Listening Skills Assessment through Computer Agents

Hiroki Tanaka  
Graduate School of Information Science, Nara Institute of Science and Technology, Japan  
hiroki-tan@is.naist.jp

Hidemi Iwasaka  
Developmental Center for Child and Adult, Shigisan Hospital, Japan  
iwasaka@heartland.or.jp

Hideki Negoro  
Center for Special Needs Education, Nara University of Education, Japan  
gorosan@nara-edu.ac.jp

Satoshi Nakamura  
Data Science Center, Nara Institute of Science and Technology, Japan  
s-nakamura@is.naist.jp

1 INTRODUCTION

Social communication skills are critical factors that influence human life. Persistent social skill deficits impede those with such afflictions from forming relationships or succeeding in social situations. Social skills training (SST), 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 would simplify the acquisition of such social skills by those who require them.

It may also be easier for those with social difficulties to use computer agents than to directly interact with a human. Using computer agents in SST is motivated by the fact that even though people with social difficulties have difficulty during social communication, they also show good or sometimes even superior systemizing skills [3, 9]. 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 can exploit the following facts: 1) such people favor 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 until the goal is achieved; and 4) interest and motivation can be maintained through computerized rewards.

Previous works conducted social skills training using computer agents [6], for instance, in the contexts of narrative, public speaking, and emotional regulation [14, 26, 27, 31].

The Bellack method (named step-by-step SST) [5] is a well-established and widely used approach. It defines the framework of SST, and defines four basic skills in SST: speaking skills, listening skills, asking skills, and expressing feeling of discomfort. Here, important aspects of listening skills are: 1) looking a partner in the eye, 2) nodding, and 3) repeating keywords of the conversation partner [5]. Listening skills considered in this study are not for hearing and understanding of speech [4, 7]. Listening skills are to explicitly express that one is listening to the partner’s speech [5, 29]. One study showed that people with social difficulties don’t tend to look conversation partners in the eye [15].

In spite of the importance of listening skills, most automated SSTs focused on users’ speaking skills. Okada et al. proposed assessing interaction skills considering listening attitudes [23]. There have been differently motivated works designed to generate (model) human-like head tilting and nodding on humanoid robots or virtual agents based on analyzing human behaviors [12, 13, 16, 19, 21]. Ward et al. proposed listening-skills training...
that produces immediate feedback, although they did not use computer agents and focused only on back-channeling behavior [30]. Also, one previous work showed that personality [10] was related to empathic listening skills [25].

In this paper, we hypothesized that the SST process in listening skills can be automated for interaction between humans and computer agents. As a first step, we analyzed a part of the automatic assessment of listening skills. We collected listening data between interaction of graduate students and computer agents, and investigated the possibility of automatically assessing user’s listening skills.

2 COMPUTER AGENTS

MMDAgent [18] was used as the computer agent. We used default parameters for the agent’s speech such as speaking rate and voice pitch.

Four Japanese people (two males and two females) created the agent’s spoken sentences. Here, one person is a license psychiatrist, who had more than three years of experience with SST, and one is a licensed speech therapist. We created three types of tasks: Speaking, Listening 1, Listening 2. We explain them in detail as follows.

(1) Speaking: The user tells a recent memorable story to the computer agent. This module follows the same procedure as a previous work [27].

(2) Listening 1: The user listens to the agent’s recent memorable story. Table 1 shows sentences we created (translated into English). This supposes casual social small talk.

(3) Listening 2: The user listens to a procedure of how to make a telephone call. These sentences are shown in Table 1 (translated into English). They are designed for a more serious situation such as job training.

Regarding Listening 1 and Listening 2, the agent spoke for about one minute. There were several three- or five-second pauses between the sentences. During the pause, the agent nodded her head if the user said something, and waited three more seconds after the final user utterance.

3 DATA COLLECTION

3.1 Participants

We recruited 27 participants (6 females and 21 males, with a mean age of 25.1, SD: 2.13) from the ***. We confirmed that participants had no hearing difficulties by directly asking them. The first author explained the experiment to the participants to obtain informed consent. The participants completed continuous Speaking (60 sec), Listening 1 (60-90 sec), Listening 2 (60-90 sec) sessions. The completion time changed according to how many times the participants spoke.

3.2 Procedure

The first author explained how to use the system by playing an example video showing head nodding and backchannel feedback.

Data was 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 placed on the bottom of the laptop screen. We turned off the light in the room to minimize external distractions (Figure 1).

After collecting data, we conducted two questionnaires explained in a later section from all participants. Total amount of time for all procedures was approximately 20 minutes. From the collected data, we calculated the following eye fixation, video, and audio features. These features were selected based on previous studies [13, 16, 21], specifically important aspects of listening skills from the Bellack method of SST [5].

3.3 Eye Fixation

An IV-filter was applied to the raw eye-gaze data. We calculated the following features: 1) standard deviation of horizontal fixation axis, 2) standard deviation of vertical fixation axis. We also manually categorized areas of interest as follows: 3) eyes, 4) mouth, 5) face, 6) other.

3.4 Video Annotation

We coded head nodding (video) and speech (audio) using the ELAN tool. The following information was coded by one male annotator: backchannel feedback (e.g. “un”, “hai” in Japanese), repetition of agent’s utterance (paraphrase), question, miscellaneous utterance, head nod (once), head nod (twice), and head nod (three or more times). Here, we defined more than one second as separate coding.

As an output of our coding, we extracted the following features: 1) the number of backchannel feedback instances, 2) the number of repetitions, 3) the number of questions, 4) the number of miscellaneous utterances, and 5) the number of nods. Here, we simplified that the number of nods is counted as the total amount of head nods of once, twice, and three or more times based on [21].

3.5 Social Responsiveness Scale and Big Five Personality Test

We conducted two questionnaires: the Social Responsiveness Scale (SRS) [8], which is related to autistic traits based on the DSM-V [1], and the Big Five Personality Test [11]. The Big Five Personality Test consists of the following subareas: extraversion, agreeableness, conscientiousness, neuroticism, and openness. The relationship between these two questionnaires was investigated in [28].
4 EXPERIMENTAL EVALUATION

This section represents our experimental evaluation of the collected data. After analyzing the relationship between each feature, we finally evaluated our prediction model toward automatic listening skills assessment. In most of this study, Spearman’s correlation coefficient was used to observe correlation. We selected good model persons who scored above five in both listening skills according to [2, 27] to be used as good examples in SST.

First, we analyzed the relationship between each question and listening skills. Then, we normalized extracted features using the z-score normalization. Regarding automatic assessment, we used multiple linear regression, which is a very simple linear approach to predicting listening skills. Leave-one-person-out cross validation was performed. We automatically selected features based on Akaike’s information criterion (AIC) in a stepwise algorithm on the training set. Moreover, the random forest regression was used to assess listening skills as a non-linear model. We set the number of variables tried at each split as four.

Finally, after confirming normality by the Kolmogorov-Smirnov test, we calculated the Pearson’s correlation coefficient between actual values and predicted values. We also calculated the root mean square error (RMSE).

5 RESULTS

5.1 Correlation Analysis

The correlation coefficient of Listening 1 and Speaking was 0.31 (p=0.10), that of Listening 2 and Speaking was 0.41 (p=0.03), and that of Listening 1 and Listening 2 was 0.54 (p=0.003). Five persons were selected as good model persons. Although we found no significant differences between other metrics (SRS and Big Five Personality) and listening, we observed SRS mean values of 47.2 \(\text{SD: } 19.8\) for good persons and 20.5 \(\text{SD: } 19.8\) for other persons (Wilcoxon rank sum test (one-tailed): p=0.03).

Table 1: Sentences spoken by computer agents. \(<\text{pause}>\) denotes three seconds of silence, and \(<\text{long pause}>\) denotes five seconds of silence.

<table>
<thead>
<tr>
<th>Table 2: Top five features. Brackets denote correlation coefficient ((**: p &lt; 0.01, *: p &lt; 0.05)).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

3.6 Ground Truth of Listening Skills

Two licensed clinical psychologists (females), who had more than three years of experience with SST, rated listening skills as well as speaking skills by watching videos exported from the Tobii video recorder. Our instruction to the raters was that you should pay attention to the participants’ impression in addition to behaviors such as eye movement, head nodding, facial expression and speech as with the usual SST. We also directed that after watching multiple videos, they should evaluate our participants’ overall listening skills with Likert scores on a scale of 1 (not good) to 7 (good) [27].

Kappa statistics of the two raters were calculated using a weighted Kappa set to be 1 (on the diagonal) and decreased weights off the diagonal [17]. The following are weighted Kappa correlation coefficients: 0.37 (Speaking), 0.47 (Listening 1), and 0.59 (Listening 2). This is a fair to moderate agreement in accordance with [17]. Also, we calculated Pearson’s correlation coefficients of two raters as follows: 0.44 (Speaking), 0.46 (Listening 1), and 0.66 (Listening 2) (all, p<0.05). Finally, we averaged the two raters’ score for further analysis.
Regarding eye fixation, we found mean values of 40% for eyes, 7% for mouth, 88% for face, and 5% for other within all fixation points in all participants. Table 2 indicates top five correlated features to listening skills. We can see that # of nods was significantly related to the listening skills. Figure 2 shows a relationship between these two attributes. Figure 3 shows two examples from persons who scored 6 (P12) and 2 (P9) in terms of head nodding and backchannel feedback in Listening 2. We confirmed that persons with a low listening score tended to nod only during the agent’s pauses. In contrast, persons with a high listening score also nodded and uttered at other times. For example, they tended to respond at positions of specific keywords, commas and periods of the agent’s transcripts within the sentence according to pitch of agent’s speech.

5.2 Listening Skills Assessment

For the linear regression, we found a correlation coefficient between the predicted value and actual values as follows: Listening 1 was 0.45 (p=0.01), and Listening 2 was 0.47 (p=0.01). The RMSE was 1.58 (Listening 1) and 1.19 (Listening 2). In contrast, the random forest regression obtained the following correlation coefficients: 0.26 (Listening 1) and 0.34 (Listening 2), which were lower performance than the linear regression. This indicates the feature set might be linearly related to listening skills.

Here, we tried to find important features identified from the linear regression. We counted remaining times of each feature in each fold. We confirmed that mouth, face, eyes, the number of nods, and the number of repetitions. were important for Listening 1, and SD of vertical fixation axis, face, the number of backchannel feedback instances, the number of miscellaneous utterances were important for Listening 2.

6 DISCUSSION

We collected data from three types of settings in human-computer interaction. Several behavioral features were coded and extracted based on [5, 13, 16, 21], and the linear regression model achieved prediction of 0.45 in Listening 1 and 0.47 in Listening 2 of correlation coefficients. We confirmed that the correlation coefficient of two raters was 0.46 (Listening 1) and 0.66 (Listening 2), and that our prediction model achieved similar prediction in Listening 1 and a previous work on speaking skills [27]. In the case of Listening 2, the human raters were more agreed than our prediction model. In Listening 2, we found that P5 (female) was predicted as 3.7 in our model; in contrast, the two raters rated 5 and 7. Although P5 responded with not very much backchannel feedback and nodding, she had appropriate timing, speech was loud and clear, and she smiled. We need to consider and extract such additional information to improve our model [16]. In the future, we should compare persons who obtained a high SRS score and others in order to know the normal range of human behavior in this regard. We also found that the amount of backchannel feedback was more important in Listening 2 than in Listening 1 because Listening 2 was a more serious type of interaction, requiring explicit cues to show one was listening.

This study did not investigate the effects of human-computer interaction and human-human interaction. A previous work suggested that people treat computers as real people, showing that people are polite to computers [24]. In contrast, a recent work found that the dynamics of facial expressions differ for users interacting with a human or a virtual agent [22]. We need to consider these effects in the future. We will integrate our listening-skills assessment into the automation framework [20] and test it on people with autism spectrum disorders.

ACKNOWLEDGMENTS

The part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101, and JP18K11437.
REFERENCES


