Teaching Social Communication Skills through Human-Agent Interaction

HIROKI TANAKA, SAKRIANI SAKTI, GRAHAM NEUBIG, and TOMOKI TODA, Nara Institute of Science and Technology
HIDEKI NEGORO and HIDEMI IWASAKA, Nara University of Education
SATOSHI NAKAMURA, Nara Institute of Science and Technology

There are a large number of computer-based systems that aim to train and improve social skills. However, most of these do not resemble the training regimens used by human instructors. In this paper, we propose a computer-based training system that follows the procedure of social skills training, a well-established method to decrease human anxiety and discomfort in social interaction, and acquire social skills. In this paper, we attempt to automate the process of social skills training by developing a dialogue system named the "automated social skills trainer," which teaches social communication skills through human-agent interaction. The system includes a virtual avatar that recognizes user speech and language information and gives feedback to users. Its design is based on conventional social skills training performed by human participants, including defining target skills, modeling, role-play, feedback, reinforcement, and homework. We performed a series of three experiments investigating 1) the advantages of using computer-based training systems compared to human-human interaction by subjectively evaluating nervousness, ease of talking, and ability to talk well, 2) the relationship between speech language features and human social skills, and 3) the effect of computer-based training using our proposed system. Results of our first experiment show that interaction with an avatar decreases nervousness and increases the user's subjective impression of their ability to talk well compared to interaction with an unfamiliar person. The experimental evaluation measuring the relationship between social skill and speech and language features shows that these features have a relationship with social skills. Finally, experiments measuring the effect of performing social skills training with the proposed application show that participants significantly improve their skill, as assessed by separate evaluators, by using the system for 50 minutes. A user survey also shows that the users thought our system is useful and easy to use, and interaction with the avatar felt similar to human to human interaction.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; K.3.1 [**Computing Milieux**]: Computers and Education

General Terms: Design, Performance, Human Factors

Additional Key Words and Phrases: Social skills training (SST), behaviour detection, dialogue system, embodied conversational avatar, computer-based training.

ACM Reference Format:
Hiroki Tanaka, Sakti Sakriani, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, Satoshi Nakamura, 2015. Teaching Social Communication Skills through Human-Agent Interaction. *ACM Trans. Interact. Intell. Syst.* 9, 4, Article 39 (July 2015), 26 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

This work is supported by the JSPS KAKEN 26540117 and Foundation for Nara Institute of Science and Technology.
Author's addresses: H. Tanaka, S. Sakriani, G. Neubig, T. Toda and S. Nakamura, Graduate School of Information Science, Nara Institute of Science and Technology, Takayama-cho 8916-5, Ikoma-shi, Nara, Japan. H. Negoro and H. Iwasaka, Center for Special Needs Education, Nara University of Education, Takabatake-cho, Nara-shi, Nara, Japan.

DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Many people have difficulties in social interactions such as presentations and job interviews [American Psychiatric Association 2013]. Persistent social skill deficits impede those afflicted with them from forming relationships or succeeding in social situations. Social skills refer to a variety of skills that are crucial for our everyday life and healthy development [Adi et al. 2007; Eisenberg et al. 2006], including skills such as those related to emotional, interpersonal, and communication skills.

Social skills training (SST) is a general cognitive behaviour therapy to train these social skills for people who have difficulties in social interaction, and is widely used by teachers, therapists, and trainers [Bauminger 2002]. However, SST requires well-trained teachers, so the number of participants joining SST program is restricted and applications are competitive.

On the other hand, if part or all of the SST process could be automated, it would become easier for those requiring SST to receive it anywhere and anytime. In addition, Donna Williams, who has an autism spectrum disorder, a severe case of social difficulties, wrote a book entitled "Nobody nowhere" (1992), in which she stated:

> *"The comprehension of words works as a progression, depending on the amount of stress caused from fear and the stress of relating directly. At best, words are understood with meaning, as with the indirect teaching of facts by a teacher or, better still, a record, television, or book. In my first three years in the special class at primary school, the teacher often left the room and the pupils responded to the lessons broadcast through an overhead speaker. I remember responding to it without the distraction of coping with the teacher. In this sense, computers would probably be beneficial for autistic children once they had the skills to use one."*

Thus, a large number of training methods using computers or other forms of technology have been proposed [Bishop 2003; Moore et al. 2000; Parsons and Mitchell 2002; Schuller et al. 2014]. However, most of these do not follow the conventional training framework used by human instructors. Slovák et al. [2015; 2015] have noted that understanding how humans learn social skills, and how this process can be supported by technology, is an important but under-researched area in human-computer interaction (HCI). These papers also summarized existing approaches to social and emotional skills learning (SEL) in a large number of evidence-based programs, and address the gap between human-based training and existing computer-based training methods, noting that much of the existing HCI work has not, so far, been connected to SST [Slovák and Fitzpatrick 2015].

In this paper, we make some first steps towards closing this gap by proposing a novel tool that tries to replicate conventional SST using a systematic and computer-based design. We develop a dialogue system named "automated social skills trainer," which is an application including video modeling of human behaviour and real-time behaviour detection as well as data visualization to help people improve their social skills (Figure 1). We investigate whether it is possible to help people who have difficulties in social interaction improve their social skills using an automated system which can be used anywhere, anytime.

The main contribution of this paper is that it is the first attempt to strictly follow the well-known and well-tested regimen of SST in the HCI context. In order to achieve this goal, we make several technical contributions:

— Creating a new dialogue system following the conventional SST framework.
— Sensing new behavioural parameters which related to the social skills, such as language and speech features.

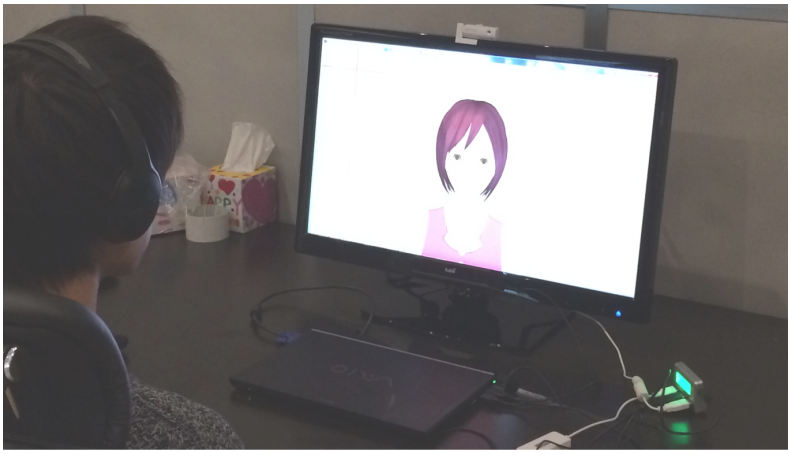

Fig. 1. SST with the automated social skills trainer.

— Providing a new feedback scheme with not only summary visualization but also quantified objective values and generated positive comments.

In order to validate the utility of the method, we perform experiments with up to 38 human participants. The experimental results show that type of interaction is related to the users subjective impression of nervousness, ease of talking, and ability to talk well, that social skill is related to automatically extracted features and has a relationship to autistic traits measured by standardized tests, and that participants improve in social skill using the automated social skills trainer.

This paper is an extended version of a conference paper [Tanaka et al. 2015]. Major extensions are the first experiment (regarding subjective impressions regarding interaction with an avatar or a human), detailed analysis of feature differences, and measuring inter-rater agreement in the second experiment. In section 2, we describe related works of this field. In section 3, we explain each module of conventional human-based SST. In section 4, we explain the correspondence between human-based SST and the automated social skills trainer. In section 5, we describe implementation details of the automated social skills trainer focusing on role-playing and feedback. In sections 6 to 8, we perform three experiments to evaluate the system. In section 9, we summarize and discuss future directions.

2. RELATED WORK

The design of an automated social skills trainer brings together several fields, including research into computers in education, intelligent virtual agents, and affective computing. The following paragraphs briefly outline work in these areas.

The use of computers in SST is motivated by the fact that while individuals with social impairments have difficulty in social interaction, they also show good and sometimes even superior skills in “systemizing” [Baron-Cohen et al. 2003]. Systemizing is the drive to analyze or build systems, and to understand and predict behaviour in terms of underlying rules and regularities. The use of systematic computer-based training for people who need to train social skills can take advantage of the facts that: 1) they favor the computerized environments because they are predictable, consistent, and free from social demands, 2) they can work at their own speed and level of understanding, 3) training can be repeated over and over again until the goal is achieved,

Table I. The basic training model of SST.

Num.	Procedure
1	Defining target skills
2	Modeling
3	Role-play
4	Feedback
5	Reinforcement
6	Homework

4) interest and motivation can be maintained through computerized rewards [Bishop 2003; Moore et al. 2000; Parsons and Mitchell 2002; Schuller et al. 2014].

There has been one previous line of work on automated conversational coaches [Hoque et al. 2013; Hoque and Picard 2014], which are dialogue systems aimed to train people for improving interview skills through real-time feature detection and feedback. They achieved 1) a realistic task involving training real users, 2) formative affective feedback that provides the user with useful feedback on the behaviours that need improvement, and 3) the interpretation or recognition of user utterances to drive the selection of backchannels or formative feedback. While this work is an excellent first step, it did not faithfully follow the traditional SST framework.

The previous work omitted steps such as 1) modeling of human behaviour e.g. [Essau et al. 2014], 2) providing feedback compared to model speakers to objectively confirm user's strengths and weaknesses, and 3) giving positive reinforcement to enhance user's self-esteem [Bellack 2004]. These steps have clear goals and definitions in traditional human-provided SST [Bauminger 2002; Liberman and Wallace 1990; Wallace et al. 1980], such as helping users to understand their current social communication skills and the path to their goals, maintaining positive motivation. We hope that by following this paradigm, the above effects can be reflected in automated social skills training. Thus, we attempt to follow the traditional SST framework as closely as possible.

3. SOCIAL SKILLS TRAINING

Conventional SST is an established method originally developed by Liberman for schizophrenics to reduce their anxiety and discomfort in social interaction [Liberman and Wallace 1990; Wallace et al. 1980]. SST is often performed with multiple sessions, and each session focuses on the training of one target skill for one or two hours. It is well known that SST can be used to effectively improve social skills for people with social disorders such as autism spectrum disorders (ASD) [Bauminger 2002].

SST can be classified into individual (one to one training) and group (one to many or many to many training) settings. One advantage of group SST is that it enables participants to observe other participants' behaviour and also receive feedback from others. On the other hand, the advantage of individual SST is that the training can be relaxed and comfortable for participants, and that lessons can be tailored to the individual's needs.

As shown in Table I, the basic training model of SST [Bellack 2004] is generally based on the following steps: defining target skills, modeling, role-play, feedback, reinforcement, and homework. We briefly describe these steps as follows:

— **Defining target skills:** The major social problem is identified, and the skills to be trained are decided based on this problem or to be selected from basic skills such as [Bellack 2004]. In order to figure out the major problems, the participants and trainers work together through discussion or trainers decide the target skills. Once the target skills are decided the trainers decide the goal after intervention. In this

step, trainers sometimes use related books to help participants understand target skills and the goal of SST. Examples of target skills include presentation skills, job interview skills, self-introduction skills, or skills regarding how to decline another's offer or request.

— **Modeling:** Before participants are asked to perform an interaction, trainers act as a model, demonstrating the skill that the participants are focusing on so that participants can see what they need to do before attempting to do it themselves. For example, trainers may show a good story telling example using appropriate verbal and non-verbal cues.
— **Role-play:** Participants are asked to role-play. For example, participants tell their experience to the trainer. This allows the participants to practice their own skills in the target situation. Trainers observe participants' social skills subjectively, but mainly focus on voice quality, amplitude, facial expression, eye-gaze and other factors. This practice is a very important aspect of SST.
— **Feedback:** Trainers provide feedback at the end of role-playing (in the case of group SST, participants also receive feedback from other participants). This feedback helps participants to identify their strengths and weaknesses. For example, trainers may tell the participant that the role-play was very good because he/she used appropriate voice amplitude.
— **Reinforcement:** Trainers give positive reinforcement, praising the participant about their achievement of targeted behaviours, in addition to feedback. This is a very important aspect of the SST, because the participants often do not have confidence in social interaction, and tend to have low self-esteem. Therefore, positive encouragement helps build confidence in social environments.
— **Homework:** Trainers set little homework challenges that participants are required to do in their own time throughout the week. For example, trainers may ask the participant to tell their story to friends or family, and let the trainer know about the result.

By performing this training, participants can learn better social skills in a number of different ways, a core aspect contributing to the effectiveness of training. However, it should be noted that the human trainer plays a very involved role in the majority of these steps. As a consequence, SST requires professional or at least well-trained trainers satisfying the above abilities (e.g. being able to perform modeling and give appropriate feedback comments). The number of skilled trainers is small, and thus the number of participants joining SST program is restricted and applications are competitive.

4. AUTOMATED SOCIAL SKILLS TRAINING

In this section, we describe our proposed automated social skills trainer following the conventional individual SST framework. We replicate human-to-human individual SST by using a spoken dialogue system. While one disadvantage of individual SST is that there is no chance to see other participants' behaviour, we can provide a surrogate by playing video of others on the computer screen. The automated social skills trainer interacts with users by speech. In the current system, we simply selected fixed templates, which are filled with the appropriate content based on the sensing results.

Table II shows the correspondence between conventional SST and our proposed method. In the following few pages, we describe each proposed module corresponding to a step in conventional SST.

— **Defining target skills:** Ideally, we would define specific target skills for each user through an initial interaction with the system. We plan to tackle this in future work, but for the time being we focused on a single target skill that has been shown to be

Table II. Correspondence between conventional SST and our proposed method.

Conventional SST	Proposed Method
Defining target skills	Story telling/narrative
Modeling	Playing video of others
Role-play	Conversation with an avatar
Feedback	Generation of visual feedback
Reinforcement	Generation of positive comments
Homework	Spoken instructions

Fig. 2. An example of video modeling.

widely applicable: story-telling or narrative ability (one of four basic skills that is targeted in traditional SST [Bellack 2004]). Story telling/narrative is a task of telling memorable stories, used widely in other situations [Lieberman 1987], and related to social interaction skills such as presentations and job interviews [Davis et al. 2004]. It has also been shown useful to distinguish children with social difficulties and children with typical development [Tanaka et al. 2014]. In the step of defining this goal, if the user says a keyword "please explain," the system tells the user that "I will help you learn to tell stories well, and after training you will have more fun telling stories."

— **Modeling:** Users can watch a recorded model video (Figure 2). The recorded models are people who have relatively good narrative skills according to subjective evaluation. Users can watch and imitate the good examples. If the user says a keyword "good example," the system tells the user that "I will show you good examples" and plays the video on the screen.

— **Role-play:** The main technical part of the proposed system is the role-playing, which is performed through interaction with an avatar. When the user says "start role-playing," the system says "Please tell me your recent memorable fun story." Role-playing starts after the avatar's utterance, and continues for one minute. The system shows a small timer on the upper right of the screen, so the user can monitor how much time they have left. After one minute of role-play, the system says "That's all. Please wait for feedback." During this time, the avatar nods its head, and the system detects and analyzes language and speech features automatically. In this work, we

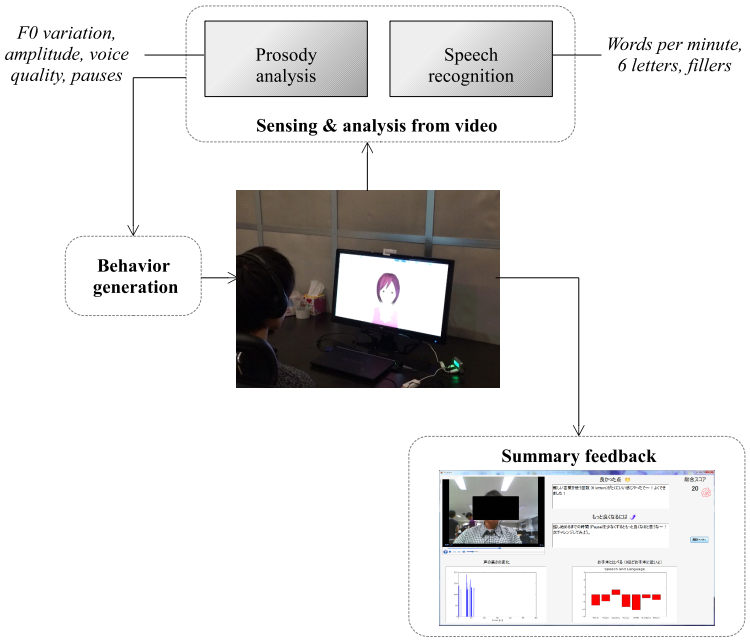

Fig. 3. The automated social skills trainer framework.

focus on features that could differentiate between Japanese people with and without social difficulties as described in [Tanaka et al. 2014]: F0 variation, amplitude, voice quality, pauses, words per minute, words with more than 6 letters, and fillers. We show a list of these features as follows:

F0 variation: F0 indicates fundamental frequency, or pitch of the voice. F0 variation therefore corresponds to the amount of variety in pitch, with less variety corresponding to a more monotone voice. We use these features because it has been widely noted that people with social difficulties have prosody that differs from that of their peers [Kanner et al. 1943; McCann and Peppé 2003; Bone et al. 2012; Kiss and van Santen 2013; Santen et al. 2013]. For instance, Kiss et al. [2012] found several differences in the fundamental frequency characteristics of people with social difficulties.

Amplitude: In human-to-human SST, trainers often focus on volume of voice because both overly small and loud voices are not appropriate for many social situations. A previous study investigating the amplitude (power) of people with social difficulties [Tanaka et al. 2014] confirmed the importance of amplitude to identify children with social difficulties.

Voice quality: People with social difficulties often exhibit abnormal voice quality, often described as clearer than their peers. Bonneh et al. [2010] quantified speech abnormalities in terms of the properties of the voice quality and was able to identify children with social difficulties with more than 80% accuracy.

Pauses: There are reports finding that children with social difficulties tend to delay responses to their parent more than children with typical development in natural conversation [Heeman et al. 2010].

Words per minute (WPM): There is a report that speaking rate was strongly correlated to interview skills [Hoque et al. 2013]. WPM is related to frequency of speech.

Words with more than 6 letters: Individuals with social difficulties use more complicated or unexpected words than typically developing people, and deficits of social difficulties affect inappropriate usage of words [Rouhizadeh et al. 2013]. Words with more than six letters may be related to complicated words [Pennebaker et al. 2007].

Fillers: The frequency of filler usage is important in story telling or presentation. Too frequent use of fillers disturbs listener focus on the contents of speech.

— **Feedback:** The system displays summary feedback according to detected features. The feedback includes comments, the user's video, the parameters compared to model speaker, and the quantified overall narrative score. The user can objectively confirm their strengths and weaknesses. In 5 seconds after finishing role-play, the system tells the user that "I will show your result" and displays feedback on the screen.

— **Reinforcement:** In addition to simply looking scores, the system also chooses good parts of the narrative, and gives positive feedback encouraging the targeted behaviour.

— **Homework:** If the user says a keyword "homework," the system tells the user that "Please tell your story to others throughout the week, and let me know about it." However, it should be noted that we did not evaluate SST across sessions in this work, so the homework is not considered in the current version of the system.

Through this framework, we can replicate to some extent conventional individual SST with the spoken dialogue system replicating each module in the framework.

5. IMPLEMENTATION DETAILS

The automated social skills trainer system works on a regular laptop, which processes the audio input in role-playing. The processed data is used to generate the behaviours of an avatar that interacts with and provides feedback to users.

The role-playing, feedback, and reinforcement consist of three modules: behaviour generation, sensing & analysis from video, and summary feedback as shown in Figure 3. The following subsections describe the modules in detail. It should be noted that the target language of our system is Japanese, and all data creation and experiments are performed with native Japanese speakers.

5.1. Data creation and subjective evaluation

As a first step towards building our system, we collected video data from a total of 19 people. This video data is used both in the modeling module and for predicting scores in the feedback module. Using this data, we assigned each dialogue an overall narrative skill score based on subjective evaluation, and the top five people were selected as models. Subjective evaluation was performed by having two raters watch the recorded participants' narrative and answer a questionnaire. This process is described in more detail in the following Experiment 2.

5.2. Dialogue agent

The automated social skills trainer was developed using MMDAgent[1], which is a Japanese spoken dialogue system integrating speech recognition, dialogue management, text-to-speech, and behaviour generation. MMDAgent works as a Windows application. We selected an animated character who is similar to an actual human, as opposed to an animal-like character, but not more realistic human-like one, as we hope that this will make the conversation interesting, and make it easier for the user to generalize learned skills in a real situation. On the other hand, we chose a clearly

[1]http://www.mmdagent.jp/

Fig. 4. The avatar used in the automated social skills training system.

animated character instead of a more realistic, life-like avatar. Teaching studies have suggested that people with social difficulties show greater improvements in emotion recognition when computer programs include cartoons rather than photographs of real faces [Silver and Oakes 2001]. In contrast, students who don't have social difficulties had greater transfer of learning when the agents had more realistic images [Baylor and Kim 2004], and Hoque et al., [2013] made human-like agents to elicit stress that can help in creating a realistic interview experience. A comparison of different types of agents is beyond the scope of our work, but is an interesting direction for the future.

The avatar is displayed from the front, and there are no distractions in background (Figure 4). The user can operate and interact with the avatar by using speech throughout the training. All dialogue system utterances were created using templates written by the first author. A dialogue example is shown in Table III.

Table III. A dialogue example.

Speaker	Content
User	Hello
System	Hello
User	Please explain today's skill?
System	I will help you learn to tell stories well.
User	Can I watch a good example?
System	I will show you good examples. When you say stop, it will be closed.
User	(Watching) Stop.
User	Please start the role-playing.
System	Please tell me about a recent memorable event. Go ahead.
User	(Speaking)
System	That's all. Please wait for feedback.

In addition, the system performs a number of behaviours to keep the user engaged. It blinks its eyes once every three seconds, and reacts to users during the role-playing. When the system recognizes an utterance, after a few seconds the system nods its head. The blinking and nodding behaviour motions were created by MikuMikuDance[2].

[2]http://www.geocities.jp/higuchuu4/

Fig. 5. The summary feedback provided by the automated social skills trainer (translated from Japanese).

5.3. Sensing and analysis from video

To calculate the linguistic features, we performed automatic speech recognition (ASR) using the Julius dictation kit[3]. We used Mecab[4] for part-of-speech tagging in Japanese utterances. For speech feature extraction, we used the Snack sound toolkit[5].

The implementation of features is as follows: 1) F0 variation: We used the coefficient of variation for fundamental frequency with a minimum pitch of 100 Hz. We did not use mean, maximum and minimum values because there are individual and gender differences in terms of these features, 2) Amplitude: We used the mean value of amplitude, 3) Voice quality: We extracted the spectral tilt by calculating the difference between the first harmonic and the third formant "h1a3" as a feature expressing voice quality [Hanson 1995], 4) Pauses: We calculated values of pauses before new turns as time between the end of the avatar's utterance and the start of the user's utterance, 5) WPM: In the automated social skills trainer, the narrative continues for one minute, and we counted the number of words in one narrative, 6) Words with more than 6 letters: We extracted percentage of words with more than 6 letters as a feature, 7) Fillers: We calculated percentage of fillers such as "umm," or "eh" in Japanese. This feature is automatically extracted using the output of the part-of-speech tagger.

5.4. Summary feedback

Based on the calculated features, we provide feedback to the users about their social skills (Figure 5). Our goal was to design visualizations so that it would be easy for users to understand and interpret their narrative skill. The summary feedback provides the following information.

— **User video:** Participants can watch the recorded video and audio of the narrative. In doing so, the user can confirm their verbal and non-verbal information, including speech contents, facial expression, and posture [Hoque et al. 2013].

[3]http://julius.sourceforge.jp/index.php
[4]https://code.google.com/p/mecab/
[5]http://www.speech.kth.se/snack

— **Overall score:** It has been noted that people who have social communication difficulties often prefer to check their improvement quantitatively [Baron-Cohen et al. 2003]. The system displays the predicted overall score, which may help motivate the user to practice more and improve their score. We predict the overall score using the multiple regression method on a scale of 0 to 100, in which 0 is the minimum and 100 is the maximum score for overall narrative skill. More details of the prediction model are described in Section 7.6.
— **Pitch variation:** Participants can see their pitch movement corresponding to the time. This also doubles as visualization of how frequently they spoke.
— **Comparison with models:** The system visualizes a comparison of extracted features between the user's current narrative and model persons' narratives in terms of z-score, which is a statistical measurement of a score's relationship to the mean in a group of scores. The users are informed that they should attempt to emulate the model in all aspects.
— **Good points:** The system generates positive comments that reinforce the user's motivation with encouraging words [Bauminger 2002]. The comments are generated based on the features that have values close to those of the models, and are embedded into a fixed template (*** is very good! Well done!)
— **Points to be improved:** The system generates comments about points to be improved for the next trial. The comments are generated based on the features that have values far from the models, and are embedded into a fixed template (You should reduce/increase ***. Try next time!)
— **Screenshot:** Participants can save the feedback by clicking a button, and this is used for checking improvements over the course of training.

6. EXPERIMENT 1: SUBJECTIVE IMPRESSION OF INTERACTION WITH AGENTS AND HUMANS

Human-based SST is usually performed by interaction between participants and trainers, who are likely to be unfamiliar to the participants in the earlier sessions. This interaction with an unfamiliar person may cause nervousness, anxiety and discomfort. Previous work has noted differences of heart rate regulation and temporal-parietal electroencephalogram (EEG) activity during viewing of familiar and unfamiliar persons [Van Hecke et al. 2009], with children with social communication difficulties having lower overall heart rate levels when viewing videos of a unfamiliar person, compared to control children.

In the first experiment, we sought to examine the effect of the interaction partner in both human-human and human-computer interaction by investigating the relationship between three types of interaction (interaction with an avatar, a familiar person, and an unfamiliar person) and subjective impressions of ease and quality of interaction.

6.1. Procedure

15 graduate students (14 males and 1 female) participated in the experiment[6]. All participants performed interactions with the proposed avatar and human counterparts, and were told that their speech and video would be recorded. A webcam (ELECOM UCAM-DLY300TA) placed on top of the laptop and headset (ELECOM HS-HP168K) recorded the video and audio of participants. In the human-human interaction setting, familiar or unfamiliar persons listened to the speaker's story and nodded according to the speaker's utterances. The familiar person (male) was selected from same insti-

[6]Note that the Research Ethic Committee of our institution has reviewed and approved both this and the following experiments. Written informed consent was obtained from all subjects before the experiments.

tution, and he knows each participant. The unfamiliar person (male), who was also selected from same institution, did not know any of the participants.

In both of HHI and HCI settings, the participants were directed to consider a story in advance and keep speaking for 1 minutes. The system shows a timer, and a small timer was placed in front of the participants in the case of HHI.

It should be noted that like Van Hecke et al. [2009], we only have a single interaction partner for each of the familiar and unfamiliar classes, but the unfamiliar and familiar persons were matched for gender, age, race, hair color and style, presence of glasses, and social communication skills. Social communication skills were measured by the Autism-spectrum quotient (AQ) [Baron-Cohen et al. 2001], and we confirmed that the unfamiliar and familiar persons obtained scores of 4 and 6 for the sums of the AQ subarea scores for communication and social skills. This suggests that they have similar levels of social skills. Figure 6 shows the recording setting of the interaction. Participants waited in a quiet room and then the human counterpart entered the room. Participants and human counterparts were instructed to avoid speaking and making eye contact before interaction. Participants also interacted with the avatar described in Section 5. During all interactions, we also measured eye-tracking patterns and electrodermal skin activity (EDA) using the Tobii X2-30 and the Affectiva Q sensor respectively, but we reserve a detailed analysis of this data for future publications. An eye-tracker was used only in the HCI setting, and in the HHI setting participants were directed to look at the person in front of them. We also collected data of HHI where the familiar person was shown on the laptop screen. Ratings according to the questionnaire described in the following were not significantly different ($p > .05$) between these settings according to Student's t-test. We reserve a detailed analysis of this data for future publications.

Based on this interaction, we examined whether the subjective impression of the participant about ease and quality of interaction is related to the three types of interaction (interaction with the avatar, the familiar person, and the unfamiliar person). To reduce bias, the order or each type of interaction was randomly selected for each participant. After each interaction, the participants were asked to answer a questionnaire about the following, rated on a scale of 1 (I did not feel so, or I disagree) to 7 (I feel so, or I agree).

Q1 Nervousness
Q2 Ease of talking
Q3 Ability to talk well

The effect of type of interaction was analyzed using repeated-measures analysis of variance (ANOVA). Post-hoc comparisons between each of the interaction types were performed using Bonferroni's method [Dunn 1961].

6.2. Subjective Evaluation

Figure 7 shows the relationship between subjective evaluation and type of interaction.

6.2.1. Nervousness. With regards to nervousness (left side of Figure 7), the results showed that type of interaction has a significant effect on subjective impressions ($F[2,28]=14.72$, $p < .05$) with $\eta_p^2 = .51$ according to ANOVA. Comparisons showed that subjects were significantly less nervous when interacting with either the avatar or the familiar person than they were when interacting with the unfamiliar person ($p < .05$). The difference between interaction with the avatar and interaction with the familiar person was not judged as statistically significant ($p > .05$).

These results showed that the participants felt nervousness when interacting with the unfamiliar person, and were less nervous when interacting with the avatar and

Fig. 6. Data recording settings of interaction between participant (seated left side) and familiar/unfamiliar person (seated right side).

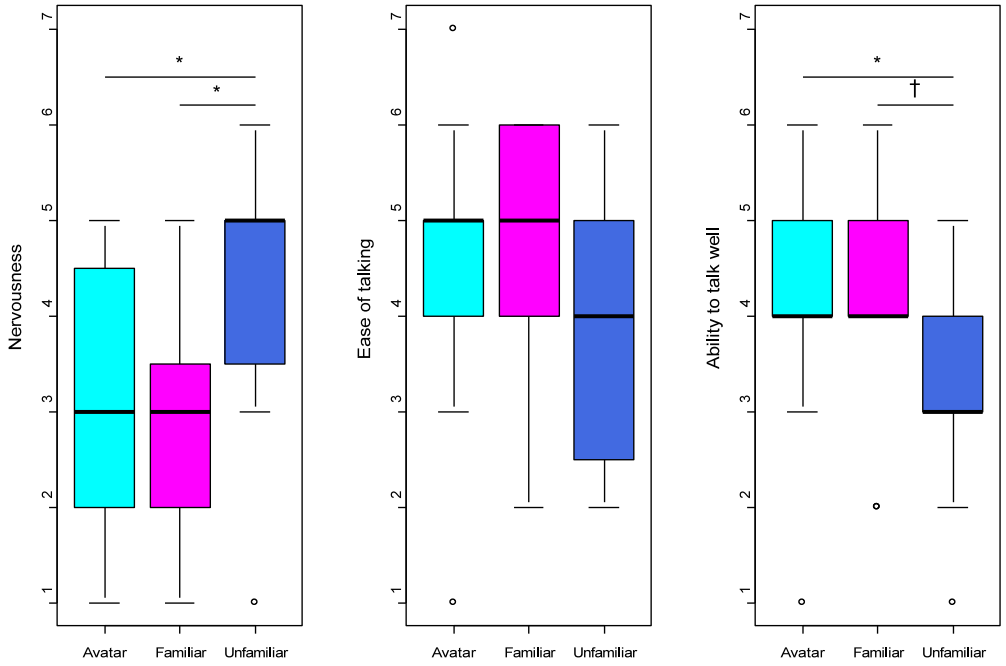

Fig. 7. Boxplot of subjective impressions corresponding to type of interaction. The bottom and the top of the box are the 25th and 75th percentile (the lower and upper quartiles, respectively), and the band near the middle of the box is the 50th percentile (the median). The lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile. Outliers were plotted as individual circle points (*: $p < .05$, † : $p < .1$).

the familiar person. This suggest that computer-based training may allow users to feel more comfortable than human-human interaction with an unfamiliar person, and that human-computer interaction may be an alternative for people with social communication difficulties to become involved in SST for the first time.

6.2.2. Ease of talking. With regards to ease of talking (middle side of Figure 7), the results showed that type of interaction did not have a significant effect on subjective impressions (F[2,28]=2.73, $p > .05$) according to ANOVA. Comparisons also showed that the differences of type of interaction were not significant ($p > .05$).

These results showed that type of interaction was not related to the ease of talking, because ease of talking is sometimes affected by counterpart's behaviours (e.g. timing of nodding and smiling), which were not carefully controlled for in the experiment.

6.2.3. Ability to talk well. With regards to ability to talk well (right side of Figure 7), the results showed that type of interaction has a significant effect on subjective impressions (F[2,28]=4.41, $p < .05$) with $\eta_p^2 = .24$ according to ANOVA. Comparisons showed that subjects were significantly more able to talk well when interacting with the avatar than when interacting with the unfamiliar person ($p < .05$), and tended to have more ability to talk well when interacting with the familiar person than when interacting with the unfamiliar person ($p < .1$). The difference between interaction with the avatar and interaction with the familiar person was not judged as statistically significant ($p > .05$).

These results showed that the participants felt they were able to talk better when interacting with the avatar compared to interacting with the unfamiliar person. The results suggest that using avatars in SST allowed the participants to use their conversational skills more effectively.

7. EXPERIMENT 2: DEFINING MODEL PERSONS

In the second experiment, we sought to answer the following questions:

1) Does narrative skill relate to linguistic, acoustic, and other information?
2) Is there a difference between talking to humans (human-human interaction: HHI) and talking to avatars (human-computer interaction: HCI) in terms of narrative?
3) Does narrative skill relate to autistic traits?
4) Are the extracted features effective for identifying narrative skill?

The result of the second experiment is used in data collection and summary feedback of the automated social skills trainer.

7.1. Procedure

We recruited 19 graduate students (16 males and 3 females), all of whom were native Japanese speakers. All subjects used the proposed system and were told that their speech and video would be recorded. A webcam (ELECOM UCAM-DLY300TA) placed on top of the laptop and headset (ELECM HS-HP168K) recorded the video and audio of participants. We recorded not only HCI but also an HHI setting in which the first author listened to the speaker's story and nodded his head according to the speaker's utterances. The small timer was placed in front of the participants in the case of HHI. The same 19 participants participate in the HCI and HHI. For HCI, participants interact with the same avatar.

To get a grasp of each subject's social skills independent of the proposed system, or the narrative setting in general, we also administered a social skills test for each subject. Specifically, we measured the sum of subarea scores for communication and social skills of Japanese version of AQ[Wakabayashi et al. 2006] which is a standard tool to measure autistic traits with a total of 50 questions including 5 subareas.

Next, we had raters watch the interactions of each participant and rate their narrative skill. Although it would be ideal for raters to be professional social skills trainers, they are few and far between, so it is difficult to recruit them for the experiment. Thus, as a proxy, we selected raters from members of the general population. Because raters

are required to have good social skills to recognize users' non-verbal expressiveness, we selected two people (male and female) with good social skills as annotators. Specifically, the annotators were selected to have low sums of the AQ subarea scores for communication and social skills (where lower indicates better social skills). The sums of both areas were 1 and 4, which is lower than the mean value of 7.6 for Japanese students [Wakabayashi et al. 2006]. The raters did not participate in the experiments as subjects. The raters did not know the recorded participants, and were trained by rating several examples prior to the evaluation. Two raters watched recorded participants' narrative for both HHI and HCI, and answered a questionnaire[7], which is based on [Hoque et al. 2013]. The questionnaire included the following items related to the participant's overall narrative performance and use of non-verbal cues such as intonation, amplitude, and lexicon usage, rated on a scale of 1 (not good, not appropriate, or small (few)) to 7 (good, appropriate, or large (frequent)).

Q1 Overall narrative skill: How good overall narrative skill of him/her?
Q2 Concentration: Did he/she focus on talking?
Q3 Friendliness: Did he/she speak in a friendly manner?
Q4 Appealingness: Did you find the way that he/she talked to be appealing?
Q5 Speaking rate: Was his/her speaking speed? (fast $< - >$ slow)
Q6 Usage of fillers: Did he/she frequently used fillers?
Q7 Intonation: Did he/she speak a story with appropriate intonation?
Q8 Voice quality: Was his/her narrative easy to listen to?
Q9 Amplitude: Did he/she speak a story with appropriate power?
Q10 Usage of easy words: Did he/she use easy words?

7.2. Agreement

The agreement of the two raters was measured by Cohen's kappa-coefficient, which calculates agreement beyond chance by distinguishing the observed agreement (A_{obs}) from the agreement by chance (A_{ch}), as follows:

$$\kappa = (A_{obs} - A_{ch}) \,/\, (1 - A_{ch})\,. \tag{1}$$

The two raters answered the ten questions for each speaker's narrative. For each rater, each subject was assigned a class of being either above or below the average score for the rater, and agreement between the classes was used to calculate the coefficient. The Kappa coefficient of two classes for two raters (κ_{Q1}) was 0.58, which corresponds to moderate agreement according to the scale proposed by [Rietveld and Van Hout 1993].

Agreement of all other questions is listed in Table IV. This table shows that agreements of questions of speech and language were moderate. Specifically, the questions about usage of easy words and voice quality have relatively good agreement. In contrast, the questions about concentration, friendliness, and attractiveness have poorer agreement, which is understandable given the subjective nature of these terms.

7.3. Correlation between questions

Figure 8 shows the correlation matrix of each question. For Q6, because usage of fillers can be assumed to be inversely proportional to social skill, we inverted the ratings before measuring correlation. The result showed that questions especially asking about the speech and language features were significantly related to overall narrative skill. On the other hand, questions asking about concentration were not related to overall narrative skill and other features. We can also see that questions related to speech and language features were correlated each other.

[7]http://goo.gl/forms/F1qDLnWyY3

Table IV. The Kappa coefficient between the two raters.

Question	Kappa coefficient
Q1: Overall narrative skill	0.58
Q2: Concentration	0.26
Q3: Friendliness	0.32
Q4: Intriguing	0.32
Q5: Speaking rate	0.43
Q6: Usage of fillers	0.40
Q7: Intonation	0.49
Q8: Voice quality	0.63
Q9: Amplitude	0.45
Q10: Usage of easy words	0.75

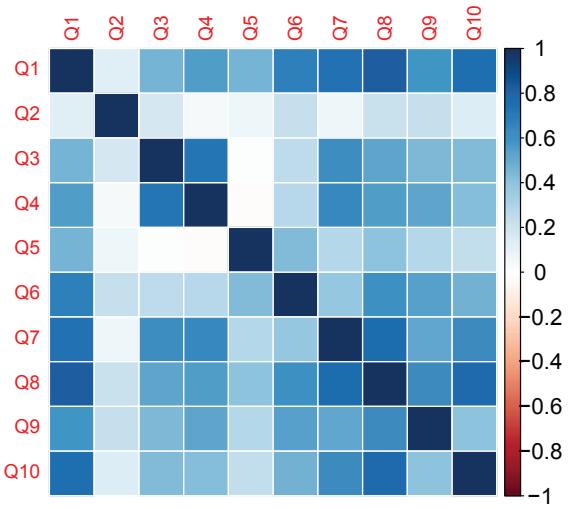

Fig. 8. Pearson's r correlations between various questions. Color indicates the strength of statistically significant correlations, and white indicates zero. Rows and columns represent the questions in the same order, so the diagonal is self-correlation.

7.4. Differences between human and computer interaction

To examine the difference between HHI and HCI, we show averaged rater scores for HHI and HCI in Figure 9. We can see that there were differences between HHI and HCI, and raters' scores of overall narrative skill in HHI were slightly higher than HCI. However, we did not find a statistical difference ($p > .05$) by Student's t-test. It is likely that if differences exist between interaction with our proposed system and interaction with an actual human in terms of overall narrative skills, they are small.

7.5. Model people and autistic traits

Based on the raters' scores, we determined the top 5 of 19 subjects to be our models for additional experiments including the modeling step of SST. As shown in Figure 10, the median value of the AQ was 1 in the case of model persons, and 13 in the case of the others. This indicates that there is also a strong relationship between the raters' assessment of narrative skill and the subjects' answers on the AQ test.

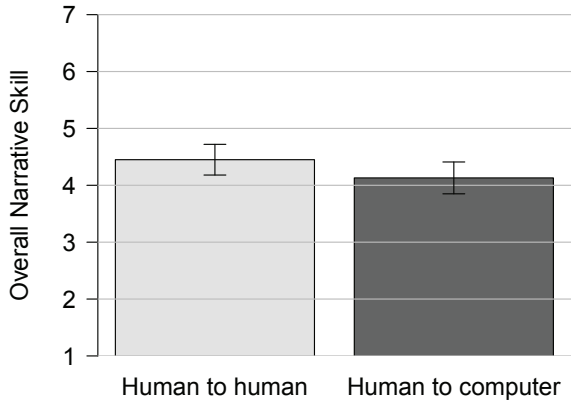

Fig. 9. The difference of raters' scores between HHI and HCI. Error bars indicate standard error.

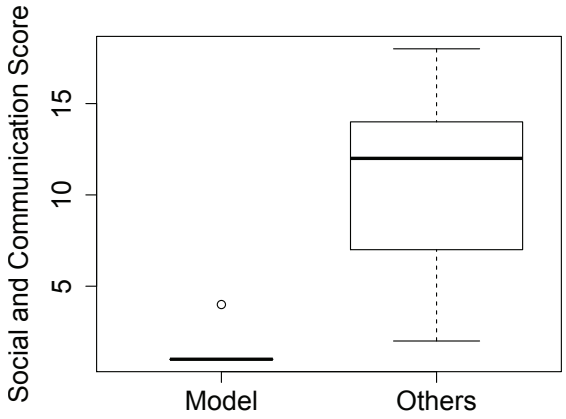

Fig. 10. The ranges of the AQ for model persons and others. Zero indicates high social and communication skills, and 20 indicates low social and communication skills.

7.6. Regression

We calculated the statistical differences of the automatically extracted features between those chosen as models due to their good speaking skills and others using Student's t-test. We found that WPM, words of more than 6 letters, and amplitude were significantly different between the groups ($p < .05$), and other features were not significantly different ($p > .05$). Table V shows that model persons speak more loudly and frequently. In contrast, the non-models more frequently use words with more than 6 letters.

Table V. Difference of mean values between model persons and others based on language and speech features from their utterances. Each table cell notes which of the two classes has the greater mean on the corresponding feature (*: $p < .05$).

F0 variation	Amplitude	Voice quality	Pauses	WPM	Words with more than 6 letters	Fillers
-	Model*	-	-	Model*	Others*	-

We used these three features to predict overall narrative skill using the multiple regression method. To evaluate how well data fit the regression model, we calculated the multiple R-squared value, which was 0.51 indicating a good prediction model based on [Cohen et al. 2013]. We also found that the correlation between the predicted value and actual value using leave-one-user-out cross validation was also 0.51 ($p < .05$), showing a weak correlation.

This regression model was integrated into the system as the feedback module's overall score.

8. EXPERIMENT 3: SOCIAL SKILLS TRAINING

In the third experiment, we examined whether the automated social skills trainer is effective to train social skills, specifically:

1) How effective is the automated social skills trainer in helping users improve their narrative skills?
2) Do users find the automated social skills trainer easy to use and helpful?

8.1. Procedure

We recruited a total of 36 graduate students (27 males and 9 females) all of whom were native Japanese speakers, different from those who participated in the first and the second experiments. Participants first entered the experiment room, and were given instructions by the first author. All subjects were told that their speech and video would be recorded. A webcam (ELECOM UCAM-DLY300TA) placed on top of the laptop and headset (ELECM HS-HP168K) recorded the video and audio of participants.

We separated participants into 3 groups: the reading book group (11 males and 1 female), the video modeling group (8 males and 4 females), and the feedback group (7 males and 5 females). The reading book group, which serves as a control, read two types of social skills books which were related to story telling/narrative skills. The book titles were "Social skills training: collection of cases" and "The easiest guide to presentations" (in Japanese). These two books can be read within 50 minutes. We marked the most important chapters in the books, and the participants were directed to read from the marked chapters. The video modeling group and the feedback group used the automated social skills trainer for their training. The video modeling group only watched the model videos, while the feedback group performed role-play and received automated feedback. The feedback group can also watch the model videos which include same contents of the video modeling group. Because we did not control the duration of watching videos or performing role-play, the participants can select the content by themselves. As shown in Figure 11, all subjects spoke their narratives to the agent (pre), received training for 50 minutes, and spoke their narratives to the agent again (post).

The same two raters from the second experiment evaluated the subjects' narrative skill by only answering overall narrative skill (Q1 of the second experiment) rated on a scale of 1 to 7. Raters did not know subjects, and the order of the pre- and post-training narratives was randomized to prevent bias. We averaged the two raters' scores and calculated improvement in score (post - pre) for each group. Note that the initial scores of the reading book group (mean: 4.79, SD: 1.41), the video modeling group (mean: 4.25,

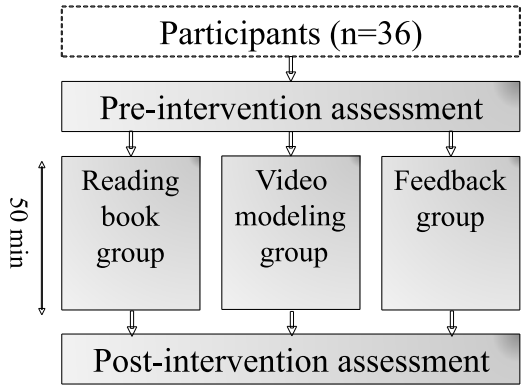

Fig. 11. Study design and participant assignment to experimental groups in the third experiment.

SD: 0.89), and the feedback group (mean: 3.75, SD: 0.87), which were not significantly different (F[2,33]=2.77, $p > .05$).

The effect of intervention type was analyzed using one-way factorial ANOVA. Post-hoc comparisons between the feedback group and the reading book group, and the feedback group and the video modeling group were performed using Bonferroni's method. We examined the case excluding the participants who initially scored high (6 to 7). We also investigated relationship between initial scores and improvement in scores, and gender differences using correlation coefficient and Student's t-test respectively.

After using the automated social skills trainer, the feedback group answered a questionnaire to evaluate usability and effectiveness of the system[8]. The questionnaire included the following items related to the system usability and training effect, rated on a scale of 1 (disagree) to 7 (agree). The users were also asked to provide comments about each question.

Q1 The system was easy to use.
Q2 I would like to use this system frequently.
Q3 The trainer looks like a human.
Q4 Watching my own video and feedback were useful.
Q5 Watching model video was useful.

8.2. Agreement

We calculated agreement according to the same procedure described in the previous section. The Kappa coefficient of two classes for two raters was 0.68, which indicates good agreement based on [Rietveld and Van Hout 1993]. The agreement of the two raters was almost the same as the second experiment.

8.3. Training effect

Figure 12 shows the improvement of overall narrative skills in each group. These results show that intervention type significantly affected the change in raters' scores (F[2,33]=6.41, $p < .05$) with $\eta_p^2 = .28$ according to ANOVA. Comparisons showed that the change in raters' scores of participants in the feedback group (mean: 0.79, SD: 0.58) who used the automated social skills trainer was significantly higher than the reading book group (mean: -0.04, SD: 0.54) and the video modeling group (mean: 0.17, SD:

[8]http://goo.gl/forms/Za4w3kfKUx

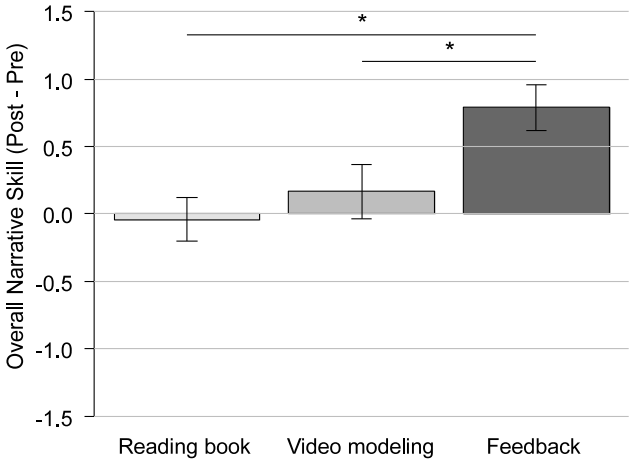

Fig. 12. The overall narrative score of each group. Error bars indicate standard error (*: $p < .05$).

0.65) ($p < .05$). The difference between the video modeling group and the reading book group was not judged as statistically significant ($p > .05$). We also excluded the participants who scored 6 to 7 in the pre score (5 participants in the reading book group), and analyzed the initial scores of the reading book group (mean: 3.79, SD: 0.86) and improvement in the scores (mean: 0.07, SD: 0.34). ANOVA and posthoc comparisons found same results as the case including all participants.

Figure 13 shows the improvement of overall narrative skills and initial scores in each group. The correlation coefficient between overall narrative skills prior to training and improvement was -0.53 ($p < .05$) showing a weak negative correlation. This is a natural result, because people who have difficulties in social interaction have more space to improve.

Figure 14 shows the improvement of overall narrative skills and initial scores in each gender. The initial scores and improvement in the scores between genders were not significantly different ($p > .05$).

8.4. Subjective evaluations

The paragraphs below describe findings from the participants' subjective evaluations of the automated social skills trainer and their feedback on their experience. We analyzed qualitative and quantitative results to represent user experience and system usability.

— **The system was easy to use:** The usability of the automated social skills trainer was rated an average of 5.4 (SD = 0.9). Most participants found the system is easy to use.

"It is easy to operate the system using only speech. My voice was recognized and I felt comfortable."

"The content of training was separated according to purpose (e.g. modeling, feedback, and homework), and it was easy for me."

— **I would like to use this system frequently:** The question regarding whether the user would like to use the system again was rated an average of 5.0 (SD = 0.7). Most participants would like to use the system frequently.

"I would like to talk to system with more variation. I want to use this system every day, and also record a life log."

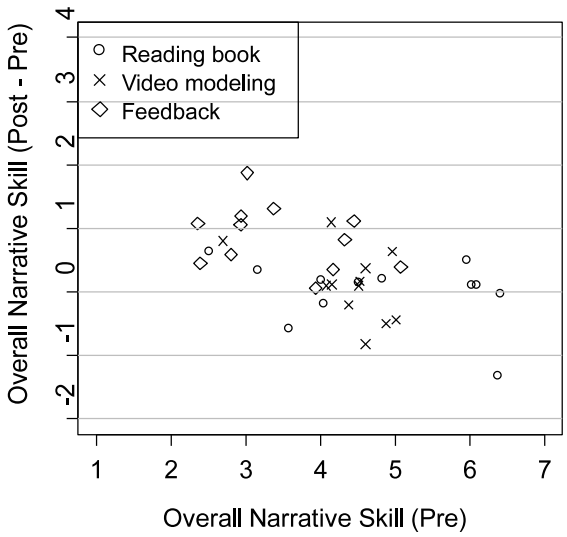

Fig. 13. The relationship between initial and improvement in scores for each group. Small amount of noise were added to each point in order to separate overlapping points.

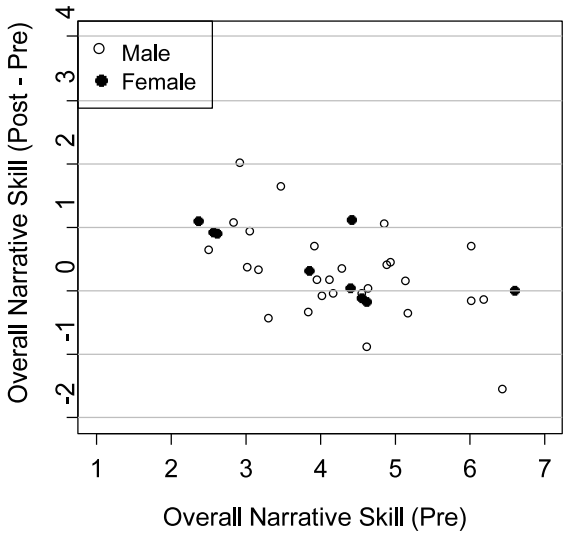

Fig. 14. Gender differences between initial and improvement in scores. Small amount of noise were added to each point in order to separate overlapping points.

"It is interesting to watch my score with helpful comments."

— **The avatar looks like a human:** The question about whether the avatar looks like a human was rated an average of 4.8 (SD = 0.4). Some participants thought the character looks like a human.

"I thought avatar's behaviour is natural, and I did not feel unnaturalness in interaction."

"I felt like I spoke to real human."

However, some participants thought the avatar did not seem like a human specifically in terms of speech synthesis.

"I felt the synthesized speech is robot-like, the intonation was unnatural."

— **Watching my own video and feedback were useful:** According to the participant's responses to the questionnaire, the feedback and watching the user's own video was rated an average of 5.6 (SD = 1.1). Most participants thought the feedback and video were useful.

"It was easy to train my skill because the system indicated the points to be improved."

"I was happy to be encouraged."

"Conversation is abstract, but the system displayed the concrete values. It is very interesting and helpful."

— **Watching model video was useful:** Overall, participants rated their preference toward watching the models' video an average of 5.2 (SD=1.5), suggesting the usefulness of model video.

"After watching the role model, I easily started to talk because of the good reference."

"I was interested in the variation of the good examples."

However, the result also showed the SD value was large. Some participants did not say that model video was helpful.

"I think I already had good skill so the modeling is not useful for me."

"I would like to see the good points of the model persons."

9. DISCUSSION

In this paper, we developed a dialogue system named "automated social skills trainer" which provides social skills training in the context of human-agent interaction. The automated social skills trainer is based on conventional SST including defining target skills, modeling, role-play, feedback, reinforcement, and homework. We focus on story telling/narrative skill as a target skill because this skill is widely used in many situations, and training can be relatively easily implemented using dialogue with an automated agent. The system includes several modules: behaviour generation, sensing & analysis from recorded video, and summary feedback. In this study, our focus was to assess the effectiveness of an automated social skills trainer that follows human-to-human SST as closely as possible. To evaluate effectiveness of the automated social skills trainer, we performed three experiments examining 1) the advantages of using computer-based training systems compared to human-human interaction by subjectively evaluating nervousness, ease of talking, and ability to talk well, 2) the relationship between speech language features and human social skills, and 3) the effect of computer-based training using our proposed system.

In our first experiment, we confirmed that nervousness and ability to talk well were related to type of interaction. Although there was no significant difference between interaction with the familiar person and interaction with the avatar with regards to these aspects, these two settings were significantly different in terms of nervousness and ability to talk well compared to interaction with the unfamiliar person. These results suggest that computer-based training may allow users to feel more comfortable than human-human interaction with an unfamiliar person, and human-computer interaction may be an alternative for people with social communication difficulties to become involved in SST for the first time. In addition, the results showed that using avatars in SST allowed the participants to use their conversational skills more effectively. However, how the skills learned while interaction with the avatar can generalized to the stressful situation of interacting with an unfamiliar person (such as interview with a professional career counselor [Hoque et al. 2013]) was not evaluated.

This is also a common problem of human-to-human SST performed by familiar trainer [Bellack 2004], so in the future we plan to quantify the generalization effect in further experiments.

In our second experiment, we confirmed that the relationship between overall narrative skill and speech and language information, confirmed that there was no significant difference between HHI and HCI with respect to overall narrative skill, set model persons according to the evaluation of two raters, and found a relationship between observed narrative ability and AQ. Baron-Cohen and their colleagues reported that the AQ value was widely distributed among members of the general population and that it is related to autistic traits [Baron-Cohen et al. 2001]. Our result showing a relationship between AQ and overall narrative skill is consistent with the above report. As a result, we found that WPM, words of more than 6 letters, and amplitude were significantly important to predict narrative skill. However, the number of participants were limited to select good examples and predict overall narrative skill. For the future, we can apply these analyses to the data acquired from different recordings such as the experiment 3.

In our third experiment, we confirmed a training effect particularly for participants in the case of the feedback group rather than the reading book group. It showed that the system could help people who have difficulties in social interaction improve their social skills. The video modeling group also improved in their scores, which is consistent to the previous work [Essau et al. 2014]. The video modeling of others was also helpful in social skills training. In this experiment, we did not set a group that watched their own video and did not watch the feedback. There is previous work reporting that subjects dislike looking at their own video during interview skill training and the skills did not change [Hoque et al. 2013], so we plan to investigate these elements separately in the future. We also confirmed a weak negative correlation between initial narrative skills and improvements in scores. This shows that training effects are found more strongly in people who have difficulties in social interaction than others. In subjective evaluation, we confirmed most participants of the feedback group were satisfied with the system in terms of usability and the feedback.

While this study was targeted for Japanese language, strictly language-dependent features of the system are minimal, and thus there are possibilities to adapt to other languages such as English. In particular, the system uses fixed utterances that are easily translated to other languages. However, because the features we extracted might be dependent on language, or more likely culture, examining related behavioural features in other languages is an interesting avenue for future work.

We summarize the limitations of the paper. First we did not consider agent's gender (we used only a female character). Taking into account the study that students perceived the male agents as significantly more interesting, intelligent, useful, and leading to greater satisfaction than the female agents [Baylor and Kim 2004] is an important avenue for future work. Second, our three experiments were performed with an imbalanced male to female ratio (with a larger number of the male participants). It has been noted that gender plays a role in the training effect of conversational coaching (e.g. [Hoque et al. 2013]), and might be related to social difficulties [Baron-Cohen et al. 2001], and thus investigating this effect in the context of automated social skills training is an important avenue for future work. Third, the current system did not consider the interactive aspects of a dialogue system. The system was targeted to teach narrative skill (a type of one-way story telling) and was used a simple strategy for nonverbal behaviour generation. In future work, we will combine other interactive models such as timing of nodding and blinking [Lee et al. 2010], and use not only a rule-based dialogue system but a more interactive conversation partner (for example, one that asks different questions) and generates different comments for feedback. We hope that this

modification will improve the user's desire to continue to use the system. Fourth, the system did not attempt to understand the content of user utterances. Although most SST focuses on non-verbal aspects of social interaction [Bellack 2004] and do not take the content of user utterances into consideration, a recent study showed the effectiveness of topic modeling in the context of job interview training [Naim et al. 2015]. We hope to consider the content of user utterances in future iterations of automated social skills training, although these will be dependent on the accuracy of speech recognition.

For other future directions, we would like to confirm the training effect over a longer period, and recruit special-need populations such as people with ASD. We also plan to add other target social skills in the automated social skills trainer, and compare with human-to-human SST. In order to do so, we will more thoroughly examine SST from the viewpoint of HHI and HCI including types of agent (anime-like agent vs. human-like agent). We have collected multi-modal information in the course of running experiments in this paper, such as eye-tracking, EDA, and interaction on the screen, and a detailed analysis of this data could be insightful. In addition, we plan to incorporate visual image processing and eye-gaze analysis into the automated social skills trainer.

ACKNOWLEDGMENTS

We would like to thank Kazuyo Iida for her helpful comments regarding the system.

REFERENCES

Yaser Adi, Amanda Kiloran, Kulsum Janmohamed, and Sarah Stewart-Brown. 2007. Systematic review of the effectiveness of interventions to promote mental wellbeing in primary schools Report 1: Universal approaches which do not focus on violence or bullying. *National Institute for Health and Clinical Excellence, London* (2007).

APA American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub.

Simon Baron-Cohen, Jennifer Richler, Dheraj Bisarya, Nhishanth Gurunathan, and Sally Wheelwright. 2003. The systemizing quotient: an investigation of adults with Asperger syndrome or high–functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358, 1430 (2003), 361–374.

Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. 2001. The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders* 31, 1 (2001), 5–17.

Nirit Bauminger. 2002. The facilitation of social-emotional understanding and social interaction in high-functioning children with autism: Intervention outcomes. *Journal of autism and developmental disorders* 32, 4 (2002), 283–298.

Amy L Baylor and Yanghee Kim. 2004. Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In *Intelligent tutoring systems*. Springer Berlin Heidelberg, 592–603.

Alan S Bellack. 2004. *Social skills training for schizophrenia: A step-by-step guide*. Guilford Press.

J Bishop. 2003. The Internet for educating individuals with social impairments. *Journal of Computer Assisted Learning* 19, 4 (2003), 546–556.

Daniel Bone, Matthew P Black, Chi-Chun Lee, Marian E Williams, Pat Levitt, Sungbok Lee, and Shrikanth Narayanan. 2012. Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist.. In *INTERSPEECH*.

Yoram S Bonneh, Yoram Levanon, Omrit Dean-Pardo, Lan Lossos, and Yael Adini. 2010. Abnormal speech spectrum and increased pitch variability in young autistic children. *Frontiers in human neuroscience* 4 (2010).

Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. 2013. Applied multiple regression/correlation analysis for the behavioral sciences. (2013).

Megan Davis, Kerstin Dautenhahn, CL Nehaniv, and SD Powell. 2004. Towards an Interactive System Facilitating Therapeutic Narrative Elicitation in Autism. (2004).

Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.

Nancy Eisenberg, William Damon, and Richard M Lerner. 2006. *Social, emotional, and personality development*. John Wiley & Sons.

Cecilia A Essau, Beatriz Olaya, Satoko Sasagawa, Jayshree Pithia, Diane Bray, and Thomas H Ollendick. 2014. Integrating video-feedback and cognitive preparation, social skills training and behavioural activation in a cognitive behavioural therapy in the treatment of childhood anxiety. *Journal of affective disorders* 167 (2014), 261–267.

MH Hanson. 1995. *Glottal characteristics of female speakers. Harvard University*. Ph.D. Dissertation. Ph. D. dissertation.

Peter A Heeman, Rebecca Lunsford, Ethan Selfridge, Lois Black, and Jan Van Santen. 2010. Autism and interactional aspects of dialogue. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 249–252.

Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 697–706.

Mohammed Ehsan Hoque and Rosalind W Picard. 2014. Rich nonverbal sensing technology for automated social skills training. *Computer* 47, 4 (2014), 28–35.

Leo Kanner and others. 1943. *Autistic disturbances of affective contact*. publisher not identified.

Géza Kiss and Jan PH van Santen. 2013. Estimating speaker-specific intonation patterns using the linear alignment model.. In *INTERSPEECH*. 354–358.

Géza Kiss, Jan PH van Santen, Emily Tucker Prud'hommeaux, and Lois M Black. 2012. Quantitative Analysis of Pitch in Speech of Children with Neurodevelopmental Disorders.. In *INTERSPEECH*.

Jina Lee, Zhiyang Wang, and Stacy Marsella. 2010. Evaluating models of speaker head nods for virtual agents. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 1257–1264.

RP Liberman and CJ Wallace. 1990. Social and independent living skills: Basic conversation skills module. *Camarillo, Calif: Author* (1990).

Robert Paul Lieberman. 1987. *Social and independent living skills*. UCLA Psychiatric Rehabilitation Consultants.

Joanne McCann and Sue Peppé. 2003. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders* 38, 4 (2003), 325–350.

David Moore, Paul McGrath, and John Thorpe. 2000. Computer-aided learning for people with autism–a framework for research and development. *Innovations in Education and Teaching International* 37, 3 (2000), 218–228.

Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 1. IEEE, 1–6.

Sarah Parsons and Peter Mitchell. 2002. The potential of virtual reality in social skills training for people with autistic spectrum disorders. *Journal of Intellectual Disability Research* 46, 5 (2002), 430–443.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: LIWC [Computer software]. *Austin, TX: liwc. net* (2007).

Toni Rietveld and Roeland Van Hout. 1993. *Statistical techniques for the study of language and language behaviour*. Walter de Gruyter.

Masoud Rouhizadeh, Emily Prud'hommeaux, Brian Roark, and Jan Van Santen. 2013. Distributional semantic models for the evaluation of disordered language. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, Vol. 2013. NIH Public Access, 709.

Jan PH Santen, Richard W Sproat, and Alison Presmanes Hill. 2013. Quantifying repetitive speech in autism spectrum disorders and language impairment. *Autism Research* 6, 5 (2013), 372–383.

Björn Schuller, Erik Marchi, Simon Baron-Cohen, Helen O'Reilly, Delia Pigat, Peter Robinson, and Ian Daves. 2014. The state of play of ASC-Inclusion: an integrated Internet-based environment for social inclusion of children with autism spectrum conditions. *arXiv preprint arXiv:1403.5912* (2014).

Miriam Silver and Peter Oakes. 2001. Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others. *Autism* 5, 3 (2001), 299–316.

Petr Slovák and Geraldine Fitzpatrick. 2015. Teaching and developing social and emotional skills with technology. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 4 (2015), 19.

Petr Slovák, Ran Gilad-Bachrach, and Geraldine Fitzpatrick. 2015. Designing Social and Emotional Skills Training: The Challenges and Opportunities for Technology Support. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2797–2800.

Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2014. Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. *ACL 2014* (2014), 88.

Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2015. Automated Social Skills Trainer. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 17–27.

Amy Vaughan Van Hecke, Jocelyn Lebow, Elgiz Bal, Damon Lamb, Emily Harden, Alexis Kramer, John Denver, Olga Bazhenova, and Stephen W Porges. 2009. Electroencephalogram and heart rate regulation to familiar and unfamiliar people in children with autism spectrum disorders. *Child development* 80, 4 (2009), 1118–1133.

Akio Wakabayashi, Simon Baron-Cohen, Sally Wheelwright, and Yoshikuni Tojo. 2006. The Autism-Spectrum Quotient (AQ) in Japan: a cross-cultural comparison. *Journal of autism and developmental disorders* 36, 2 (2006), 263–270.

Charles J Wallace, Connie J Nelson, Robert Paul Liberman, Robert A Aitchison, David Lukoff, John P Elder, and Chris Ferris. 1980. A review and critique of social skills training with schizophrenic patients. *Schizophrenia Bulletin* 6, 1 (1980), 42.

Received July 2015; revised *** 2015; accepted *** 2016