Automated Social Skills Coach with Real-Time Feedback on Nonverbal Cues

Mohammad Rafayet Ali
Computer Science, University of Rochester, Rochester, NY, United States

***Abstract*—Nonverbal cues are considered the most important part in social communication. An automated social skills coach has the potential to help overcome the difficulties of having the social skills training. In this work, we aim to explore the different aspects of the real-time feedback on nonverbal cues during a conversation. We developed an initial prototype and using a Wizard-of-Oz technique we conducted a study with 47 individuals. We collected videos of the participants having a conversation with a virtual agent. This study also revealed that the participants who used our system improved their gesture in a face-to-face conversation. Our goal is to explore different machine learning techniques on the facial and prosodic features to automatically generate feedback on the nonverbal cues. Also, in future, this work will uncover the effectiveness of different ways of delivering real-time feedback during a conversation.**

I. Introduction

Social skills are essential for social competence. Effective nonverbal communication is a key component of social skills. In order to form and maintain [1], [2] social communication, effective nonverbal communication is essential. Practicing conversations with other people, preferably an expert, is considered to be one of the best ways to improve the social skills. However, due to social stigma, expense and time constraints it is often difficult to have such training. Computers and internet based technologies have the potential to help individuals improve their social skills. Imagine we have a conversational agent in our computer with whom we can have a conversation and receive feedback on nonverbal cues. The deficit of effective nonverbal communication includes lack of eye contact, soft tone of voice, lack of proper facial expression (i.e., smile), bad posture etc. It would be helpful if computers could remind the users about the appropriate use of the nonverbal behavior. Generating feedback to help improve these deficits requires sensing human nonverbal behaviors, and understanding the nonverbal cues. For many years researchers are working on detecting the facial expression [3], [4]. However, generating real-time feedback on nonverbal cues based on the facial and prosodic features is more difficult problem than recognizing the facial expressions. Also, it is necessary to deliver the feedback in a meaningful, personalized, nondistractive, and generalizable way.

Face-to-face conversations are dynamic in nature. During a conversation, we have little time to respond. In the response period, we need to think about our response as well as exhibit proper nonverbal behavior. For a computer, conducting conversations and giving feedback simultaneously would be very hard since it has limited computational power, no subconscious mind, and finite memory.

We propose an automated system which allows users to have a conversation with a virtual agent and receive feedback on the nonverbal behaviors. In addition to the real-time feedback, we want to explore the effects of post feedback. As a preliminary exploration, we have developed a Wizard-of-Oz driven prototype of the system. To test the viability of our system design we conducted a speed-date study on 47 individuals. We found that the participants who used our system showed significant improvement in head nodding and marginal improvement in eye contact and gesturing. We collected audio and video data from the study for automation. Our goal is to build an automated system that can conduct a natural face-to-face conversation, provide non-distracting real-time feedback, summarize the whole conversation in an easily interpretable way, and enable users to apply the skills in real life.

II. Related Work

Numerous studies focused on conversational agents and giving feedback to help people improve their social skills. Tanaka et al. [5] showed the effects of nonverbal behaviors in training social skills. They presented an automated social skills trainer [6],[7] that uses a virtual agent to have conversations with users and capture features of the user's speech (e.g. amplitude, voice quality, words per minute, etc.). Their system gives post feedback after analyzing the speech features which helped people who have autism spectrum disorder. The TARDIS framework [8] also worked with virtual agents to provide realistic socio-emotional interaction in the context of simulated job interviews. It senses nonverbal behaviors of participants, including gaze direction, prosody, and gesture, in a pre-determined job interview scenario. Cicero [9] explored the possibility of using an interactive virtual audience for public speaking training using the speaker's nonverbal features. In MACH [10], a virtual agent asks a series of interview questions, providing neutral acknowledgments by mirroring smiles and head nods, in a simulated interview, and follows up with detailed feedback on the user's performance at the end. This project demonstrated the viability of using virtual agents and summarized feedback in the context of a job interview.

Many researchers have worked on minimizing the distraction of having real-time feedback during social interactions. Kulyk et al. [11] presented a tool for visualizing the nonverbal behaviors (i.e., gaze behavior, speaking time) of

humans in small group meetings. They found that the real-time feedback is perceived as a useful feature and that it helps to balance participation. Some past strategies include providing chat-based visual feedback on language use, such as the proportion of agreement words and overall word count, to improve group collaboration [11]. Real-time feedback was also used for training classroom teachers using bug-in-ear technology to increase teachers' rate of praise statements and their use of proven effective instructional purposes [14].

Another significant challenge is to automate conversation with a virtual agent in an open-ended scenario. Many studies used a Wizard-of-Oz technique to simulate the conversation. Boujarwah et al.[12] presented a tool to enable non-expert humans to generate conversational scenarios, which can be used to teach children with ASD appropriate behaviors in different social scenarios. SimSensei [13] used a virtual agent in the context of the healthcare decision support system. The goal of this system is different than our work, as it aims to identify psychological distress indicators through a conversation with a patient in which the patient feels comfortable sharing information. This system has both nonverbal sensing and a dialogue manager. The dialogue manager of SimSensei used four classifiers to understand the users' speech and, based on the classifiers' output, it generates the response. The dialogue manager, called FLoReS [14], was used by the same team in SimCoach [13], which used hundreds of sub-dialogue networks to generate a response by maximizing future reward.

III. Methodological Approach and Key Idea

In order to help people improve their conversational skills, we want to build a system where individuals can practice conversations with a virtual agent and receive feedback. The feedback needs to be natural and non-distracting. We also want to incorporate post summary feedback, which proved to be effective in the past [10]. Designing such a system brings many challenges. The first challenge is delivering real-time feedback in a natural and non-distracting way, during a conversation. We need to extract the facial and prosodic features from the users during the conversation and analyze them in order to generate appropriate feedback. The second challenge is delivering the feedback. In our current prototype, we incorporated flashing icons for giving feedback. However, more natural ways of delivering feedback need to be explored. The third challenge is presenting the post summary feedback in a way that is easy to understand and reflect upon. Post feedback can be helpful for those population who have the difficulty of processing too many information at the same time (i.e., older adults [15]). The whole system needs to be ubiquitously available, addressing the problem of social stigma and costly training sessions. Generalizability of these systems is also a key factor in the system design. People may improve while having a conversation in the virtual environment, but they might still have trouble demonstrating the social skills in a face-to-face interaction. For this reason, it is often useful to train people with the system, then monitor their performance in real-world scenarios, gathering feedback to improve the system further.

IV. Research Carried Out

We designed a system where people can practice their conversational skills with a virtual agent and receive both real-time and post summary feedback. Our system is available online, which allows people to practice anytime, anywhere.

A. System Overview

Our system gives real-time feedback on human nonverbal behavior during a conversation. To minimize cognitive load, we chose four nonverbal cues - eye contact, volume, body movement, and smile. In the future, we will incorporate other nonverbal and verbal cues, as well. Our real-time feedback interface is shown in Fig. 1. Four icons representing four nonverbal behaviors are placed at the bottom of the interface, which can turn red or green, prompting the users to adjust their behavior during the conversation. After the conversation, users can see post summary feedback (shown in Fig. 2), which includes how many times they received feedback (Reminders), for how long they could keep the icons green (Best Streak), and how much time they took to adjust their behavior (Response Lag).

B. Study Design

In order to conduct a speed-date study, we recruited 47 male undergraduate students to participate and 8 female research assistants (RA). Each day we invited four participants and four RAs in our speed-date study. Each participant had a speed date with each of the RAs for four minutes. We then divided the participants into two groups—treatment and control—randomly. The treatment group practiced conversation with our system for 20 minutes, and the control group read a pamphlet [16] and watched a YouTube video [17] on how to improve conversational skills. After the intervention, all the participants had speed date sessions again with the same set of RAs. The RAs were asked to rate the participants using the Conversational Skills Rating Scale (CSRS) [18] scale. We found that the participants who used our system were showed significant improvement in head nodding and marginal improvement in eye contact and gesturing. Fig. 3 shows the study design.

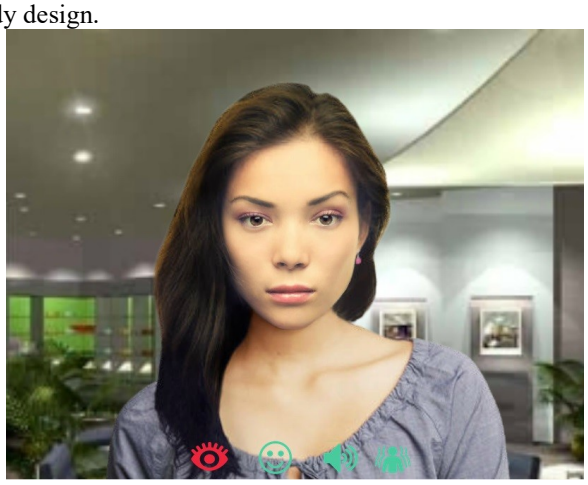

Fig. 1. Real-time feedback interface

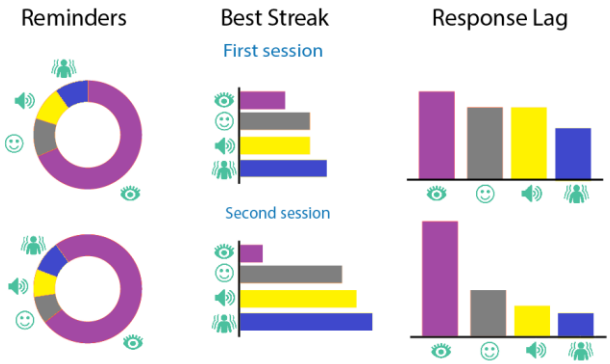

Fig. 2. Post summary feedback interface

C. Data Collection

During the intervention, we collected the videos of the participants interacting with the virtual agent. From those videos, we extracted facial and prosodic features using Praat [19] and SHORE [20]. To extract gaze direction, we used software from Visage Technologies [21].

To label the videos, we recruited the same set of RAs and asked them to label those moments in the video where they felt that the participants needed to receive feedback. We took only those moments of feedback on which at least two RAs agreed. The videos are also transcribed to automate the dialogue manager.

D. Automated System

To automatically generate the feedback, we applied a Hidden Markov Model (HMM), which is able to capture the temporal pattern in the video. For each of the nonverbal cues, two HMMs were trained using the Baum-Welch algorithm—one for the negative class (green icon) and another for the positive class (red icon). To predict the classes for a given video stream, we simply used the forward-backward algorithm to calculate the probability of each class. We used a 10-fold cross-validation method for selecting the parameters of our model and applied it to a test set. The results of the test set are shown in Table 1.

To automate the dialogue manager, we developed a topic-based dialogue manager, which is able to conduct a conversation on a predefined topic. At the top level, the dialogue manager contains the main topics in structures called Schema. Schemas are hierarchically structured, containing subschemas that allow the conversation to be more spontaneous. To capture users' responses, we used the Nuance [22] speech recognizer. For each of the schema and subschema, there are some set of rules that generate questions. After receiving input from the speech recognizer, we generated high-level interpretations depending on the keywords in the response, which then generate the next response for the virtual agent.

During our preliminary study on speed-dating, we collected data in the form of audio and video. This data contains the face of the participants and the conversation they were having with a virtual agent. Additionally, all the videos are labeled. Our initial approach was to utilize the hidden Markov model. However, there is room for improvements. In future, we will

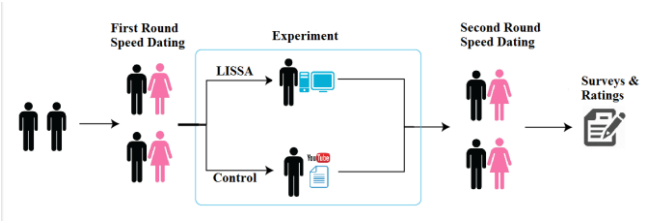

Fig. 3. Study design

V. Future Work

apply other machine learning models including support vector machines (SVM), clustering, and random forest. We also plan to apply deep neural network based classifier. The performance of these models is yet to be explored. The challenge with SVM is that it might overlook the temporal information. Deep learning–based models are robust but require a lot of training data for higher accuracy. We extracted the features using three off the shelf software [20], [21]. Other features might be more important than those which we used in our system. Also, we want to explore the automatic feature selection approaches (i.e., principal component analysis).

Our system gives feedback by flashing icons red and green. However, in a natural face-to-face conversation, this type of feedback is absent. Traditional social skills training [23] includes face-to-face conversation. In order to generalize our system and help people improve their conversational skills in real face-to-face conversations, we need to build a more natural system. One possibility is to eliminate the flashing icon and introduce the feedback in the virtual agent's dialogue.

The automated open-ended conversation is an open problem requiring much prior knowledge to be stored. We used a topic-based dialogue manager with pre-stored dialogues. In future, we want to generate dialogues extemporaneously using deep learning techniques. In addition, we will add more dialogue scenarios to be able to conduct the conversation smoothly. In natural conversations, as we progress, we gather some information from the other partner and use the knowledge to guide our conversation further. In our system, we do not store any information about the user. In the future, it will be desirable to generate dialogues depending on the information stored during the conversation.

Our dialogue manager and feedback module work independently. In an ideal scenario, it is natural that the feedback would affect the dialogues. For example, when the system is providing feedback on smiling (smile icon turns red) continuously and the participant is not paying attention, the dialogue manager can make the virtual agent say, "Are you upset about something?"

TABLE 1. PERFORMANCE MEASURE OF AUTOMATED FEEDBACK SYSTEM

	Accuracy	True Positive Rate	False Positive Rate
Volume	84.55%	72.92%	15.31%
Smile	78.14%	67.53%	37.20%
Body Movement	72.10%	70.84%	27.64%
Eye Contact	78.30%	93.20%	39.50%

Our system was tested only on male college students. In the future, we will engage other populations (e.g., children with autism.) We have also tested our system in one social scenario: speed-dating. We will conduct studies in other scenarios where people need to have good conversational skills (i.e., job interviews).

We often evaluate our work in a lab environment, which might defeat the whole purpose of allowing people use it on their own. In a lab environment, we ask participants to use our system and practice for 20 minutes. If, however, we asked them to use the system from home, we would have no control over the amount of time they practice. Once we deploy the system online there will be a lot of people using it, hence generating a lot of data. We might be able to use this data for future research.

VI. CONCLUSION

We designed a system that allows individuals to practice their conversational skills and receive real-time feedback on their nonverbal behavior. We have shown the effectiveness of our system in improving individuals' conversational skills. From the speed date study, we collected data in the form of audio and video. Using the data, we built an automated feedback module and a dialogue manager. In this paper, we have proposed some ways of exploration which might help improve the system and the overall user experience. Once we have the fully automated system it will open up endless possibilities of research on human behavior, facial expression analysis, and social skills training.

VII. REFERENCE

[1] B. M. DePaulo, "Nonverbal Behavior and Self-Presentation.," *Psychol. Bull.*, vol. 111, no. 2, pp. 203–243, 1992.

[2] C. F. Keating, "The developmental arc of nonverbal communication: Capacity and consequence for human social bonds.," in *APA handbook of nonverbal communication.*, 2016, pp. 103–138.

[3] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, 2000.

[4] J. J. J. Lien, T. Kanade, J. F. Cohn, and C. C. Li, "Detection, tracking, and classification of action units in facial expression," *Rob. Auton. Syst.*, vol. 31, no. 3, pp. 131–146, 2000.

[5] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Modality and contextual differences in computer based non-verbal communication training," in *4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013 - Proceedings*, 2013, pp. 127–132.

[6] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "NOCOA+: Multimodal computer-based training for social and communication skills," *IEICE Trans. Inf. Syst.*, vol. E98D, no. 8, pp. 1536–1544, 2015.

[7] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, and S. Nakamura, "Automated Social Skills Trainer," *Proc. 20th Int. Conf. Intell. User Interfaces*, pp. 17–27, 2015.

[8] H. Jones and N. Sabouret, "TARDIS-A simulation platform with an affective virtual recruiter for job interviews," *hazael.jones.free.fr*, 2012.

[9] L. Batrinca, G. Stratou, A. Shapiro, L. P. Morency, and S. Scherer, "Cicero - Towards a multimodal virtual audience platform for public speaking training," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8108 LNAI, pp. 116–128, 2013.

[10] M. (Ehsan) Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," *Proc. 2013 ACM Int. Jt. Conf. Pervasive ubiquitous Comput. - UbiComp '13*, p. 697, 2013.

[11] O. Kulyk, J. Wang, and J. Terken, "Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 3869 LNCS, pp. 150–161.

[12] F. A. Boujarwah, M. O. Riedl, G. D. Abowd, and R. I. Arriaga, "REACT: intelligent authoring of social skills instructional modules for adolescents with high-functioning Autism," *SIGACCESS Access. Comput.*, no. 99, pp. 13–23, 2011.

[13] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L. Morency, "SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support," in *International Conference on Autonomous Agents and Multi-Agent Systems*, 2014, no. 1, pp. 1061–1068.

[14] F. Morbini, D. DeVault, K. Sagae, J. Gerten, A. Nazarian, and D. Traum, "FLoReS: A Forward Looking, Reward Seeking, Dialogue Manager," *Nat. Interact. with Robot. Knowbots Smartphones*, pp. 313–325, 2014.

[15] S. Getzmann, E. J. Golob, and E. Wascher, "Focused and divided attention in a simulated cocktail-party situation: ERP evidence from younger and older adults," *Neurobiol. Aging*, vol. 41, pp. 138–149, 2016.

[16] D. Wendler, *Improve your social skills*. 2014.

[17] "5 Body Language Tricks To Make Anyone Instantly Like You - Personality Development & English Lessons." .

[18] A. Instructional, A. Of, and I. Competence, "Conversational Skills Rating Scale."

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." .

[20] "SHORE™ - Object and Face Recognition." [Online]. Available: http://www.iis.fraunhofer.de/en/ff/bsy/tech/bildanalyse/shore-gesichtsdetektion.html.

[21] "Visage Technologies face tracking and analysis." [Online]. Available: http://visagetechnologies.com/.

[22] "NUANCE: Speech Recognition Solutions." [Online]. Available: http://www.nuance.com/for-individuals/by-solution/speech-recognition/index.htm.

[23] Thomas E Smith, Alan S Bellack, and Robert Paul Liberman. 1996. Social skills training for schizophrenia: Review and future directions. Clinical Psychology Review 16, 7: 599–617. http://doi.org/10.1016/S0272-7358(96)00025-6