Automated Social Skills Training with Audiovisual Information

Hiroki Tanaka[1], Sakriani Sakti[1], Graham Neubig[1], Hideki Negoro[2], Hidemi Iwasaka[2] and Satoshi Nakamura[1]

Abstract— **People with social communication difficulties tend to have superior skills using computers, and as a result computer-based social skills training systems are flourishing. Social skills training, performed by human trainers, is a well-established method to obtain appropriate skills in social interaction. Previous works have attempted to automate one or several parts of social skills training through human-computer interaction. However, while previous work on simulating social skills training considered only acoustic and linguistic features, human social skills trainers take into account visual features (e.g. facial expression, posture). In this paper, we create and evaluate a social skills training system that closes this gap by considering audiovisual features regarding ratio of smiling, yaw, and pitch. An experimental evaluation measures the difference in effectiveness of social skill training when using audio features and audiovisual features. Results showed that the visual features were effective to improve users' social skills.**

I. INTRODUCTION

Many people have difficulties in social interactions such as presentations and job interviews [1]. Persistent social skill deficits impede those afflicted with them from forming relationships or succeeding in social situations. Social skills training (SST) is a general cognitive behavior therapy to obtain appropriate these social skills for people who have difficulties in social interaction, and is widely used by teachers, therapists, and trainers [2], [3]. If the SST process could be automated, it would become easier for those requiring SST to receive it anywhere and anytime.

Previous works have attempted to automate one or several parts of SST (see reviews in [4], [5]), for instance, in the context of job interviews [6], public speaking [7], or emotional expression [8]. Specifically, Tanaka *et al.,* [9] developed a system named "automated social skills trainer" that entirely follows the basic training model of SST through human-agent interaction. The system provides feedback according to extracted audio features, and the feedback was shown to improve users' social skill.

While this work is a first step, there still are a large number of gaps between human-based SST and the automated social skills training. One of the gaps is related to modality. While the previous work considered only acoustic and linguistic features, visual information (e.g., facial expression, head pose, and posture) is also an essential feature of human-based SST [2], [3] and public speaking aids [10]. In this paper, we use the automated social skills trainer as a baseline, and extend the system by adding audiovisual information to achieve more effective improvement in users' social skills. This system can be used by not only people who have difficulties in social interaction but also people with autism spectrum disorder (ASD) [8], [9].

*This work was supported by JSPS KAKEN 26540117

[1]H. Tanaka, S. Sakti, G. Neubig, and S. Nakamura are with Graduate School of Information Science, Nara Institute of Science and Technology, Takayama-cho 8916-5, Ikoma-shi, Nara, Japan {hiroki-tan,ssakti,neubig,s-nakamura}@is.naist.jp

[2]H. Negoro, and H. Iwasaka are with Center for Special Needs Education, Nara University of Education, Takabatake-cho, Nara-shi, Nara, Japan {gorosan,hiwasaka}@nara-edu.ac.jp

II. SST AND PREVIOUS WORK

The basic training model of SST [3] generally follows steps of instruction, modeling, role-playing, feedback, and homework. We briefly describe these steps and their implementation in the automated social skills trainer as follows:

1) **Instruction:** Instruction includes defining target skills and explaining their goal. After the major social problems faced by the trainee are identified, the skills to be trained are decided based on these problems. The automatic social skill trainer sets narrative as a target skill, which is a task of telling stories, and a basic skill necessary for other higher-level skills [3].
2) **Modeling:** Trainers act as a model, demonstrating the skill that the users are focusing on so that the users can see what they need to do before attempting to do it themselves. The automated social skills trainer replicates this for narrative skills by allowing users to watch a recorded model video of people who have relatively good narrative skills.
3) **Role-playing:** Users are asked to role-play. When the user says "start role-playing," the system says "please tell me about something fun you experienced recently." The role-playing starts after the system's question, and continues for one minute. During this time, the avatar nods its head, and the system automatically senses and analyzes features from the video of the user.
4) **Feedback:** At the end of role-playing, the system analyzes the features of the user's video and displays feedback based on this analysis. This feedback helps users to identify their strengths and weaknesses. Because displaying a large number of features in the feedback may confuse users, the system performs feature selection to identify the features effective in defining narrative skills. This process was conducted in discussion with a professional social skills trainer.
5) **Homework:** The system sets little homework challenges that users perform in their own time throughout the week. The system tells users to "please tell your story to others throughout the week, and let me know about it."

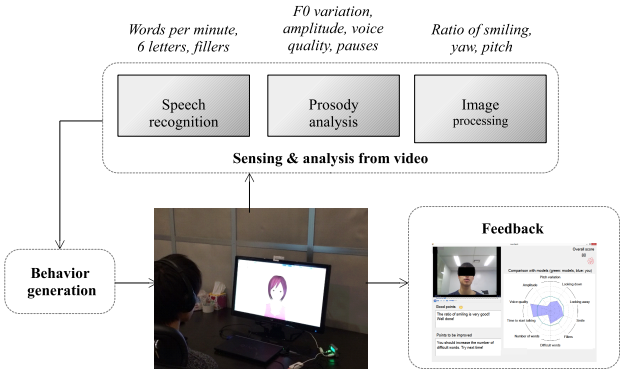

Fig. 1. System framework.

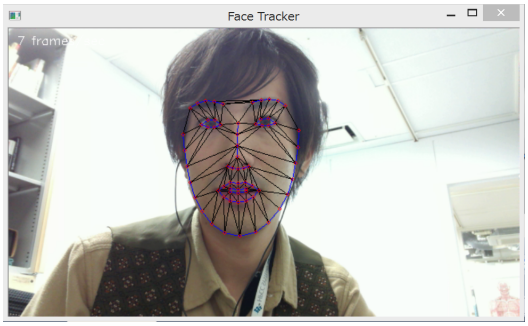

Fig. 2. Extracted points using the face tracker.

III. Implementation of The Multimodal System

The role-playing and feedback are the most important elements of the system, which consist of sensing and analysis from video, and feedback visualization (Figure 1). In this section, we describe the integration of multimodal information into each module.

A. Sensing and analysis from video

To analyze the visual information of the video, we extracted a number of facial features by using a constrained local model (CLM) [11] based face tracker[1] as shown in Figure 2.

From a total of 66 feature points, we calculated the features of outer eye-brow height (OBH), inner eyebrow height (IBH), outer lip height (OLH), inner lip height (ILH), eye opening, and lip corner distance (LipCDT) following Naim et al. [12]. Using these features, we modeled smiling faces using The Japanese Female Facial Expression (JAFFE) database [13]. The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) performed by 10 Japanese female models. Each image has been rated with regards to 6 emotion adjectives by 60 Japanese subjects. In the database, we used 31 samples of happy faces and 30 samples of neutral faces. We trained a model of two types of facial expression using SVM with linear kernel. Precision, recall, and F-measure of leave-one-out cross validation over the database were .91, .97, and .94 respectively.

For video, we predicted whether the label belongs to the smiling or neutral class in each frame, and the proportion of smiling frames among all frames was named the ratio of smiling. To verify that the model generalizes to other speakers and video, we used the NOCOA+ database [14], which contains derisive and friendly videos of four Japanese men. We extracted the ratio of smiling from these videos and confirmed that friendly videos have a significantly larger ratio of smiling compared to derisive videos (p=0.002, Cohen's d=1.89).

In addition to the smile features, we separately incorporated two head pose features (yaw and pitch), based on the corresponding elements of the global transformations associated with rotation. Here, the yaw indicates horizontal direction, and the pitch indicates vertical direction of head pose. These features reflect looking away and looking down during talking [12]. We calculated the absolute value of the yaw and took the average of the entire frame to analyze shift from the front. With regard to the pitch, both facing up (negative values) and facing down (positive values) were considered to be important, and thus we calculated the average value of the entire video without taking the absolute value. Because the system did not record images below the chest, we did not take into account non-facial gestures.

Extracted features related to speech and language follow Tanaka et al. [9] as follows: F0 variation, amplitude, voice quality, pauses, words per minute, words of more than 6 letters, fillers.

B. Feedback

Based on the calculated features, the system displays feedback to the users (Figure 3). We attempt to display this feedback in a way that is easier to understand and interpret as follows:

- **User video:** Users can watch the recorded video and audio in the narrative.
- **Overall score:** The system displays a predicted overall score, which motivates the user to practice more and improve their score. We predict the overall score using the generalized linear multiple regression method on a scale of 0 to 100 [9].
  With regard to features of the regression model, because we analyzed the data and found that ratio of smiling for models was significantly higher than others (p=0.04, d=1.21), we added ratio of smiling as input features of the regression model. We confirmed that the correlation coefficient between the predicted narrative skills and subjectively evaluated skill using leave-one-user-out cross validation was 0.55 (p=0.02), which indicates a weak correlation, when using statistically significant features (words per minute, amplitude, words of more than 6 letters, and ratio of smiling). We also confirmed that the correlation coefficient was 0.51 when using only speech and language features [9] showing that ratio of smiling provides a slight gain in correlation with subjective evaluation.

[1]https://github.com/kylemcdonald/FaceTracker

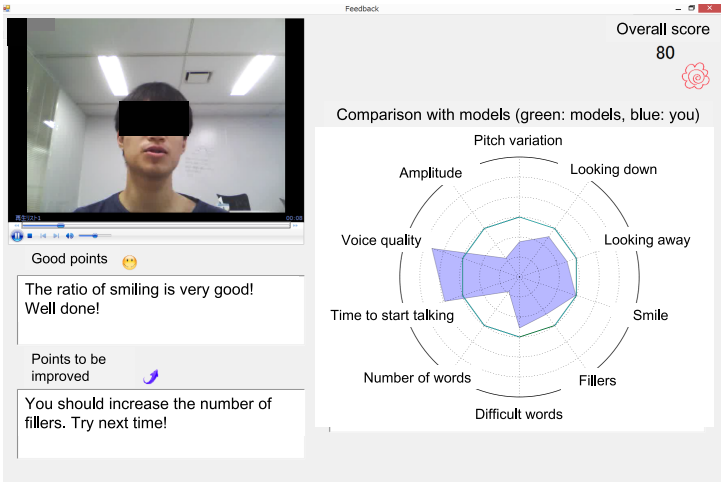

Fig. 3. Audiovisual feedback provided by the automated social skills trainer.

- **Comparison with models:** The system uses a radar chart to visualize the 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.
- **Comments:** The system generates positive comments that reinforce the user's motivation based on the features that have values closest to those of the models. In addition, the system also generates comments about points to be improved based on the feature that has a median distance from the models. This choice of the median, instead of the farthest, feature was made through discussions with a professional social skills trainer, who noted that it may be fundamentally difficult for people with communication difficulties to improve their worst points.

IV. Experiment: Training Effect

In our experiment, we examined the difference in effectiveness of social skill training when using feedback related to audio features and feedback related to audiovisual features.

A. Procedure

We recruited a total of 18 graduate students (15 males and 3 females) all of whom were native Japanese speakers. Participants were given instructions by the first author and told that their speech and video would be recorded[2]. A webcam placed on top of the laptop and headset recorded the video and audio of participants.

We separated participants into two groups: the audio group (6 males and 3 females), and the audiovisual group (9 males)[3]. All participants spoke a story to a familiar person (pre), used the automated social skills trainer for 50 minutes, and spoke a story to the same familiar person (post). In the training, participants followed the procedure of the basic training model. Because we did not control the duration of watching videos or performing role-play, the participants can select the content by themselves. The audio group received feedback regarding speech and language features, while the audiovisual group received feedback regarding not only audio but also ratio of smiling, yaw, and pitch features.

We asked an experienced social skill trainer, who has been a supervisor of young adult developmental support performing social skills training more than three years, to evaluate the overall narrative skills according to a Likert score rated on a scale of 1 (not good) to 7 (good). He watched randomly ordered pre and post videos and rated the score.

To find important features that related to narrative skills, we calculated statistical differences of each feature between recorded videos of the relatively low scores (low score group) and the relatively high scores (high score group). We also calculated $post-pre$ scores to measure the effect of training. We present the p-values of one-tailed t-tests and Cohen's d values as a measure of the effect size.

B. Result

First, we examined differences in audiovisual features between groups with lower and higher levels of social skills. Because the social skill trainer's rating was between 3 and 6 (mean: 4.5), we separated each video into the low score group (3 or 4, n=16) and the high score group (5 or 6, n=20). We found a statistical difference between the two groups in the case of amplitude (the high score group was significantly louder than the low score group, p=0.03, d=0.63), ratio of smiling (the high score group smiled significantly more than the low score group, p=0.03, d=0.66), and pitch (the high score group looked up significantly more than the low score group, p=0.04, d=0.63) in Figure 4. This showed that amplitudes, ratio of smiling, and pitch are strongly related to narrative skills.

[2]Note that the Research Ethic Committee of our institution has reviewed and approved this experiment. Written informed consent was obtained from all participants before the experiment.

[3]It should be noted that while the genders are not balanced, we also performed the forthcoming analysis without including the female participants, and it did not affect any of the statistical differences.

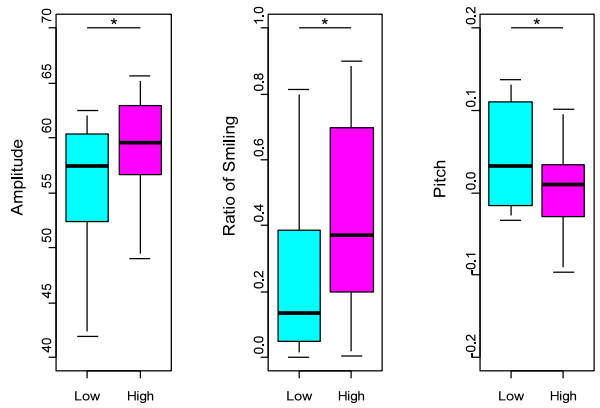

Fig. 4. Boxplot of audiovisual features between the low score group and the high score group.

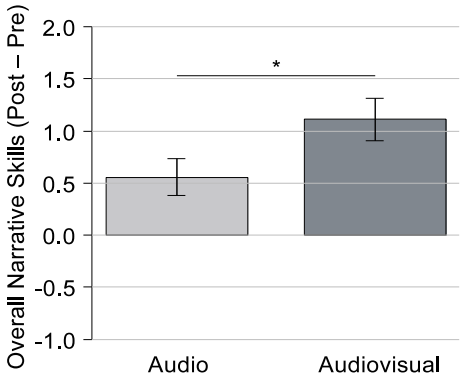

Fig. 5. The improvement of overall narrative score in two groups. Error bars indicate standard error (*: $p < .05$).

Next, we examine the effect of training using audio or audiovisual features. Before training, the initial scores of the two groups were not significantly different (p=0.96 (two-tailed t-tests), d=0.03): the audio group had a mean of 4.0 (sd: 0.91) and the audiovisual group had a mean of 4.2 (sd: 0.83). Figure 5 shows the improvement of overall narrative skills in two groups. These results show that audiovisual feedback significantly affected the increase in overall narrative skills (p=0.03, d=0.98). The overall gain in skills using audiovisual feedback was 1.1, which is comparable to, or slightly greater than similar previous work (around a 1.0 point improvement through interview skill training for one week using a virtual tutoring agent [6], and around a 0.7 point improvement through narrative skill training [9]).

The advantage of audiovisual feedback compared to audio can likely be attributed to slight improvements in ratio of smiling (p=0.09, d=0.66), with the audiovisual group achieving a mean improvement of 0.088 (sd: 0.19) and the audio group seeing a loss of -0.026 (sd: 0.15).

These results also reflect knowledge of human-based SST, which has shown the importance of smiling and facing straight to express that the speaker is having fun [2].

V. Conclusion

We extended a method for automatic social skill training by adding audiovisual information. We extracted features regarding ratio of smiling and head pose. The experimental evaluation confirmed the training effect and the relationship between narrative skills and extracted features.

For future directions, we examine the timing of each feature (e.g. smiling) in more detail. We would like to add other target social skills as referred to by Bellack [3], and compare with human-based SST. Furthermore, we would like to confirm the training effect over a longer period, and recruit special-need populations such as people with ASD who are potential users of the system.

Acknowledgment

We would like to thank children of Nara autism society, and Probono Nara for their helpful comments. This study was supported by JSPS KAKEN 26540117.

References

[1] A. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[2] R. Liberman and C. Wallace, "Social and independent living skills: Basic conversation skills module," *Camarillo, Calif: Author*, 1990.

[3] A. S. Bellack, *Social skills training for schizophrenia: A step-by-step guide*. Guilford Press, 2004.

[4] N. Aresti-Bartolome and B. Garcia-Zapirain, "Technologies as support tools for persons with autistic spectrum disorder: a systematic review," *International journal of environmental research and public health*, vol. 11, no. 8, pp. 7767–7802, 2014.

[5] J. A. Kientz, M. S. Goodwin, G. R. Hayes, and G. D. Abowd, "Interactive technologies for autism," *Synthesis Lectures on Assistive, Rehabilitative, and Health-Preserving Technologies*, vol. 2, no. 2, pp. 1–177, 2013.

[6] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 697–706.

[7] M. I. Tanveer, E. Lin, and M. E. Hoque, "Rhema: A real-time in-situ intelligent interface to help people with public speaking," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 2015, pp. 286–295.

[8] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, D. Pigat, P. Robinson, and I. Daves, "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.

[9] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, and S. Nakamura, "Automated social skills trainer," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 2015, pp. 17–27.

[10] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer, "Exploring feedback strategies to improve public speaking: An interactive virtual audience framework," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 1143–1154.

[11] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 1034–1041.

[12] I. Naim, M. I. Tanveer, D. Gildea *et al.*, "Automated analysis and prediction of job interview performance," *arXiv preprint arXiv:1504.03425*, 2015.

[13] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 200–205.

[14] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Nocoa+: Multimodal computer-based training for social and communication skills," *IEICE Transaction on Information and Systems*, vol. E98-D, no. 8, pp. 1536–1544, 2015.