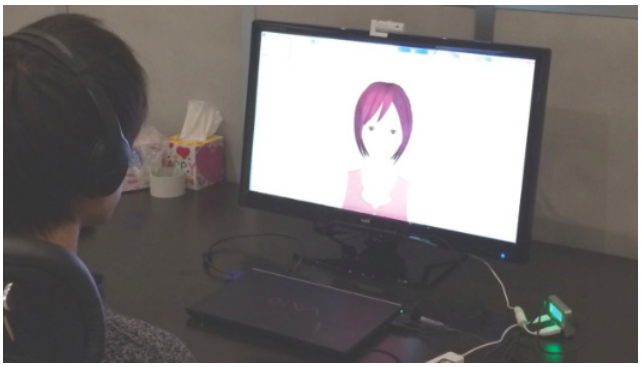

Fig. 2 Experiment's recording system in use [13]

by the agent showed almost the same effectiveness as SST by a familiar partner. Therefore, we believe the interaction with this agent is applicable to our research purpose. We recorded the user's facial images, voice, and eye movements synchronously. After the video recording, participants were asked to answer two questionnaires: the Big Five Personality Test and SRS-2. A human annotator transcribed every user's utterances.

The recording duration was one minute. Although a longer talk session might have been better to capture the characteristic behaviors of users, it would also have been a more challenging task to complete for participants. Some participants reported it was difficult to talk continuously even for one minute. Therefore, we balanced these factors and set the recording time to a one-minute duration.

4. Social Responsiveness Scale, 2th Edition (SRS-2)

SRS-2 is an evaluation metric of the severity of social impairment, and it is composed of 65 questions. Although SRS-2 was initially designed to assess potential autism spectrum disorder (ASD) sufferers, it can also differentiate various mental diseases. Furthermore, its effectiveness has been investigated not only with disabled people but also with non-disabled people [19]. Consequently, it was found suitable for evaluating non-disabled people as well.

This paper used the SRS-2 overall score and one of its treatment sub-scale scores called Social Communication (22 items out of 65) as ground-truth labels for training. Since the overall score of SRS-2 includes and can be affected by social communication skills and lifestyle factors, we decided to also evaluate the Social communication score of SRS-2, which provides more explicit information on the user's social communication skills than the overall score. In particular, it indicates the physical aspect of social interaction, which is more feasible for future SST automation because it is intuitively understandable and objectively observable [19].

Social communication score and SRS-2 overall score were highly correlated with a coefficient of 0.92. On the other hand, correlation coefficients between the scores and the subjective speaking score, rated by experts from previous research, were −0.39 and −0.33, respectively [13].

For the overall score distribution, the average and the standard deviation were 65.90 and 19.77, respectively, for males and 61.50 and 21.00 for females. Since the symptomatic population of ASD is gender imbalanced, its cutoff values are different at 65 and 52 for males and females, respectively [20]. From this distribution, we observed that this dataset includes scores in the general populations' range and those in the ASD range, although participants had never received any clinical diagnosis. Therefore, we think these subjects were suitable for this experimental research.

5. Experiment 1: SRS-2 Score Prediction Model

In our first experiment, we attempted to predict SRS-2 scores using machine learning with multimodal features.

5.1 Multimodal Features

Since multimodal information is needed to acquire better socials skills, we considered using multimodal features as inputs to the model [1].

Tanaka et al. used linear regression with the following features to estimate subjective speaking score: words per minute, words over six letters, number of fillers, vocal amplitude mean, coefficient of F0 variation, pause percentage, spectral tilt between F1 and F3 (H1A3), smile ratio, and head poses (pitch, yaw, roll) [12]. They selected these features based on their previous work on differentiating Japanese people with and without social difficulties [21]. In this experiment, we set these features as feature-set1 and evaluated whether they effectively estimated the SRS-2 score, which includes many other aspects of communication than just the speaking score. Although feature-set1 includes multimodal information, the variety of features is limited. Moreover, none of these features involve giving consideration to speech contents. However, the irregularity of speech contents is one of the critical symptoms in many psychiatric disorders such as schizophrenia and autism spectrum disorder (ASD) [2]. Therefore, in feature-set2, we decided to add text features that involve the speech contents of the user. A list of the features used in this research, with brief descriptions, is given in Table 1.

5.1.1 Text Features

Disorganized speech is one of the significant symptoms of Schizophrenia. We can find an extreme example of this in Bleuler's book: "I always liked geography. My last teacher in that subject was Professor August A. He was a man with black eyes. I also like black eyes. There are also blue and grey eyes and other sorts, too..." [22]. Although this symptom has been recognized for years, capturing it by automatic calculation has not been easy because we need to consider the context of the speech.

To capture this symptom, we propose applying BERT-based similarity features, named BERT-based similarity

Table 1 Multimodal features

Feature name	Description
Energy	Mean spectral energy
F0, F1, F2, F3 Mean	Mean frequency of F0, F1, F2, F3
F0, F1, F2, F3 SD	Standard deviation of F0, F1, F2, F3
F0 Min	Minimum F0 frequency
F0 Max	Maximum F0 frequency
F0 range	Difference between F0 MAX and F0 MIN
F1, F2, F3 BW	Average bandwidth of F1, F2, F3
F2/F1, F3/F1 Mean	Mean ratio of F2-F1 and F3-F1
F2/F1, F3/F1 SD	Standard deviation of F2/F1 and F3/F1
Int mean	Mean vocal intensity
Int Min	Minimum vocal intensity
Int Max	Maximum vocal intensity
Int range	Differences between max and min intensities
Int SD	Standard deviation of vocal intensity
Jitter	Irregularities in F0 frequency
Shimmer	Irregularities in intensity
Unvoiced %	Percentage of unvoiced regions
Breaks %	Average percentage of breaks
WPM	# of Words in one minute
Vocabulary_size	Total size of vocaburary
Six_plus	# of Words more than six letters
BERT_word	Word level feature with BERT
W2V_word	Word-level feature with Word2Vec
BERT_cont_word	Content-word-level feature with BERT
W2V_cont_word	Content-word-level feature with Word2Vec
BERT_sent	Sentence-level feature with BERT
W2V_sent	Sentence-level feature with Word2Vec
Conj%	Percentage of conjunctions
Filler%	Percentage of fillers
Pronoun%	Percentage of fillers
Nei_cont_identical	Total number of identical content words between adjacent sentences
Pose_Rx	Head angle in radian around X axis
Pose_Rx_CV	Coefficient of variation for Pose_Rx
Pose_Ry	Head angle in radian around Y-axis
Pose_Ry_CV	Coefficient of variation for Pose_Ry
Pose_Rz	Head angle in radian around Z-axis
Pose_Rz_CV	Coefficient of variation for Pose_Rz
Smile_ratio	Ratio of Smiling frames to # of total frames
AU01	Inner-brow raiser
AU02	Outer-brow raiser
AU04	Brow lowerer
AU05	Upper-lid raiser
AU06	Cheek raiser
AU07	Lid tightener
AU09	Nose wrinkler
AU10	Upper-lip raiser
AU12	Lip-corner puller
AU14	Dimpler
AU15	Lip-corner depressor
AU17	Chin raiser
AU20	Lip stretcher
AU23	Lip tightener
AU25	Lips parter
AU26	Jaw dropper
AU28	Lip sucker
AU45	Blinker

(BERT_sim). Equation (1) shows BERT_sim's calculation, where N is the total number of words, $Cos(A, B)$ is the cosine similarity between A and B, and $embed_i$ is the ith embedding.

$$BERT\_sim = \frac{1}{N}\sum_{i=1}^{N} Cos(embed_{i-1}, embed_i). \quad (1)$$

BERT is a type of neural network initially proposed in the natural language processing field [23]. Compared to previous neural network approaches like Word2Vec, BERT can capture context across an entire input text regardless of its length [24]. Therefore, BERT provides better word-level representation than previous methods like word2Vec. We used the BERT pretrained model from the *Transformers* library (bert-base-japanese) [25].

This paper uses several setups for the embedding calculations (Table 1). To show the effectiveness of BERT-based features, we calculated Word2Vec-based features as well. If the user produces only similar words sequentially, the score should be high. We implemented this text feature because physicians have reported that people with ASD tend to use only a limited variety of words compared to non-disabled people [26]. Moreover, we also aim to capture *disorganized speech* by using these features. Although we did not include data for actual patients with ASD or schizophrenia in the dataset, we can test whether our method effectively captures these symptoms because the participants used in the dataset were labeled with corresponding SRS-2 scores.

Unlike conventional topic models (e.g., Latent Dirichlet Allocation), this method does not require hyperparameters to define several topic classes. Therefore, it is more generic than other methods based on topic modeling.

Additionally, we implemented the following statistical text features based on psychiatric knowledge [2]: conjunction percentage (Conj%), filler percentage (Filler%), and pronoun percentage (Pronoun%). We calculated the frequency of identical neighboring content words (Neighbor_content_identical) to capture the strength of the semantical connection between adjacent sentences using the following procedure. First, we changed every word to its dictionary form. Second, We counted the number of identical content words between adjacent sentences. Finally, we used the sum of those numbers as the feature value.

5.1.2 Audio Features

To increase the variety of audio features, we adopted the features used in a job interview scoring system [16]. We decided to apply these features to this research because job interviews also use multimodal information in a manner similar to SST. When we communicate with others, linguistic speech content and other non-verbal information deliver the speaker's true intentions to listeners. Furthermore, in previous research on ASD and schizophrenia patients, significant differences were reported for audio features such as pause duration and pitch (F0) variability [27], [28]. These features

were calculated with the audio analysis toolkit Praat [29].

5.1.3 Visual Features

Similar to audio features, we needed to add extra visual features for better system performance. In this research, we added features related to facial expressions. It is said that facial expressions are essential for communication, and their effectiveness has been reported in previous works [10], [15], [30], [31]. To calculate facial-related features, we mainly used the facial expression analysis tool *OpenFace* [32]. It also can handle head pose and action units (AUs), which are tiny components of facial expressions [33].

OpenFace's AU recognition evaluates two different attributes: presence and intensity. Presence is a binary attribute indicating whether the AU is detected. In contrast, intensity is a continuous value indicating the strength of the AU's presence, ranging from 0 to 5. Since presence was more stable than intensity in our preliminary experiments, we chose presence as the AU feature for this study. In addition, since the number of video frames differed depending on the particular video, we used the average of these presence values for each AU as the input feature to the machine learning model. In terms of head pose, we calculated the average pose in absolute value and coefficient of variation for pitch, yaw, and roll.

In previous research, researchers reported that smile frequency is one of the important features related to social skills [21]. Therefore, we took this feature into account. We used a smile detection model based on cascade filters with the OpenCV computer vision library [34]. We averaged its presence, similarly to AU features, by dividing the data's value by the number of video frames.

5.2 Experimental Conditions

Following Tanaka's research, we used linear regression as a machine learning model [12]. Since linear regression is an easily explainable and straightforward method, it is suitable for this research, especially when we try to identify skills that should be further trained. Due to the simplicity of linear regression, it can be easily degraded by multicollinearity across different features. To avoid this, we applied a feature selection method based on Pearson's correlation scores if the total number of features exceeded 10. Pearson's correlation score was calculated with the following equation:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}, \tag{2}$$

where x_i and y_i indicate a feature value and an objective score, respectively, for a participant's data with an index i, $\overline{x}$ and $\overline{y}$ indicate average scores of feature values and objective scores, and n indicates the total number of participants. By using this correlation score, we calculated the F-statistic with the following equation:

$$F = \frac{{r_{xy}}^2}{1 - {r_{xy}}^2}(n-2), \tag{3}$$

where r_{xy} indicates the correlation score with a feature value x and an objective value y, and $(n-2)$ indicates the degree of freedom for this setting with participant data size n. We used this F-statistic to sort features in descending order. Next, we filtered out all data except for the top-10 relevant features.

As mentioned in Sect. 4, we set overall score (SRS) and social communication score (Com) as ground-truth scores for the estimation from the *Social Responsiveness Scale, 2th edition*. All input features were standardized to have an average value of zero and a standard deviation of one. Sometimes, linear regression is used with regularization such as L1, L2, and a combination of L1 and L2 called elastic net. However, such regularizations did not improve anything in our preliminary experiments. Therefore, we did not apply any regularization in this paper.

In preparing the features, we found an outlier caused by a failure of the head pose estimation. Consequently, we eliminated this from the dataset and trained the models using the remaining data of the 26 participants with 67 features each.

To deal with the limited size of the dataset, we utilized the leave-one-out cross-validation technique to minimize individuality. Therefore, we used 25 participants' data out of 26 to train the models and used 1 participant's data as test data for each model. For visual feature extraction, we used OpenFace with a default setting of 15 fps for video data. OpenFace can calculate the AU binary presence score and AU linear intensity score separately; however, these two scores might not be consistent because their predictors were trained independently [33]. Consequently, we used only the presence score since it was found to be more stable than intensity in our preliminary experiment. At the same time, eliminating one of these scores can avoid collinearity between them. For audio feature extraction, we used data with a sampling rate of 22 kHz. We set the target pitch frequency range from 100 to 600 Hz for pitch extraction. The Berg method was used with a window length of 0.025 seconds for the formant estimation. For the remaining conditions, we used the default settings of Praat [4].

In addition to the linear regression model, we experimentally trained the partial least square (PLS) regression model and XGBoost regression model, anticipating better prediction performance.

5.3 Results

Table 2 shows the results for social skills level estimation, where *SetX* indicates feature-setX, *select.* indicates feature selection, *Correl* indicates Pearson's correlation score between ground-truth values and estimated values, and *RMSE* indicates root mean squared error.

For the prediction using feature-set1, we confirmed that feature-set1 was influential not only for the subjective speaking score used in the previous research but also for the social communication score of SRS-2. By comparing the results of feature-set1 and feature-set2, there was no significant difference between estimation results for either over-

Table 2 Results of social skills level estimation [a]

Name	SRS		Com	
	Correl	RMSE	Correl	RMSE
Set1 [12]	0.28	32.13	0.45*	11.33
Set2	0.42*	18.80	0.30	10.54
Set2 with select.	0.76*	13.49	0.63*	8.42

[a] *Correl* with * indicates a significant difference in the test of no correlation

Table 3 Selected features in descending order of feature coefficient

SRS	Com
F0_SD	F0_SD
AU17	AU45
AU45	F3_SD
F3_SD	AU17
Pose_Rx_CV	AU28
Pose_Rx	AU23
AU28	Vocabulary_size
AU15	Pose_Rx_CV
AU01	F3_mean
Pose_Rz_CV	BERT_cont_word

all score or social communication score (Wilcoxon signed-rank test, p-value threshold at 0.05). In contrast, feature-set2 with feature selection showed the best correlation score for both overall score prediction and social communication score prediction. In terms of RMSE, we confirmed that the model trained on feature-set2 with feature selection produced the lowest score.

Table 3 shows the selected features arranged in descending order of feature coefficient magnitude. In both overall score prediction and social communication score prediction, the standard deviation of F0 had the highest coefficient. Moreover, the effectiveness of text features such as *vocabulary_size* and *BERT_content_word* was seen only in social communication prediction.

Although we trained PLS regression and XGBoost regression, those models could not surpass the linear regression model in RMSE or the correlation coefficient between ground-truth and predicted values.

5.4 Discussion

From the prediction results shown in Table 2, we found feature-set1 was adequate for predicting the social communication score since its correlation coefficient was significantly correlated to the ground-truth scores ($p<0.05$). Despite its limited number of multimodal features, surprisingly, it demonstrated its predictive ability. We believe this was because the subjective speaking and social communication scores were highly correlated, with a coefficient of -0.39. Again, although the subjective speaking score was rated by experienced SST trainers, SRS-2 is more reliable because its effectiveness has been proven by previous research [19], [35]–[37]. However, it was not significantly correlated with overall score prediction using feature-set1. We think this is because the overall score includes communication aspects of social skills and daily habits as well as other non-communicational information that was not considered in the subjective speaking score [13].

Compared to feature-set1, feature-set2 with feature selection was more effective in predicting scores with a correlation coefficient of 0.76 for overall score and 0.63 for social communication score. In Table 3, *F0_SD* had the highest feature coefficient of linear regression. In previous research, the authors reported that an irregularity of F0 movement was found in ASD patients and schizophrenia patients [27], [28]. The results from this experiment support those earlier findings. Interestingly, the effectiveness of text features was only seen in the results of social communication score prediction, not in overall score prediction. These results indicate that text content (i.e., what was said) was unimportant for overall score prediction.

As a result, XGBoost regression and PLS regression could not surpass linear regression. In addition to this performance difference, we believe its interpretability is an additional reason to choose the linear regression model for the final goal of our research. Although we can use decision tree-based algorithms such as XGBoost for the feature selection with importance scores, it is challenging to interpret and quantize the results instantly due to their complex decision paths. Such instant interpretation and quantization for each skill are critical for our future usage of this model in giving users feedback right after the role-play (called summary feedback in SST) or within the role play (called immediate feedback in SST). Similarly, the PLS regression model is also difficult to interpret since its principal component axis is ambiguous, and researchers' subjectivity is needed to determine the meaning of the axes. Therefore, we believe the most straightforward linear algorithm, based on correlation coefficients, is the best choice for our research purpose. In addition, a recent paper achieved high prediction performance on debate skill score prediction using linear regression [38]. Since its target is similar to social skills in terms of the necessity of multi-modality communication, we believe this result supports our usage of linear regression for the social skills score prediction.

6. Experiment 2: Subjective Evaluation

In Sect. 4, we confirmed that SRS-2 scores could be predicted using linear regression with multimodal features and the method of feature selection. Although its score prediction was accurate, it remained unclear what kind of linguistic information that *BERT_cont_word* captured. To investigate this, we conducted a second experiment that focused on subjective evaluation.

6.1 Method

We collected 15 participants using a web-based crowd worker service. Participants were asked to rate text examples from 1 to 7 subjectively for three different tasks. They

Table 4 Items of subjective rating

Name	Description
Content_word	Similarity between adjacent content words
Sentence	Similarity between adjacent sentences
Word_choice	Proper word choice for talking with psychiatrists
Fluency	Fluency as a native speaker
Structure	Story structure

	content_word	sentence	word_choice	fluency	structure
BERT_word	0.26	0.08	-0.37	**-0.84**	**-0.63**
BERT_content_word	0.26	0.10	**-0.42**	**-0.70**	**-0.65**
Pronoun%	-0.07	-0.03	**-0.40**	**-0.41**	**-0.51**

Fig. 3 **Correlation between features and subjective scores**. Darker red indicates a stronger correlation. Values with underlines were significantly correlated.

were asked to rate the scores absolutely, not relatively. All rating items are shown in Table 4. We selected these items based on the measured symptoms of schizophrenia [2]. *Content_word* was used for task1. *Sentence* was used for task2. *Word_choice*, *Fluency*, and *Structure* were used for task3. In task1, participants were shown a series of content words used in the one-minute talk. Then, they were asked to rate the overall subjective score for correlations between adjacent content words one by one for each subject in the dataset. This score is named *Content_word* in Table 4. In task2, participants were asked to perform similar activities to those in task1 except for using a series of sentences instead of content words. This score is named *Sentence* in Table 4. In task3, participants were shown a raw text transcription for the one-minute talk. Then, they were asked to rate the overall subjective score for three different aspects: appropriateness of word choice if the participants talked to a trainer (*Word_choice*), fluency as a native speaker (*Fluency*), and structure level of the talk (*Structure*).

Using the collected scores, we calculated Pearson's correlation between the features in feature-set2 and the scores. The original target of this experiment was only *BERT_cont_word*. However, it would be better for future research to find text features that strongly reflect the user's subjective evaluation metrics. Therefore, we decided to calculate the correlation for every feature in feature-set2. Furthermore, the intra-score correlation was calculated to investigate the correlations among the subjective scores by the crowd workers.

6.2 Results

Figure 3 shows Pearson's correlation coefficients between scores and features, where scores with underlines indicate significant correlation ($p<0.05$). Although we calculated the correlation coefficients for all features in feature-set2, we confirmed that only *BERT_word*, *BERT_cont_word*, and *Pronoun%* were significantly correlated out of the 13 text features. Additionally, we confirmed positive correlations between *BERT_word* and *content_word* and between *BERT_cont_word* and *content_word*.

	content_word	sentence	word_choice	fluency	structure
content_word	1.00	0.32	-0.04	-0.08	-0.11
sentence	0.32	1.00	-0.15	0.17	-0.06
word_choice	-0.04	-0.15	**1.00**	**0.49**	**0.72**
fluency	-0.08	0.17	**0.49**	**1.00**	**0.73**
structure	-0.11	-0.06	**0.72**	**0.73**	**1.00**

Fig. 4 **Intra-score correlation of subjective scoring**. Darker red indicates a stronger correlation. Values with underlines were significantly correlated.

Figure 4 shows intra-score correlation of subjective scores, where scores with underlines indicate significant correlation ($p<0.05$). From the results, we found that *word_choice*, *fluency*, and *structure* were significantly correlated.

6.3 Discussion

Originally, we designed *BERT_cont_word* to capture similar information to the subjective score of *content_word*. Although there was no significance, we confirmed the correlation between them with a coefficient of 0.26. Therefore, it seems our method could capture similar characteristics to those that human raters did. Moreover, unexpectedly, the text features shown in Fig. 3 were strongly and negatively correlated with *word_choice*, *fluency*, and *structure*. The results indicate that these features tend to be low if the subjective scores are high.

In Fig. 4, *word_choice*, *fluency*, and *structure* showed similar trends. According to the figure, their subjective scores were strongly correlated to each other. These scores were rated in the same task, with ratings executed after reading the entire text of the talk. Therefore, unfortunately, these subjective scores might have interfered with each other. However, this may also indicate that the trends of these scores are related to the *overall quality* of the speech.

Therefore, these solid negative correlations between BERT-based features and subjective scores suggest that people who speak well tend to talk about a wide variety of topics and contents. Conceptually, BERT-based features can be high if the adjacent embeddings are similar. Hence, BERT-based features should be low if the words in the talk are semantically diverse. This assumption should be evaluated carefully from several different perspectives in future work.

7. Conclusion

Using a feature-selection method, we showed that SRS-2 scores could be estimated accurately using linear regression from multimodal features. In addition, we proposed BERT-based features to handle the context of speech. The results of estimating the social skills level suggest that *BERT_cont_word* and *Vocabulary_size* are important for estimating the social communication score. In contrast, the results also indicate that text features are not crucial for overall score prediction. In addition, we attempted to clarify the kinds of aspects that BERT-based features captured from the text. The results suggest that people who speak well tend to

talk about a wide variety of topics and contents.

Although we mainly discussed the effects of text features, we could not investigate the effects of non-verbal features. This non-verbal effect should be researched in the future. In addition, the extension from one-to-one communication to social interaction with three or more people should also be investigated since such group interaction is far more complex than the setting in this paper.

Acknowledgements

Funding was provided by the Core Research for Evolutional Science and Technology (Grant No. JPMJCR19A5) and the Japan Society for the Promotion of Science (Grant No. JP18K11437).

References

[1] A.S. Bellack, K.T. Mueser, S. Gingerich, and J. Agresta, Social Skills Training for Schizophrenia: A Step-by-Step Guide, 2 ed., Guilford Press, 2004.
[2] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders: Dsm-5, Amer Psychiatric Pub Inc, May 2013.
[3] A. Salter, Conditioned reflex therapy, Creative Age Press, 1949.
[4] J. Wolpe, Psychotherapy by reciprocal inhibition, Stanford University Press, 1958.
[5] A. Bandura, Principles of behavior modification, Holt, Rinehart and Winston, 1969.
[6] K.T. Mueser and A.S. Bellack, "Social skills training: Alive and well?," Journal of Mental Health, vol.16, no.5, pp.549–552, 2007.
[7] J. Hunt and D. Eisenberg, "Mental health problems and help-seeking behavior among college students," Journal of Adolescent Health, vol.46, no.1, pp.3–10, 2010.
[8] B. et al., "The mental health and well-being of ontario students, 1991-2015: Detailed osduhs findings," Tech. Rep., 43, Toronto: Centre for Addiction and Mental Health, 2016.
[9] S.S. Co., "Social skills co.." Accessed Sept. 17, 2020.
[10] M.R. Ali, S.Z. Razavi, R. Langevin, A. Al Mamun, B. Kane, R. Rawassizadeh, L.K. Schubert, and E. Hoque, "A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons," Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA '20, New York, NY, USA, Association for Computing Machinery, pp.1–8, 2020.
[11] H. Tanaka, S. Sakriani, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, and S. Nakamura, "Teaching social communication skills through human-agent interaction," ACM Trans. Interact. Intell. Syst., vol.6, no.2, pp.1–26, Aug. 2016.
[12] H. Tanaka, H. Negoro, H. Iwasaka, and S. Nakamura, "Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders," PLOS ONE, vol.12, no.8, pp.1–15, Aug. 2017.
[13] H. Tanaka, H. Iwasaka, H. Negoro, and S. Nakamura, "Analysis of conversational listening skills toward agent-based social skills training," Journal on Multimodal User Interfaces, vol.14, no.1, pp.73–82, March 2020.
[14] T. Saga, H. Tanaka, H. Iwasaka, and S. Nakamura, "Objective prediction of social skills level for automated social skills training using audio and text information," Companion Publication of the 2020 International Conference on Multimodal Interaction, ICMI '20 Companion, New York, NY, USA, pp.467–471, Association for Computing Machinery, 2020.
[15] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, and L.P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, pp.1061–1068, 01 2014.
[16] I. Naim, M.I. Tanveer, D. Gildea, and M.E. Hoque, "Automated analysis and prediction of job interview performance," IEEE Transactions on Affective Computing, vol.9, no.2, pp.191–204, 2018.
[17] S. SC, Psychiatric interviewing: The Art of Understanding: A Practical Guide for Psychiatrists, Psychologists, Counselors, Social Workers, Nurses, and Other Mental Health Professionals, with online video modules, Elsevier Health Sciences, 2016.
[18] J.J. Silverman, M. Galanter, M. Jackson-Triche, D.G. Jacobs, J.W. Lomax, M.B. Riba, L.D. Tong, K.E. Watkins, L.J. Fochtmann, R.S. Rhoads, and J. Yager, "The american psychiatric association practice guidelines for the psychiatric evaluation of adults," American Journal of Psychiatry, vol.172, no.8, pp.798–802, 2015. PMID: 26234607.
[19] M. John N. Constantino and P. Christian P. Gruber, Social Responsiveness Scale, Second Edition (SRS-2) Back, Western Psychological Services, 2012.
[20] R. Takei, J. Matsuo, H. Takahashi, T. Uchiyama, H. Kunugi, and Y. Kamio, "Verification of the utility of the social responsiveness scale for adults in non-clinical and clinical adult populations in japan," BMC Psychiatry, Nov. 2014.
[21] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative," Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp.88–96, June 2014.
[22] E. Bleuler, Dementia Praecox, or the Group of Schizophrenias, New York: International Universities Press, 1911/1950.
[23] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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), Minneapolis, Minnesota, pp.4171–4186, Association for Computational Linguistics, June 2019.
[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," CoRR, vol.abs/1301.3781, 2013.
[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A.M. Rush, "Transformers: State-of-the-art natural language processing," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, pp.38–45, Association for Computational Linguistics, Oct. 2020.
[26] S.T. Kover, A.S. McDuffie, R.J. Hagerman, and L. Abbeduto, "Receptive vocabulary in boys with autism spectrum disorder: Cross-sectional developmental trajectories," Journal of Autism and Developmental Disorders, vol.43, no.11, pp.2696–2709, Nov. 2013.
[27] Y. Nakai, R. Takashima, T. Takiguchi, and S. Takada, "Speech intonation in children with autism spectrum disorder," Brain and Development, vol.36, no.6, pp.516–522, 2014.
[28] F. Martínez-Sánchez, J. Muela-Martínez, P. Cortés-Soto, J. Meilán, J. Ferrándiz, D. Egea-Caparrós, and I.M. Valverde, "Can the acoustic analysis of expressive prosody discriminate schizophrenia?," The Spanish Journal of Psychology, vol.18, pp.1–9, Oct. 2015.
[29] V. van Heuven, "Praat, a system for doing phonetics by computer," Glot International, vol.5, no.9/10, pp.341–345, 2001.
[30] M.(E.) Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R.W. Picard, "Mach: My automated conversation coach," Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp.697–706, Sept. 2013.

[31] M.R. Ali, K. Van Orden, K. Parkhurst, S. Liu, V.-D. Nguyen, P. Duberstein, and M.E. Hoque, "Aging and engaging: A social conversational skills training program for older adults," 23rd International Conference on Intelligent User Interfaces, IUI '18, New York, NY, USA, pp.55–66, Association for Computing Machinery, 2018.
[32] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp.59–66, 2018.
[33] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp.1–6, 2015.
[34] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.
[35] R. Takei, J. Matsuo, H. Takahashi, T. Uchiyama, H. Kunugi, and Y. Kamio, "Verification of the utility of the social responsiveness scale for adults in non-clinical and clinical adult populations in japan," BMC psychiatry, vol.14, no.1, pp.1–9, 2014.
[36] M.M. Channell, "The social responsiveness scale (srs-2) in school-age children with down syndrome at low risk for autism spectrum disorder," Autism & Developmental Language Impairments, vol.5, p.2396941520962406, 2020.
[37] K. Briot, F. Jean, A. Jouni, M.-M. Geoffray, M. Ly-Le Moal, D. Umbricht, C. Chatham, L. Murtagh, R. Delorme, M. Bouvard, M. Leboyer, and A. Amestoy, "Social anxiety in children and adolescents with autism spectrum disorders contribute to impairments in social communication and social motivation," Frontiers in psychiatry, vol.11, pp.710–710, July 2020.
[38] T. Sen, G. Naven, L.M. Gerstner, D. k Bagley, R.A. Baten, W. Rahman, K. Hasan, K. Haut, A.A. Mamun, S. Samrose, A. k Solbu, R.E. Barnes, M.G. Frank, and E. Hoque, "Dbates: dataset of debate audio features, text, and visual expressions from competitive debate speeches," IEEE Transactions on Affective Computing, p.1, 2021.

Takeshi Saga received his B.E. from the National Institute of Technology, Hachinohe College, Japan. He has completed a research internship at the Active and Attentive Vision Laboratory in York University, Canada, under Dr. John K. Tsotsos. Takeshi is now a master's candidate at the Nara Institute of Science and Technology, Japan, and an assistant researcher on the Tourism Information Analytics Team at the Center for Advanced Intelligence Project (AIP), Riken, Japan.

Hiroki Tanaka received his master's and Ph.D. degrees from the Nara Institute of Science and Technology, Japan, in 2012 and 2015, respectively. He is an Assistant Professor in the Graduate School of Information Science, Nara Institute of Science and Technology. His research interest is assisting people with disabilities through human-computer interaction.

Hidemi Iwasaka received his Ph.D. degrees from Nara Medical University. He is a physician working in child psychiatry. He was a Professor at the Nara University of Education. Currently, he is head of the Center of Child and Adult Development at Shigisan Hospital. His research interest is developing and evaluating psychosocial treatment, such as Parent Training or Social Skills Training for Developmental Disorders.

Satoshi Nakamura is a professor at the Nara Institute of Science and Technology, Team Leader of the Tourism Information Analytics Team, AIP Center, RIKEN, and Honorary Professor at the Karlsruhe Institute of Technology, Germany. He received his B.S. degree from the Kyoto Institute of Technology in 1981 and a Ph.D. from Kyoto University in 1992. He was the Director of ATR Spoken Language Communication Research Laboratories in 2000–2008 and Vice President of ATR in 2007–2008. He was the Director-General of Keihanna Research Laboratories, National Institute of Information and Communications Technology, in 2009–2010. He is currently a professor of the Augmented Human Communication Laboratory, Graduate School of Science and Technology, and the director at the Data Science Center, Nara Institute of Science and Technology, Japan. He is working on modeling and systems of spoken language processing, including speech-to-speech translation. He is one of the leading scientists in speech-to-speech translation research and has been serving various speech-to-speech translation research projects globally. He received the Yamashita Research Award, the Kiyasu Award from the Information Processing Society of Japan, the Telecom System Award, the AAMT Nagao Award, the Docomo Mobile Science Award in 2007, and the ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology from the Minister of Education, Science and Technology, and the Commendation for Science and Technology from the Minister of Internal Affairs and Communications. He also received the LREC Antonio Zampolli Award 2012. He was an elected Board Member of the International Speech Communication Association, ISCA, in 2011–2019, an IEEE Signal Processing Magazine Editorial Board member in 2012–2015, and an IEEE SPS Speech and Language Technical Committee Member in 2013–2015. He is an IEEE Fellow, ISCA Fellow, IPSJ Fellow, and ATR Fellow.