Noname manuscript No.
(will be inserted by the editor)

Analysis of Conversational Listening Skills toward Agent-based Social Skills Training

Hiroki Tanaka · Hidemi Iwasaka · Hideki Negoro · Satoshi Nakamura

Received: date / Accepted: date

Abstract Listening skills are critical for human communication. Social skills training (SST), performed by human trainers, is a well-established method for obtaining appropriate skills in social interaction. Previous work automated the process of social skills training by developing a dialogue system that teaches speaking skills through interaction with a computer agent. Even though previous work that simulated social skills training considered speaking skills, the SST framework incorporates other skills, such as listening, asking questions, and expressing discomfort. In this paper, we extend our automated social skills training by considering user listening skills during conversations with computer agents. We prepared two scenarios: Listening 1 and Listening 2, which respectively assume small talk and job training. A female agent spoke to the participants about a recent story and how to make a telephone call, and the participants listened. We recorded the data of 27 Japanese graduate students who interacted with the agent. Two expert external raters assessed the participants' listening skills. We manually extracted features that might be related to the eye fixation and behavioral cues of the participants and confirmed that a simple linear regression with selected features correctly predicted listening skills with a correlation coefficient above 0.50 in both scenarios. The number of noddings and backchannels within the utterances contributes to the predictions because we found that just using these two features predicted listening skills with a correlation coefficient above 0.43. Since these two features are easier to understand for users, we plan to integrate them into the framework of automated social skills training.

H. Tanaka
Nara Institute of Science and Technology, Japan
E-mail: hiroki-tan@is.naist.jp

H. Iwasaka
Nara Medical University / Heartland Shigisan, Japan

H. Negoro
Nara University of Education, Japan

S. Nakamura
Nara Institute of Science and Technology, Japan

1 Introduction

Social skills are critical factors that influence human life. Persistent social skill deficits hamper those with such afflictions from forming relationships or succeeding in social situations. Social skills training (SST) [6], a general psychosocial treatment through which people with social difficulties can obtain appropriate social skills, is widely used by teachers, therapists, and trainers. Automating the SST process will simplify the acquisition of such social skills by those who require them.

Autism spectrum disorder (ASD) is a spectrum condition [1,12], meaning it has a broad range of characteristics from mild to severe. Using computer agents in SST is motivated by the fact that even though some people with high-functioning autism experience difficulty during social communication, they also show good or even superior systemizing skills [3,13]. Systemizing is the drive to analyze or build systems and understand and predict behavior in terms of underlying rules and regularities. The use of systematic computer-based training for people who need to improve their social skills exploits the following criteria: 1) such people favor computerized environments because they are predictable, consistent, and free from social demands; 2) they can work at their own pace and level of understanding; 3) training can be repeated until the goal is achieved; and 4) interest and motivation can be maintained through computerized rewards. It may also be easier for those who suffer from social difficulties to use computer agents than to directly interact with humans [44,10]. A recent paper suggested that people with such social difficulties as ASD feel safer and more comfortable in a virtual environment than during interaction with an actual person [35].

Previous works conducted SST using computer agents [7,30], including in speaking and emotional regulation contexts [45,42,18,52]. A special issue of the Journal on Multimodal User Interfaces on SST addressed the Theory of Mind (ToM) reasoning [50], gaze leading [36], virtual discrete trial training for teacher trainees [40], and bad news conversations [33].

Specifically, Tanaka et al., [45] proposed an automated social skills trainer that consists of the following aspects: 1) instruction and motivation of target skills, 2) modeling, 3) role-playing, 4) feedback, and 5) homework. In terms of interactive application design, these processes mostly include two challenging parts: assessment of the target skills and giving feedback about those target skills. The system follows the evidence-based training scheme called the Bellack method [5].

The Bellack method (or step-by-step SST) is a well-structured and widely used evidence-based approach. It is a cognitive-behavioral SST inspired by Social Learning Theory's five core principles: modeling, shaping, reinforcement, overlearning, and generalization [2,39]. The Bellack method defines the SST framework and its four basic skills: speaking, listening, asking questions, and expressing discomfort. These skills are beneficial for all people [5]. Training should be done in the above order; since expressing discomfort is more difficult, it should be the last. These skills are helpful and basis to obtain other skills such as how to be more assertive, and how to maintain conversation. In this paper, we focus on listening skills as second most important skills. The following are the critical aspects of listening skills: 1) looking conversation partners in the eye (eye contact), 2) nodding, and 3) repeating the keywords of the conversation partner [5]. The listening skills investigated by this study are not for hearing and understanding speech [4,8]. Listening skills explicitly express that one is actively listening to the partner's speech [5,48]. For example, a previous study showed that people with social difficulties tend to avoid looking conversation partners in the eye [20].

Despite the importance of listening skills, most automated SSTs focus on speaking skills. Okada et al. assessed interaction skills that addressed listening attitudes [34]. Many differently motivated works have been designed to generate (model) human-like head tilting, backchannels, and nodding on humanoid robots or computer agents based on analyzing human behaviors [25, 31,16,17,22,29,19]. Ward et al. proposed listening-skills training that produces immediate feedback, although they did not use computer agents and focused only on backchanneling behavior [51]. Another previous work argued that personality [14] is related to empathic listening skills [38].

In this paper, we hypothesized that the SST process in listening skills can be automated for interaction between humans and computer agents. First, we analyzed part of the automatic prediction of listening skills by collecting the listening data of the interaction of graduate students and computer agents and investigated the possibility of automatically assessing user's listening skills from multimodal aspects. To the best of our knowledge, this is the first work to analyze human conversational listening skills in human-agent interaction.

This paper is an extended version of previously published work [43]. We extend that previous work by scrutinizing correlation and making predictions with a few features toward the SSTs of actual users.

2 Computer Agents

We used an MMD agent [24] as a computer agent and show an example in Fig. 1. We used default parameters for its speech such as speaking rate and voice pitch. For example, the articulation rate, which divides the number of morphemes (measured in MeCab [21]) by voiced seconds, is 1.85 seconds for the following utterance of the agent: "hazime ni denwa o kake te." The pitch was around 270 Hz.

Four Japanese people (two males and two females) created the agent's spoken sentences: one person was a licensed psychiatrist with more than three years of experience with SST and another was a licensed speech therapist. We created the following three tasks: Speaking, Listening 1, and Listening 2:

1. Speaking: The user describes a recent story/experience to the computer agent. This module follows the same procedure as a previous work [45].
2. Listening 1: The user listens to the agent's story. Table 1 shows the sentences that we created (translated into English). This assumes a casual, small-talk situation.
3. Listening 2: The user listens to a procedure that explains how to make a telephone call. These sentences are shown in Table 1 (translated into English). They are designed for more serious situations such as job training.

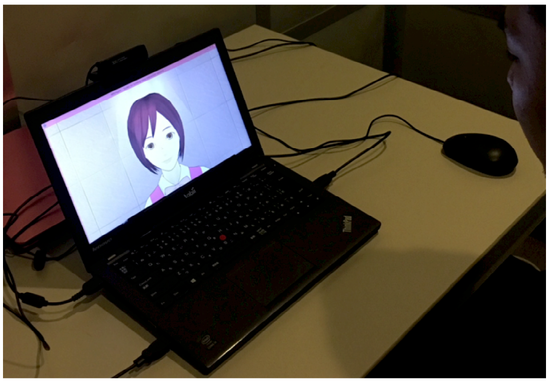

Fig. 1 Interaction with computer agents.

We provided them in the same order for all the participants because we assumed the order of actual training. We aimed to develop from the basic skills from speaking to listening and then to practical skills in actual job-training situations. Since we are examining the relationship between autistic traits and speaking/listening skills, we placed the speaking task before the listening tasks.

In Listening 1 and Listening 2, the agent spoke for about one minute, including several four- or five-second pauses between the sentences. These pause lengths were empirically determined by two people (one of whom was a licensed psychiatrist with over three years of experience with SST) for the naturalness of the conversation and knowledge that even elderly persons can respond to the agent's question in an average of less than four seconds [41]. During the pauses, the agent nodded if the user said something and waited over three seconds after the user's final utterance. The collected data had few overlap phenomena because the system explicitly waited for three seconds after speaking sentences and waited for three more seconds after the user utterance. Since backchannel and turn-taking generally impact overlapped speech, we must consider such phenomena for our assessment and training in the future. At the end, if the user answers yes to the questions, "Is there anything you would like to ask?" and "Do you have any more questions?", the agent gives no answer, waits for the end of the user's utterance, and finally answers "Thank you."

3 Methods

3.1 Participants

We recruited 27 participants (six females and 21 males whose mean age was 25.1, SD: 2.13) from the Nara Institute of Science and Technology. We payed 1000 yen for each participant. The first author explained the experiment to them and obtained their informed consent. They completed consecutive Speaking (60 sec.), Listening 1 (60-90 sec.), and Listening 2 (60-90 sec.) sessions. The completion time varied based on how long the participants spoke (about four minutes of completion time).

3.2 Procedure

The first author explained how to use the system by playing an example video that showed head nodding and backchannel feedback. This was a video of an experimenter testing the system.

Data were collected in a soundproof room using a laptop PC (IBM ThinkPad). A webcam (ELECOM UCAM-DLY300TA) was placed on top of the laptop, and an eye-tracker (Tobii X2-30) was set at the bottom of the laptop screen. We turned off the light in the room to minimize external distractions (Fig. 1).

After collecting data, we gave two questionnaires (explained below) to all of the participants. The total amount of time for all the procedures was approximately 20 minutes. From the collected data, we calculated the following multimodal aspects: eye fixation, image, and speech features. We selected these features based on previous studies [17,31,22], specifically influenced by the critical aspects of listening skills from the Bellack method's SST [5].

3.3 Eye Fixation

We used a table-mounted eye tracker (Tobii X2-30) because it can obtain focal points with high resolution accuracy when head movements are relatively small [11].

We applied an IV-filter to the raw eye-gaze data and manually categorized the following areas of interests (AOIs): 1) eyes, 2) mouth, 3) face, 4) other. Face includes eyes and mouth regions. Finally, we extracted the ratio of each AOI. These regions are represented in Fig. 2.

3.4 Speech- and Image-based Features

We manually coded head nodding (video) and speech (audio) using the ELAN tool. The following information was coded by one male annotator: backchannel feedback (defined as the following Japanese vocalizations: un, un un, un un un, hu un, hai, hai hai, and hai hai hai), backchannel feedback between an agent's utterances,

Table 1 Sentences spoken by computer agents: pause denotes three seconds of silence, and long pause denotes five seconds of silence.

*Listening*1
The other day, my friends and I went to a trendy cafe in Kyoto that was advertised in a magazine. *pause*
I drank a caffe latte. The foam on top of it had these really cute animal designs.
So I took a picture and uploaded it to Instagram. *pause*
After that, my friend and I drove to Kiyomizu Temple.
I found a souvenir shop on a side street on the way.
After carefully considering all the choices, I bought a very delicious roll cake. *pause*
At Kiyomizu Temple, the autumn leaves were so beautiful.
Since I also think that cherry blossoms are beautiful, I'm hoping to return in the spring.
That's all. Is there anything you would like to ask? *long pause*
Thank you.

*Listening*2
First, dial the number, and after someone answers, give your name and affiliation. *pause*
After that, say the name and affiliation of the person to whom you want to speak
and ask to be connected. *pause*
When you are connected, briefly state your purpose. *pause*
If the person in charge is unavailable, explain that you will call back, and then hang up. *pause*
It is important to make a phone call at a proper time.
Since calling late at night or early in the morning will probably annoy people, avoid doing so. *pause*
This concludes the explanation. Do you have any more questions? *long pause*
Thank you.

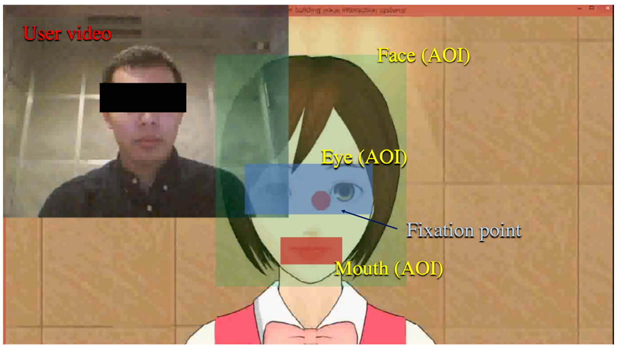

Fig. 2 Exported video, AOI (rectangles), and fixation point (red circle) examples: these are inexact areas.

backchannel feedback within an agent's utterances, repetition of an agent's utterance (paraphrase), questions, miscellaneous utterances, head nod (once), head nod (twice), and head nod (three or more times). We manually coded one nod differently from two or more than three nods. In this study, we simplified how we counted the number of nods as the total amount of head nods as once, twice, and three or more times based on a previous work [31]. We defined more than one second as separate coding. This means an interval between two consecutive nods. Here the backchannel feedback is a verbal behavior that expresses something, e.g., I'm adjusting to my conversation partners. Nodding, a non-verbal behavior, is represented by vertical head movements that express acknowledgement and agreement, e.g.

As coding output, we extracted the following features: 1) the number of backchannel feedback instances (also the backchannels within and between utterances), 2) the number of repetitions/paraphrase utterances, 3) the number of questions, 4) the number of miscellaneous utterances, 5) the number of nods, 6) gap (the timing between the end of an agent's utterance and the beginning of a user's utterance).

3.5 Social Responsiveness Scale and Big Five Personality Test

We collected answers to two questionnaires: the Social Responsiveness Scale (SRS) [9], which is related to autistic traits based on the DSM-V [1], and the Big Five Personality Test [15]. We collected answers from these two questionnaires because we hypothesized that autistic traits and personality [14] are related to each other as well as to speaking and listening skills. The latter consists of the following sub-areas: extraversion, agreeableness, conscientiousness, neuroticism, and openness. The relationship between these two questionnaires was previously investigated [47]. Personality is related to empathic listening skills [38]. We found a significantly negative correlation between SRS and extraversion (Spearman's $\rho = -0.5$, $p < 0.05$) (Fig. 3).

3.6 Clinical Psychologists' Ratings of Listening Skills

Two licensed clinical psychologists with over three years of experience with SST rated both the listening and speaking skills by watching videos exported from the Tobii video recorder (Fig. 4). Their impressions focused on the participants in addition to such behaviors as

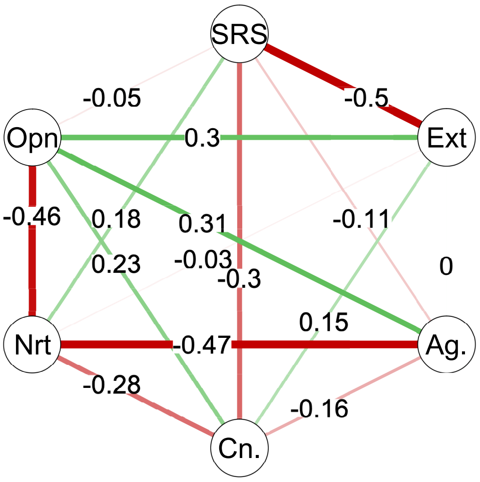

Fig. 3 Relationships among each social psychological scale.

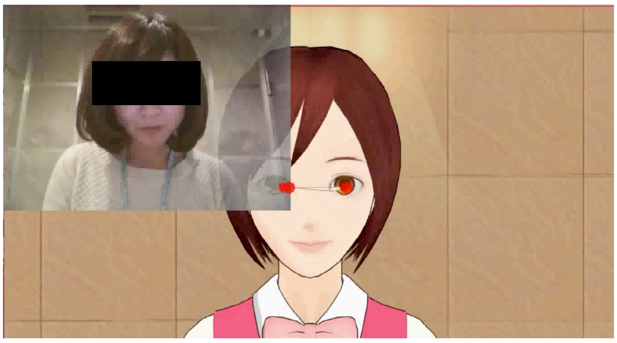
[redacted]

Fig. 4 Screenshot of video recording for ratings.

eye movements, head noddings, facial expressions, and speech (as with the usual SST). After watching multiple videos, they evaluated the overall listening skills of the participants with Likert scores on a scale of 1 (not good) to 7 (good) [45].

We calculated the Kappa statistics of the two raters using a weighted Kappa set to 1 (on the diagonal) and decreased weights off the diagonal [23]. The following are the weighted Kappa correlation coefficients: 0.37 (Speaking), 0.47 (Listening 1), and 0.59 (Listening 2). They show fair to moderate agreement [23]. We also calculated the Pearson's correlation coefficients of the two raters: 0.44 (Speaking), 0.46 (Listening 1), and 0.66 (Listening 2) (all, p<0.05). Finally, we averaged the two raters' scores for further analysis. Although we tried to separately analyze their individual scores, averaging their results led to more consistent results. For example, as reported below in section 4.2, we confirmed larger variation and fewer mean values of correlation coefficient than the averaged results beforehand. Demographic information of all participants is shown in Table 2.

3.7 Statistical Analysis

This section presents our experimental evaluation of the collected data. After analyzing the relationship among each feature, we evaluated our prediction model for automatic listening-skill prediction. In this study, we used Spearman's correlation coefficient to observe correlation. We defined good models as people who scored above five in both raters [45] and will be used them as good examples in SST.

First, we analyzed the relationship between each question and the two listening skills. Then we normalized the extracted features using z-score normalization. Regarding automatic assessment, we used multiple linear regression (LR), which is a very simple linear approach to predict listening skills. Leave-one-person-out cross validation evaluated the generalizability. We automatically selected features based on Akaike's information criterion (AIC) in a stepwise algorithm (forward) on the training set.

Finally, after confirming normality by the Kolmogorov-Smirnov test, we calculated the Pearson's correlation coefficient between the actual and predicted values as well as the root mean square error (RMSE). We also evaluated the statistical significance using the Wilcox rank sum test and the Cohen's D value between good models and the others in terms of SRS scores and multimodal behaviors.

4 Results

4.1 Statistical Analysis

The following are the correlation coefficients: Listening 1 and Speaking were 0.31 (p=0.10), Listening 2 and Speaking were 0.41 (p=0.03), and Listening 1 and Listening 2 were 0.54 (p=0.003). Eight people were selected as good models for Speaking. Seven were selected as good models for Listening 1. Five were chosen for Listening 2. Specifically, participants ID5 and ID12 were good for all three tests (speaking task and two listening tasks). Fig. 5 represents the SRS score distribution between the good models and the others. Good models have lower SRS scores than the others. Specifically, we found a significant difference for Listening 2.

Regarding eye fixation, we found mean values of 40% for the eyes, 7% for the mouth, 88% for the face, and 5% for the others within all of the fixation points in

Table 2 Participant demographics: ID, gender, age, SRS scores (min: 0, max: 195), extraversion, agreeableness, conscientiousness, neuroticism, and openness (min: 2, max: 14), averaged scores of speaking skills, listening skills 1, and listening skills 2 (min: 1, max: 7). Bold fonts indicate good models.

ID	Gender	Age	SRS	Extra.	Agree.	Consc.	Neuro.	Open.	Speaking	Listening 1	Listening 2
1	m	23	80	5	9	7	12	7	**6**	5.5	4
2	m	23	34	4	13	4	5	8	**5**	**6**	3.5
3	f	24	78	5	9	2	12	5	3.5	**5.5**	3
4	m	23	61	6	8	7	14	7	3.5	2.5	2.5
5	f	29	70	6	7	5	12	8	**7**	**7**	**6**
6	m	25	53	11	8	11	3	13	5	3	1.5
7	m	24	105	6	11	7	14	9	3	4	4
8	m	28	101	6	11	3	10	8	**6**	4	3
9	f	24	37	11	9	8	12	5	3	1	1
10	m	29	63	13	11	5	8	9	4	4	4.5
11	m	27	87	4	12	5	4	8	3.5	2	3
12	m	23	54	11	12	5	6	8	**6.5**	**6.5**	**6**
13	m	26	59	4	12	4	8	8	5	4	3.5
14	m	26	77	4	12	3	5	6	4.5	2	2
15	m	23	75	3	14	5	7	9	4	**5**	**5.5**
16	m	29	59	14	13	2	8	14	**6.5**	2.5	5
17	m	24	57	5	11	4	12	5	**6.5**	3.5	**5.5**
18	m	23	67	5	12	3	9	13	4.5	3	3.5
19	m	24	73	3	9	6	6	8	4.5	2	5
20	m	26	39	6	12	9	4	5	4	4	4
21	m	28	19	14	14	10	2	14	4.5	**5.5**	**5.5**
22	m	25	80	2	11	10	2	12	3	4.5	4
23	m	28	52	7	8	7	13	8	3.5	5	4
24	m	24	75	9	9	6	5	8	5	4	4
25	f	24	34	11	10	6	13	5	**5**	**5.5**	4.5
26	f	23	97	5	12	7	9	8	3	2.5	2
27	f	24	52	6	12	10	8	10	5	3.5	2

Table 3 Top five features: Brackets denote correlation coefficient (**: $p < 0.01$, *: $p < 0.05$, †: $p < 0.10$).

Rank	Listening 1	Listening 2
1	Nods (0.51**)	Back. (0.55**)
2	Questions (0.42*)	Back. w/ (0.48*)
3	Back. w/ (0.36†)	Nods (0.42*)
4	Repetitions (0.25)	Back. b/ (0.25)
5	Miscellaneous (0.22)	Repetitions (0.23)

every participant. We did not find any significant correlations in terms of eye fixation. The averaged # of nods was 4.89 in Listening 1 and 4.85 in Listening 2. The averaged # of backchannels was 5.26 in Listening 1 and 5.44 in Listening 2. We did not find any effects of elapsed time; nods and backchannels were not more frequent at the beginning or at the end of the video.

Table 4 indicates the top five features that are correlated to listening skills. The # of nods was significantly related to listening skills as was the # of backchannels. The backchannels between an agent's utterances are more important than those within an agent's utterances. Fig. 6 shows the examples of the actual timings of the head nodding and backchannel feedback in Listening 1 and Listening 2. We confirmed that persons with low listening scores tended to nod only during the agent's pauses. In contrast, those with high listening scores nodded and uttered at other times. They tended to respond at the positions of specific keywords, commas, and periods of the agent's transcripts within the sentence based on the pitch of the agent's speech. It was also discussed: backchannel occurs everywhere in Japanese, e.g., in near sentence-final particles and interjection particles [28]. Further study needs to investigate these effects.

4.2 Predicting Listening Skills

4.2.1 Scoring based on regression models

For linear regression, we identified the following correlation coefficients between the predicted and actual values: Listening 1 was 0.50 ($p < 0.01$), and Listening 2 was 0.51 ($p < 0.01$). Their RMSEs were 1.52 (Listening 1) and 1.26 (Listening 2).

Decreasing the number of features is crucial to reduce user's cognitive complexity when an automated system provides feedback in an actual SST. To do so, we extracted two important features found by previous results. We chose the number of noddings and backchannels within agent utterances to predict the listening

Table 4 Prediction results of listening skills (**: $p < 0.01$, *: $p < 0.05$).

Model	Correlation	RMSE
LR with AIC (Lis 1)	0.504**	1.52
LR with AIC (Lis 2)	0.511**	1.26
LR with two features (Lis 1)	0.481**	1.34
LR with two features (Lis 2)	0.438*	1.23

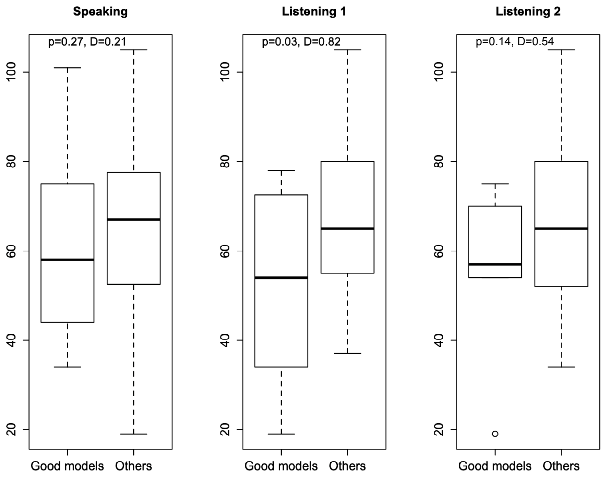

Fig. 5 Good models and SRS scores: We represent p-values of Wilcox rank sum test and Cohen's D values.

skills with these two features and obtained a 0.48 correlation coefficient in Listening 1 ($p < 0.01$) and a 0.43 correlation coefficient in Listening 2 ($p < 0.05$). Even though this slightly lowered the prediction values, they remain significant.

4.2.2 Good models and others

This subsection represents the differences between good models and the others in terms of the number of nodding and backchannels within an agent's utterances. We also confirmed a relationship between good models and SRS scores in Figure 5.

As shown in Table 5, we found a significant difference of nodding (Wilcox rank sum test (one-tailed), p=0.02). Good models nod with a mean of 7.0 in Listening 1. In contrast, the others had a mean of 5.09. The number of backchannel feedbacks within the utterances was not significantly different between the two groups in Listening 1. We did find a significant difference of backchannel feedback within an agent's utterances (Wilcox rank sum test (one-tailed), p=0.007). Good models gave backchannels with a mean of 5.0 in Listening 1; the mean of the others was 1.95. The number of nodding was not significantly different between the two groups in Listening 2.

Table 5 Mean and SD values of good models and others: Wilcoxon rank sum test with p-values.

	Model	Others	W	p-value
Listening 1				
back /w	3.6 (2.5)	0.68 (1.2)	91.5	0.13
nod	7.0 (4.0)	5.09 (3.8)	106.5	0.02
Listening 2				
back /w	5.0 (4.8)	1.85 (1.9)	91	0.007
nod	9.85 (6.8)	3.65 (3.1)	72.5	0.14

In Listening 1, the total amount of time that agents spoke was 45 seconds (total interaction time was around 60 seconds). In Listening 2, the total amount of time that agents spoke was 38 seconds (the total interaction time was around 56 seconds). Thus, good models nodded once every 8-9 seconds in Listening 1, and in Listening 2 they gave backchannel feedback when the agent spoke every 7-8 seconds. This seems relevant to human-human interaction in Japanese and demonstrates that backchannels occur on average every 7-8 seconds [28].

5 Discussion

This paper analyzed listening skills from conversational behaviors by interaction with computer agents. We prepared two scenarios: Listening 1 and Listening 2, which respectively are related to small talk and job training. We collected data from three types of settings in human-computer interaction. Several multimodal behavioral features were coded and extracted based on previous work [17,31,22,5], and a linear regression model achieved predictions of 0.50 in Listening 1 and 0.51 in Listening 2 of correlation coefficients. We confirmed that the correlation coefficients of the two raters were 0.46 (Listening 1) and 0.66 (Listening 2), and our prediction model achieved similar or slightly better predictions in Listening 1 and a previous work on speaking skills [45]. For Listening 2, the human raters agreed more than our prediction model. However, our study did not consider such behavioral effects as loud/clear speech, smiling, or posture. We need to investigate and extract such additional information to improve our model [22]. We also found that the amount of backchannel feedback was more important in Listening 2 than in Listening 1, probably because the former denotes a more serious type of interaction that requires explicit cues to indicate that one is listening. For fully automated listening skills analysis, we need to evaluate each behavior and our prediction model using backchannel recognition and nod detection [26].

This study investigated evidence-based findings that suggest that the following aspects are critical for listening skills: 1) looking conversation partners in the eye,

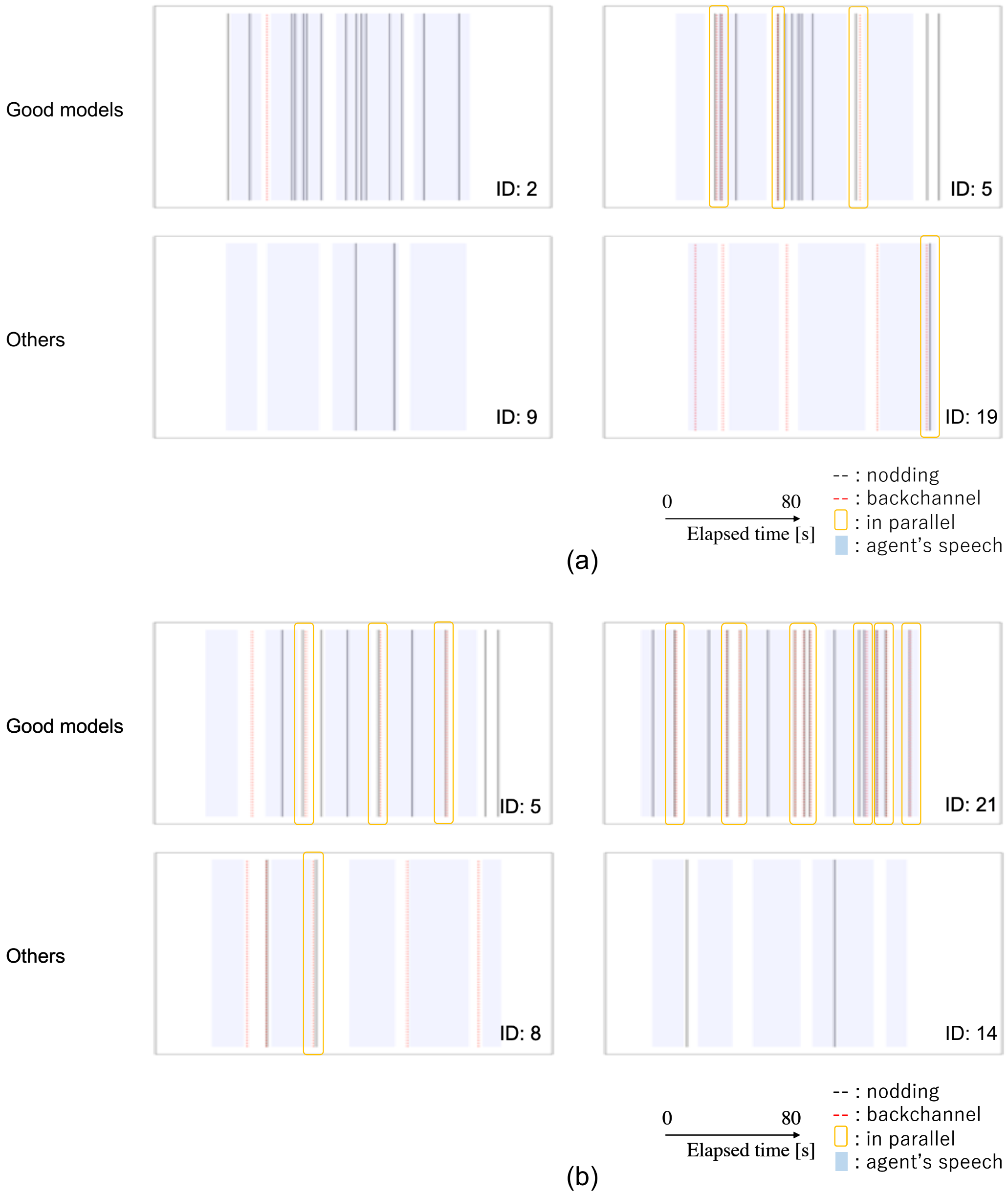

Fig. 6 (a) Timing is displayed in Listening 1; (b) Timing is displayed in Listening 2. User nodding and backchannels, and agent's speech: Graphs are lined horizontally. Rectangles show that nodding and backchannel occur in parallel.

2) nodding, and 3) repeating the keywords of conversation partners [5]. Our results confirmed that the second and third points are important. Although the number of repetitions is not significantly different in our study, we identified greater values of repetition in people who can be described as good models. However, the first point was not found in this study. This difference might reflect interaction types. We did not find any significant correlations in terms of eye fixation; the mean values for every participant were 40% for the eyes, 7% for the mouth, and 88% for the face. Most participants made eye contact in the context of human-agent interaction that was unrelated to listening skills.

This study did not control the training's order effects. Since the participants were all trained by speaking and listening 1 before listening 2, listening 2 might have been easier for them in the sense that they were already experienced. During the speaking task, the agent expressed backchannels. This is important since agent behavior might induce delayed imitation by the participants in the listening tasks. We will consider such order effects in future work.

Also, this study did not investigate the effects of human-agent interaction or human-human interaction as well as cultural aspect. A previous work suggested that people treat computers as real people and exhibit politeness to them [37]. In contrast, another recent work found that the dynamics of facial expressions differ for users interacting with a human and a virtual agent [32]. After the recording, one participant (ID 11) commented that human-agent interaction was safer than human-human interaction because people are complex and the feelings of conversational partners should be estimated in real time. As explained in the introduction, people with social difficulties favor computerized environments because they are predictable and consistent (not complex). The major difference between Japanese and American speakers was in the frequency and the discourse contexts in which backchannels occurred [27]. A previous study also showed that American speakers provided backchannel every 19-20 seconds. We need to consider such cultural aspects in the future.

The present study was conducted with neurotypical participants (without ASD). Thus training with multiple modalities was simplified. However, training participants with ASD with multisensory information is more complicated because they might get overwhelmed by the quantity and the multisensory nature of the stimuli/motor skills [42]. One possibility is to train each modality individually to enable ASD users to iteratively build upon the learned modalities one after the other.

We will integrate our listening-skills assessment into an automation feedback framework [26,45] and test it on people with ASD to measure the effects of the training on physiological attributes [49,46].

Acknowledgements This is a post-peer-review, pre-copyedit version of an article published in the Journal on Multimodal User Interfaces. The final authenticated version is available online at: https://doi.org/10.1007/s12193-019-00313-y

Funding was provided by Core Research for Evolutional Science and Technology (Grant No. JPMJCR19A5) and Japan Society for the Promotion of Science (Grant Nos. JP17H06101 and JP18K11437).

References

1. Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders: Dsm-5. Diagnostic and Statistical Manual of Mental Disorders. Amer Psychiatric Pub Incorporated (2013). URL https://books.google.co.jp/books?id=EIbMlwEACAAJ
2. Bandura, A.: Social learning theory of aggression. Journal of communication **28**(3), 12–29 (1978)
3. Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., Wheelwright, S.: The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. Philos. Trans. R. Soc. Lond., B, Biol. Sci. **358**(1430), 361–374 (2003)
4. Barry, J.G., Tomlin, D., Moore, D.R., Dillon, H.: Use of Questionnaire-Based Measures in the Assessment of Listening Difficulties in School-Aged Children. Ear Hear **36**(6), 300–313 (2015)
5. Bellack, A., Mueser, K., Gingerich, S., Agresta, J.: Social Skills Training for Schizophrenia, Second Edition: A Step-by-Step Guide. Guilford Publications (2013). URL https://books.google.co.jp/books?id=TSMxAAAAQBAJ
6. Bohlander, A.J., Orlich, F., Varley, C.K.: Social skills training for children with autism. Pediatric Clinics of North America **59**(1), 165 – 174 (2012). DOI https://doi.org/10.1016/j.pcl.2011.10.001. Autism Spectrum Disorders: Practical Overview for Pediatricians
7. Cassell, J.: Embodied conversational agents: Representation and intelligence in user interfaces. AI Mag. **22**(4), 67–83 (2001). URL http://dl.acm.org/citation.cfm?id=567363.567368
8. Cigerci, F., Gultekin, M.: Use of digital stories to develop listening comprehension skills **27**, 252–268 (2017)
9. Constantino, J.N., Davis, S.A., Todd, R.D., Schindler, M.K., Gross, M.M., Brophy, S.L., Metzger, L.M., Shoushtari, C.S., Splinter, R., Reich, W.: Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. Journal of autism and developmental disorders **33**(4), 427–433 (2003)
10. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., Morency, L.P.: Simsensei kiosk: A virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems, AAMAS '14, pp. 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2014). URL http://dl.acm.org/citation.cfm?id=2617388.2617415
11. Duchowski, A.T.: Eye Tracking Methodology: Theory and Practice. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2007)
12. Frith, U., Happe, F.: Autism spectrum disorder. Curr. Biol. **15**(19), R786–790 (2005)
13. Golan, O., Baron-Cohen, S.: Systemizing empathy: teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia. Dev. Psychopathol. **18**(2), 591–617 (2006)
14. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in Personality **37**, 504–528 (2003)
15. Gosling, S.D., Rentfrow, P.J., Swann Jr, W.B.: A very brief measure of the big-five personality domains. Journal of Research in personality **37**(6), 504–528 (2003)
16. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating Rapport with Virtual Agents. In: Lecture Notes in Artificial Intelligence; Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA), vol. 4722, pp. 125–128. Paris, France (2007)
17. Heylen, D.: Listening heads. In: Proceedings of the Embodied Communication in Humans and Machines, 2Nd ZiF Research Group International Conference on Modeling Communication with Robots and Virtual Humans, ZiF'06, pp. 241–259. Springer-Verlag, Berlin, Heidelberg (2008). URL http://dl.acm.org/citation.cfm?id=1794517.1794530

18. Hoque, M.E., Courgeon, M., Martin, J.C., Mutlu, B., Picard, R.W.: Mach: My automated conversation coach. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, pp. 697–706. ACM, New York, NY, USA (2013). DOI 10.1145/2493432.2493502. URL http://doi.acm.org/10.1145/2493432.2493502
19. Huang, L., Morency, L.P., Gratch, J.: Learning backchannel prediction model from parasocial consensus sampling: A subjective evaluation. In: J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, A. Safonova (eds.) Intelligent Virtual Agents, pp. 159–172. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
20. Klin, A., Jones, W., Schultz, R., Volkmar, F., Cohen, D.: Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. Arch. Gen. Psychiatry **59**(9), 809–816 (2002)
21. KUDO, T.: Mecab : Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/ (2005). URL https://ci.nii.ac.jp/naid/10019716933/
22. Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., Kawahara, T.: Attentive listening system with backchanneling, response generation and flexible turn-taking. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pp. 127–136. Association for Computational Linguistics (2017). URL http://aclweb.org/anthology/W17-5516
23. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1) (1977)
24. Lee, A., Oura, K., Tokuda, K.: Mmdagent - a fully open-source toolkit for voice interaction systems. In: ICASSP (2013)
25. Liu, C., Ishi, C.T., Ishiguro, H., Hagita, N.: Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In: ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 285–292 (2012)
26. Liu, F., Surendran, D., Xu, Y.: Classification of statement and question intonations in mandarin. Proc. 3rd Speech Prosody pp. 603–606 (2006)
27. Maynard, S.K.: Conversation management in contrast: Listener response in japanese and american english. Journal of Pragmatics **14**(3), 397–412 (1990)
28. Maynard, S.K.: Kaiwa bunseki (discourse analysis) [written in japanese] (1993)
29. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. **3**(1), 5–17 (2012). DOI 10.1109/T-AFFC.2011.20. URL http://dx.doi.org/10.1109/T-AFFC.2011.20
30. Milne, M., Raghavendra, P., Leibbrandt, R., Powers, D.M.W.: Personalisation and automation in a virtual conversation skills tutor for children with autism. Journal on Multimodal User Interfaces **12**, 257–269 (2018)
31. Nori, F., Lipi, A.A., Nakano, Y.: Cultural difference in nonverbal behaviors in negotiation conversations: Towards a model for culture-adapted conversational agents. In: Proceedings of the 6th International Conference on Universal Access in Human-computer Interaction: Design for All and eInclusion - Volume Part I, UAHCI'11, pp. 410–419. Springer-Verlag, Berlin, Heidelberg (2011). URL http://dl.acm.org/citation.cfm?id=2022591.2022639
32. Ochs, M., Libermann, N., Boidin, A., Chaminade, T.: Do you speak to a human or a virtual agent? automatic analysis of user's social cues during mediated communication. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, pp. 197–205. ACM, New York, NY, USA (2017). DOI 10.1145/3136755.3136807. URL http://doi.acm.org/10.1145/3136755.3136807
33. Ochs, M., Mestre, D., de Montcheuil, G., Pergandi, J.M., Saubesty, J., Lombardo, E., Francon, D., Blache, P.: Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. Journal on Multimodal User Interfaces **13**, 41–51 (2019)
34. Okada, S., Ohtake, Y., Nakano, Y.I., Hayashi, Y., Huang, H.H., Takase, Y., Nitta, K.: Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, pp. 169–176. ACM, New York, NY, USA (2016). DOI 10.1145/2993148.2993154. URL http://doi.acm.org/10.1145/2993148.2993154
35. Poyade, M., Morris, G., Taylor, I., Portela, V.: Using mobile virtual reality to empower people with hidden disabilities to overcome their barriers. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 504–505. ACM, New York, NY, USA (2017). DOI 10.1145/3136755.3143025
36. Recht, S., Grynszpan, O.: The sense of social agency in gaze leading. Journal on Multimodal User Interfaces **13**, 19–30 (2019)
37. Reeves, B., Nass, C.I.: The media equation: How people treat computers, television, and new media like real people and places. Cambridge university press (1996)
38. Sims, C.M.: Do the big-five personality traits predict empathic listening and assertive communication? International Journal of Listening **31**(3), 163–188 (2017). DOI 10.1080/10904018.2016.1202770. URL https://doi.org/10.1080/10904018.2016.1202770
39. Skinner, B.: Science And Human Behavior. Free Press paperback. Psychology. New York, Macmillan (1953)
40. Sveinbjornsdottir, B., Johannsson, S.H., Oddsdottir, J., Siguroardottir, T.P., Valdimarsson, G.I., Vilhjalmsson, H.H.: Virtual discrete trial training for teacher trainees. Journal on Multimodal User Interfaces **13**, 31–40 (2019)
41. Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., Nakamura, S.: Detecting dementia through interactive computer avatars. IEEE Journal of Translational Engineering in Health and Medicine **5**, 1–11 (2017). DOI 10.1109/JTEHM.2017.2752152
42. Tanaka, H., Negoro, H., Iwasaka, H., Nakamura, S.: Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. PLOS ONE **12**(8), 1–15 (2017). DOI 10.1371/journal.pone.0182151. URL https://doi.org/10.1371/journal.pone.0182151
43. Tanaka, H., Negoro, H., Iwasaka, H., Nakamura, S.: Listening skills assessment through computer agents. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18, pp. 492–496. ACM, New York, NY, USA (2018). DOI 10.1145/3242969.3242970. URL http://doi.acm.org/10.1145/3242969.3242970
44. Tanaka, H., Sakriani, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H., Nakamura, S.: Teaching social communication skills through human-agent interaction. ACM Trans. Interact. Intell. Syst. **6**(2),

18:1–18:26 (2016). DOI 10.1145/2937757. URL http://doi.acm.org/10.1145/2937757

45. Tanaka, H., Sakti, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H., Nakamura, S.: Automated social skills trainer. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15, pp. 17–27. ACM, New York, NY, USA (2015). DOI 10.1145/2678025.2701368. URL http://doi.acm.org/10.1145/2678025.2701368
46. Tanaka, H., Watanabe, H., Maki, H., Sakriani, S., Nakamura, S.: Electroencephalogram-Based Single-Trial Detection of Language Expectation Violations in Listening to Speech. Front Comput Neurosci **13**, 15 (2019)
47. Tsai, M.N., Wu, C.L., Tseng, L.P., An, C.P., Chen, H.C.: Extraversion Is a Mediator of Gelotophobia: A Study of Autism Spectrum Disorder and the Big Five. Front Psychol **9**, 150 (2018)
48. Tyagi, B.: Listening: An important skill and its various aspects. The Criterion An International Journal in English **12**, 1–8 (2013)
49. Van Hecke, A.V., Stevens, S., Carson, A.M., Karst, J.S., Dolan, B., Schohl, K., McKindles, R.J., Remmel, R., Brockman, S.: Measuring the plasticity of social approach: a randomized controlled trial of the effects of the PEERS intervention on EEG asymmetry in adolescents with autism spectrum disorders. J Autism Dev Disord **45**(2), 316–335 (2015)
50. Veltman, K., de Weerd, H., Verbrugge, R.: Training the use of theory of mind using artificial agents. Journal on Multimodal User Interfaces **13**, 3–18 (2019)
51. Ward, N.G., Escalante, R., Bayyari, Y.A., Solorio, T.: Learning to show you're listening. Computer Assisted Language Learning **20**(4), 385–407 (2007). DOI 10.1080/09588220701745825. URL https://doi.org/10.1080/09588220701745825
52. Zhao, R., Li, V., Barbosa, H., Ghoshal, G., Hoque, M.E.: Semi-automated 8 collaborative online training module for improving communication skills. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(2), 32:1–32:20 (2017). DOI 10.1145/3090097. URL http://doi.acm.org/10.1145/3090097