Feedback Strategies on Verbal and Nonverbal Cues to Improve Communication Skills

By

Mohammad Rafayet Ali

Submitted in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

Supervised by Dr. Ehsan Hoque

Department of Computer Science Arts, Sciences and Engineering Edmund A. Hajim School of Engineering and Applied Sciences University of Rochester 2020
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biographical Sketch</td>
<td>vi</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>viii</td>
</tr>
<tr>
<td>Abstract</td>
<td>ix</td>
</tr>
<tr>
<td>Contributors and Funding Sources</td>
<td>xi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Challenges and Contributions</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Thesis Outline</td>
<td>9</td>
</tr>
<tr>
<td>2 Related Work</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Real-time Feedback</td>
<td>12</td>
</tr>
<tr>
<td>2.1.1 Utilization of Verbal and Nonverbal Behaviors</td>
<td>13</td>
</tr>
<tr>
<td>2.1.2 Virtual Agent-based Interventions</td>
<td>14</td>
</tr>
<tr>
<td>2.1.3 Virtual Environment and Virtual Reality</td>
<td>15</td>
</tr>
<tr>
<td>2.1.4 Multiple Modalities</td>
<td>16</td>
</tr>
<tr>
<td>2.2 Post Feedback</td>
<td>17</td>
</tr>
</tbody>
</table>
4.4 Dialogue Module ........................................... 53
  4.4.1 Dialogue Schemas ...................................... 53
  4.4.2 Receiving Input ....................................... 54
  4.4.3 Understanding input ................................... 55
  4.4.4 Reacting to input ..................................... 56

5 LISSA Autism Study ........................................... 57
  5.1 Study Design ........................................... 58
  5.2 Survey Results .......................................... 59
  5.3 Qualitative Analysis ..................................... 60
  5.4 Discussion ............................................... 66
  5.5 Limitations and Future Work ............................ 70

6 Aging and Engaging ........................................... 72
  6.1 Interface ................................................. 73
  6.2 Pilot Study ............................................. 74
  6.3 Results .................................................. 75
    6.3.1 Survey Results ...................................... 75
    6.3.2 Feature Analysis ..................................... 78
    6.3.3 Interview Results .................................... 79
    6.3.4 Discussion .......................................... 82
  6.4 Longitudinal Study ...................................... 85
    6.4.1 Study Design ........................................ 88
    6.4.2 Methods ............................................. 90
    6.4.3 Results ............................................. 93
    6.4.4 Discussion .......................................... 101

7 SOPHIE - Standardized Online Patient for Human Interaction Edu-
cation ........................................................... 108
7.1 Design of SOPHIE .................................................. 110
  7.1.1 Scenario ..................................................... 110
  7.1.2 Feedback Interface ........................................... 111
7.2 Pilot Study ...................................................... 113

8 Limitations and Future Work ........................................... 118
  8.1 Nonverbal Behavior .............................................. 118
  8.2 Feedback Modalities ............................................ 119
  8.3 Study Limitations ............................................... 120
  8.4 Technical Limitations ........................................... 120
  8.5 Generalizability ................................................ 122

9 Conclusion .......................................................... 124

References .............................................................. 138
Biographical Sketch

Mohammad Rafayet Ali received his bachelor’s degree in 2013 from Bangladesh University of Engineering and Technology. He received Dean’s list award and a bachelors’ thesis award after his graduation. He served as a lecturer at Ahsanullah University of Science and Technology from 2013 to 2014. In August 2014, he moved to Rochester to pursue a Ph.D. degree in Computer Science under the supervision of Dr. Ehsan Hoque. During his graduate studies, he published papers (listed below) in several prestigious conferences including UbiComp, IUI, and FG. He collaborated with professors across different fields including, psychology, psychiatry, and family medicine. He interned at Microsoft Research in Summer 2019. He is a student member of SIGCHI and IEEE. Below is the list of his publications.


Sen, Taylan et al. 2018. "Modeling Doctor-Patient Communication with Affective Text Analysis." In 2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017,

Acknowledgements

I would like to thank my advisor, Dr. Ehsan Hoque for his continuous support and guidance throughout my Ph.D. Dr. Hoque provided excellent mentorship to me, which helped me formalize my research. Besides mentorship, he helped me improve my soft skills such as presentation skills. He took the time to review all my papers multiple times before any submission. He set an example of actively involving himself in students’ projects. He always encouraged me to attend conferences and connect with the researchers in my field. Additionally, he helped me connect with my collaborators from outside of the Department of Computer Science.

I would like to acknowledge Dr. Lenhart Schubert who has supported me from the beginning of my Ph.D. studies. His knowledge in dialogue helped us develop prototypes. He helped me in writing and formulating research papers. I am very fortunate to have him as a member of my thesis committee. I am also lucky to have the opportunity to work with Dr. Ronald Epstein. His experience with cancer care and medical education helped a lot in developing the algorithms and interpret the findings. He allowed me to shadow him in his clinical visits which helped me understand the underlying challenges of communication between patients and physicians.

Next, I would like to thank Dr. Zhen Bai for guiding me in writing the thesis. I would like to acknowledge all the help from Dr. Kim Van Orden and Dr. Ronald Rogge, for supporting me with my research and helping me to conduct studies. Dr. Orden gave me the opportunity to work directly with the study population. Without her support, I can not imagine the work with older adults coming into reality. Finally, I would like to thank my colleagues, Taylan Sen, Dev Crasta, Zahra Razavi, Ben Kane, Raina Langevin, Agustin Baretto, and Facundo Ciancio for their continuous support and help.
Abstract

In this thesis, we present findings on designing and validating real-time and post feedback on nonverbal skills in face to face communication skills with a humanoid agent. The technical challenges included a real-time machine learning framework that can automatically process the audio-video data via a webcam, allowing the users to converse in natural language and receive live and post feedback on smile intensity, volume modulation, pauses, synchronicity, body language, eye-contact, sentiment and turn-taking.

Our initial exploration included designing a wizard-of-oz experiment validating the form factors (i.e., flashing icons using the traffic light analogy) for real-time feedback using 46 college students. Using the data, we trained a hidden Markov model to generate feedback. The feedback on verbal cue was generated by performing sentiment and word category analysis. For post feedback, we summarized the nonverbal feedback using the support vector machine.

The technical contributions were validated in three unique contexts: 1) helping individuals with autism; 2) helping elderly with their social skills; 3) helping physicians improve their interactions skills with patients.

Applications to speed-dating and autism: In a randomized control study with 47 college students, we found that the feedback helped improve eye contact and gesture. In a preliminary study with nine teenagers with autism, we identified several design guidelines which include, briefing the users, making positive acknowledgments, and personalizing dialogue.

Applications to aging: In a pilot study with 25 older adults, participants found the feedback useful and were able to reflect on the feedback. In a subsequent longitudinal study with 18 older adults, participants improved their eye contact and smiling.

Applications to patient-physician communication: In the context of patient-physician communication, we conducted a study with eight clinicians where they found the feedback intuitive and easy to follow. Additionally, we identified two communication behaviors of
physicians that help improve patients’ prognosis understanding – 1) lecturing style of a conversational structure by maximizing entropy, and 2) the positive language patterns (i.e., sentiment trajectory) using k-means clustering. We used a data set that includes conversations between physicians (N=38) and late-stage cancer patients (N=382). With statistical analysis, we show that physicians who were lecturing their patients and did not vary their positive sentiment had patients with prognosis misunderstanding.

During global pandemics (e.g., COVID-19), when social distancing is recommended, most communication is taking place online. This indicates the need for online communication training programs that can overcome social and global boundaries.
Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Professors M. Ehsan Hoque (advisor), Lenhart Schubert, and Zhen Bai of the Department of Computer Science and Professor Ronald Epstein of the Department of Family Medicine. The data analyzed and study conducted for Chapter 4 was provided by Professor Ronald Roggee of the Department of Psychology. The study presented in Chapter 6 were conducted by Professor Kim Van Orden of the Department of Psychiatry. The data analyzed in Chapter 7 was provided by Professor Ronald Epstein of the Department of Family Medicine. All other work conducted for the dissertation was completed by the student independently. Work presented here was supported by NSF EAGER grant IIS-1543758, NSF Award IIS-1464162, a Google Faculty Research Award (Ehsan Hoque, PI), Microsoft Azure for Research grant, and by a University Research Award from the University of Rochester (Van Orden, PI).
List of Tables

3.1 Study Data: Counts and Prognosis Survey Options ......................... 34
3.2 Average Prognosis Misunderstanding scores in High and Low LECT-UR, Average Sentiment, and Sentiment Synchronicity Groups .......................... 38
3.3 Prognosis misunderstanding % in three sentiment trajectory clusters. ... 42
3.4 Confounder-Adjusted Logit Model ........................................... 43
6.1 Table 3: Accuracy of feedback compared to human judges ............... 77
6.2 Examples of conversation topics and their emotional intensity in Aging and Engaging Program ....................................................... 87
6.3 Conversational skills rating questions and response options. ............. 94
6.4 Participant characteristics. Notes: a1 participant identified as black, 1 as multiracial, and 1 declined to report on race. b7 participants reported being divorced, 1 legally separated, 3 widowed, and 0 as never married. cPROMIS scales are computerized adaptive tests with population normed T-scores with a mean of 50 and standard deviation of 10. dThe full title for the ‘social satisfaction’ scale is ‘satisfaction with social roles and activities.’ eThe full title for the ‘social self-efficacy’ scale is ‘self-efficacy to manage social interactions’. 106
6.5 Average rating of the conversational skills in role-play session. ........ 107
List of Figures

3.1 LISSA real-time conversation and feedback interface .................................. 23
3.2 Feedback generation algorithm using HMM. .................................................. 26
3.3 Sample LISSA feedback summary for two sessions. This user has improved both volume and body movement, while neglecting eye contact ......................... 28
3.4 An example of feedback interface for each conversation phase. Users can receive either positive or negative feedback for each of the four conversational skills cues. For example, a user can receive positive feedback on eye contact and speaking volume, and negative feedback on smiling and content. ........... 30
3.5 Final feedback interface ................................................................................. 31
3.6 Classification of Physician (D) and Patient (P) Turn Lengths over Window as Lecturing and Not-Lecturing a) Regions of Lecturing and Not-Lecturing as P length vs. D length b) example transcript turn lengths ................................. 33
3.7 Finding the Optimal Lecturing Threshold and window size Based on Entropy. a) Contour plot of Entropy, b) Heatmap of Entropy .................................................. 37
3.8 Distribution of Prognosis Misunderstanding Levels in the High and Low LECT-UR Groups ................................................................................................. 38
3.9 Resulting Sentiment Trajectory Clusters for best K=3. ................................. 41
3.10 Logit model weights for predicting whether the prognosis misunderstood. ... 43
4.1 Keeping conversations on track. Multiple types of responses are pre-loaded to help lead users to the desired follow-up questions. ................................. 46
4.2 a) Researchers monitoring the participant from a remote control room. Feedback and dialogues are being generated by the wizards. b) A speed-dating scenario.

4.3 Effect sizes of condition on post-treatment skills ratings after adjusting for correlation with baseline ratings

4.4 System evaluation results. Green bars indicate significantly (p < .05) favorable evaluations, red bars indicate significantly unfavorable evaluations, and grey bars indicate that mean evaluations were not significantly different from the neutral option for that item (p >= .05). Individual responses ranged from -2 (Strongly Disagree) to +2 (Strongly Agree), figure is expanded to highlight differences between means.

4.5 Outline of the functionality of dialogue module

5.1 Teenagers with Autism Spectrum Disorder interacting with LISSA.

5.2 Survey questions and responses

6.1 Aging and Engaging virtual conversational agent

6.2 The overall system. Users initiate a conversation, which is driven by a dialogue controller. Conversation audio and video are processed in a server and feedback is generated. Users receive the feedback one by one and move to the next conversation phase. After four rounds of conversation, users receive final feedback summarizing the previous

6.3 Average ratings from participants in different groups based on interpersonal needs. (1 = strongly agree, 5=strongly disagree)

6.4 Average ratings from participants in different groups based on their diagnosis (1=strongly agree, 5=strongly disagree)
6.5 (a) The trend line of the percentage of participants annotated as positive feedback recipients after each conversation. (b) The average length of the participants’ speeches for each of the small conversation.

6.6 The study design. Participants first have a face-to-face conversation with a rater and then they are randomly assigned a condition - treatment or control. All participants in both conditions complete their online program (Aging and Engaging for treatment, an online pamphlet for control) in eight sessions. After the intervention, all participants have another round of role-play sessions. In both role-play session the rater rates the participant on their conversational skills.

6.7 Change of Composite score from baseline to follow-up.

7.1 A physician conversing with SOPHIE

7.2 Feedback interface of SOPHIE. On the left side, the conversation transcript is shown. On the right (from top to bottom) speech rate, and the number of questions are shown. Turn-taking shows the turn length and at the bottom, the sentiment trajectory of both physician and SOPHIE are shown with the ideal/suggested sentiment trajectory.
Chapter One

Introduction

Face to face communication plays an important role in our life. However, many individuals lack the communication skills which are essential for making the face to face interaction effective. Although training for communication skills exists, they are often expensive, non-repeatable, and non-scalable. The growing trend of using interactive conversational technologies such as SIRI or Alexa can become a potential augmentation to traditional communication skills training. In order to help individuals, technology first needs to understand the requirements of the target outcome, understand the cues of communication, and generate feedback. The feedback needs to be interpretable, meaningful, and generalizable. In this thesis, we discuss how computers can give feedback to help improve communication skills. Specifically, we explore the real-time feedback and post summary feedback on verbal and nonverbal cues. In addition to the generation of feedback, we have explored the application of such feedback among college students, teens with autism, older adults, and oncologists. We present the design of different types of feedback, the algorithms for understanding behaviors and generating feedback, and the application of such feedback among the target population through several user studies.
1.1 Motivation

To understand the unique challenges with computer-mediated communication skills training we present three scenarios. To design effective feedback strategies we first need to identify the unique problems in these scenarios. We then discuss how different feedback strategies would be applicable in these scenarios.

Case 1: Mike, a 21 y/o male computer science major college student, who finds it difficult to engage in a face-to-face conversation with his peers. Although Mike is very good at understanding the taught materials in his class he craves for the group studies where all of his peers share interesting ideas outside the textbook. The nonverbal behaviors such as making eye contact, smiling appropriately appears very confusing when he engages in a conversation. Despite knowing how to adjust these behaviors, Mike believes that the lack of practicing nonverbal communication skills is the key factor in his difficulties. Even though Mike understands the importance of practicing, he feels shy to ask his busy friends to have a conversation to practice his communication skills. The social skills practice sessions offered at the local debating club also very expensive and their focus is mostly on public speaking.

Case 2: Betty is a 73 y/o female, who worked as a receptionist, retired eight years ago. After the death of her husband two years ago, she feels lonely and isolated. She decided to move into a senior living community where she hopes to make new friends. However, after spending three days at her new place she found that she is not making any friends despite being friendly. She realized that when she initiates a conversation she talks about the negative things that happened in her life such as losing her husband. In addition to the negative things, Betty finds it hard to smile back when the other person is making a joke. As a result, Betty feels frustrated and hopes to get some help from someone who would not be judgmental.

Case 3: Dr. Hall is an oncologist practicing for over five years. Every week he sees
more than thirty cancer patients in his outpatient suite and palliative care unit at the local hospital. He knows that he won’t be seeing many of his patients next month due to the severity and progression of their disease. However, more than half of his patients actually do not understand their situation and are willing to go for aggressive treatment options such as chemotherapy, which Dr. Hall knows will not be effective. After seeing so many patients throughout his career, Dr. Hall still feels frustrated when he sees his patients not understanding their prognosis. Dr. Hall feels that he needs communication training. He knows that he could improve his communication skills by practicing with a standardized patient but that could happen once every two weeks.

The above three cases show three key problems.

**Need for communication skills training.**

Cases like Mike, Betty, and Dr. Hall are not unique. There are many individuals who can use help from technology to improve communication skills. Like Mike, one in 59 individuals in the US shows a deficit in nonverbal communication (Baio, 2014; Georgescu et al., 2014). They often fail to make appropriate gestures, eye contact, and smile to make their verbal communication compelling. These deficits are not only common in young adults, but they are also common in older adults. Up to 20 percent of community-dwelling older adults demonstrate difficulties with nonverbal communication (Geurts, Stek, and Comijs, 2016), and yet effective intervention strategies are unavailable to them. Also, the number of older adults is projected to reach 1.5 billion by the year 2050. Due to the communication skills deficit, individuals are having trouble making friends and lose opportunities such as getting a job or engage in a romantic relationship.

Communication skills also play an important role in serious situations such as information transfer between doctor and patient. Due to the weak verbal and nonverbal communication, more than 60% of cancer patients walk out of their doctors’ offices without fully understand-
ing their prognosis. As a result, the late-stage cancer patients, who have the most important and time-critical decisions to make end-up choosing aggressive treatments. Effective doctor-patient communication is important in enabling patients to understand their prognosis and make informed decisions. In addition, good communication skills may facilitate physicians’ understanding of patients’ symptoms, concerns, and treatment wishes. Indeed, patients with advanced cancer have indicated their desire for frank discussions with their oncologists on sensitive topics including making major health care decisions and prognosis. However, doctors often overlook the concerns patients hold, perhaps because only a few physicians have been exposed to in-depth communication skills training.

To improve communication skills we need tools that can augment traditional training but will be readily available and repeatable.

There exist communication skills training programs for individuals with deficits. However, the number of experts is not enough compared to the number of individuals who need help. Also, there are social stigma and costs associated with traditional training. In addition to these limitations, traditional training is not repeatable in high-frequency.

Communication skills training should be customized to the needs of the end-users.

We need to understand the differences in these three cases. Dr. Hall needs to practice a specific skill that is making his patients understand their prognosis. In contrast, Betty and Mike both need to improve their nonverbal skills. In the case of Betty, in addition to the nonverbal skills, she needs feedback on her verbal and spoken language. It should also be noted that a similar type of feedback may not work for all these cases. Mike might be able to process and reflect on real-time feedback but Betty, with her declined cognitive capability, would not be able to focus on real-time feedback and conversation simultaneously. In the
case of Dr. Hall, the feedback mechanism needs to be very specific (such as, “be more positive at the end of your conversation”) and should be grounded by past research and data due to the sensitivity of his work.

1.2 Challenges and Contributions

Computers nowadays can sense many behaviors which include nonverbal and verbal behaviors such as, smile, sentiment, eye gaze, etc. To help individuals with communication skills the computers need to go beyond sensing to generate meaningful and useful feedback. The feedback can be delivered right when the conversation is taking place, in real-time or after the conversation (i.e., post-feedback). While designing feedback for the communication skills program we faced several challenges -

- The raw audio and video signals need to be processed to generate meaningful feedback on verbal and nonverbal behavior.
- The feedback needs to be easy to understand and non-distracting.
- While designing feedback we should consider individual differences and cognitive capability.
- The design decisions should be backed up by data, expert opinions, and user experience.
- The feedback designed to help improve communication skills need to be generalized in real-world scenarios.
- To allow individuals to practice communication skills in a computer-mediated platform, a conversational agent is needed that is capable of conducting naturalistic conversations.

In this thesis, we aim to find answers to the following research questions –
• RQ1: How to design feedback that is easy to understand and useful for improving communication skills.

• RQ2: How can we demonstrate the viability of the feedback.

• RQ3: How computers can conceptualize sensed nonverbal and verbal behavior to generate meaningful and accurate feedback.

• RQ4: Will the feedback help individuals to demonstrate measurable improvements in communication skills that are generalizable in face-to-face conversations?

To address the research questions we made the following contributions –

• **Design of Feedback:** We have designed the real-time and post summary feedback aiming to improve communication skills. The feedback is delivered while having a conversation with a virtual agent and after conversation. The real-time feedback is given on four nonverbal behaviors – eye contact, speaking volume, smile, and body movement. We utilized colored icons for each of the behaviors to give feedback. The post feedback has additional verbal feedback – the sentiment (use positive language) of conversation, lecturing style of conversation, and speech rate. The design of the feedback is informed by a team of experts and focus groups. Our design team includes UX experts, psychologists, gerontologists, and pediatric psychiatrists. We presented the design considerations of the feedback. Additionally, through the analysis of the study interviews, we present a few design guidelines for future feedback designers (Addressing RQ1).

• **Viability of Feedback:** We have conducted studies to test the acceptability and viability of the feedback. We developed an online communication training program – LISSA (see chapter 4), featuring a virtual agent capable of having conversations. This program utilizes both the real-time and post feedback we have designed. Using LISSA,
we first conducted a randomized control study with a wizard of oz controlled environment with 47 undergraduate students. In the study, participants had face-to-face conversations with research assistants in the context of speed-dating. The speed-dating session took place before and after the LISSA intervention. The study results show that the participants who received feedback through LISSA had better ratings on their eye contact, and head nods. Additionally, we found that the feedback was easy to understand and easy to reflect on. To extend our knowledge of the applicability of the feedback We conducted additional studies with nine teenagers with Autism Spectrum Disorder (ASD). From the study, we identified a few design guidelines allowing further improvements. These guidelines include briefing users at the outset about the purpose and limitations of the system, avoiding unrealistic expectations, incorporating positive acknowledgment of behavior change, making the appearance of the virtual agent more realistic and responsive, personalizing the topics better engagement. We next developed a communication training program – Aging and Engaging, for Older adults which utilized the post feedback mechanism. We conducted a pilot study with 25 older adults to test usability and acceptability. From the study, we found that the participants started conversing with longer turns indicating that they were getting comfortable with the virtual agent as the session progressed. We also found that the participants were adjusting their nonverbal and verbal behaviors during subsequent conversational sessions. We then conducted a randomized control study with 18 older adults in the span of 8 week period. This longitudinal study showed that participants were able to retain their training on communication skills in a face-to-face conversation with a real human. This demonstrates the generalizability of the feedback (Addressing RQ2). To find the viability of feedback that is very specific for physicians we developed another program that features a virtual agent as a patient. After a conversation with the virtual patient, the program gives feedback on the pattern of the positive language
usage, how long the user had spoken without giving a chance to ask questions, speech rate, and how many questions were asked. We conducted a study with eight clinicians. We found that the feedback was easy to understand and accurate. However, the participants in the study suggested adding explanations to the terminologies we used in the feedback and allow the user to go over the history of the patient before starting a conversation.

• **Algorithmic Contribution:** Through the studies with college students and older adults, we collected data to generate feedback. We have developed several algorithms to sense and process verbal and nonverbal behaviors. Specifically, we have collected audio and video data from users conversing with a virtual agent and gathered labels from external experts. The labels were given on the appropriate usage of verbal and nonverbal behaviors. To understand the temporal patterns of the nonverbal cues we used a hidden Markov model technique. In section 3.1.4 we discuss the specifics of the algorithm and the trade-off. To generate feedback targeting physicians' communication skills we used a data set of physician-patient conversations Hoerger et al., 2013. To identify the conversation styles we used sentiment analysis with k-means clustering algorithm (section 3.3.3). This analysis identifies the common patterns of positive language usage in conversations that are associated with better prognosis understanding. To detect lecturing related conversational structure, we used entropy maximization and Maximum likelihood estimation algorithms (section 3.3.2). We also present the empirical evidence of the effectiveness of these detected behaviors in terms of prognosis understanding. We found that overall the nonverbal feedback was accurate. We also found that the feedback on lecturing and sentiment style was correlated with effective communication (Addressing RQ3).

• **Applications:** In this thesis, we have shown the applicability of the feedback through
several studies in different contexts and population groups. We have simulated conversations with virtual agents in several contexts, which include, speed-dating, end-of-life communication, and small talk. We have developed independent communication programs around the simulated conversational tools with the feedback strategies. We first developed the LISSA program for college students and teens with ASD. This program focuses on both real-time and post feedback on nonverbal cues. Next, we developed a social skills development program for older adults – Aging and Engaging, which utilized the post feedback strategies acknowledging the cognitive impairment of the older adults. Our third program – SOPHIE, is for training the physicians with the target of improving prognosis understanding. One of the core contributions of this thesis is the diverse set of applications that were built around the feedback strategies. These applications also demonstrate the generalizability of the feedback (Addressing RQ4).

1.3 Thesis Outline

- **Chapter 2 Related Work:**

  In this chapter, we discuss the work on computer-mediated communication skills training. We present how previous research has designed feedback for both real-time and post conversation. Additionally, we discuss the studies conducted in the past with different population group which includes, teenagers with autism, older adults, medical students, and college students. We present works involving virtual agents in training communication skills. Finally, we discuss the potential areas where we find further research is necessary and our contributions in those areas.

- **Chapter 3 Feedback Design For Communication Skills Training:**

  In this chapter, we present the design of our real-time and post feedback. We first present the design of real-time feedback. We show the design guidelines and a speed-
dating study to prove the concept. Through the study, we collected data. We present how we used the data to automate the feedback. Next, we show the design of post feedback. We show the design considerations and implementations of the feedback. Next, we present a task-specific feedback design. Specifically, the task was to improve communication with patients with advanced cancer. We present the data set we have used and the algorithms to detect certain behaviors to give feedback on. Finally, we show empirical evidence of the effectiveness of our feedback design.

• Chapter 4 LISSA – Live Interactive Social Skills Assistance:

In this chapter, we present a virtual agent-based communication skills program LISSA – Live Interactive Social Skills Assistance. This program utilizes the real-time and post feedback strategies we presented in chapter 3. We present a randomized control study with 47 college students in the context of speed-dating. Through the analysis, we show that the LISSA was useful in improving communication skills among college students. We then present the data we collect through the study which we used to train machine learning models to automatically generate feedback.

• Chapter 5 LISSA Autism Study:

In this chapter, we present a study with LISSA where the target population was teenagers with autism spectrum disorder. In this study, we aim to identify the design guidelines for such a communication skills training program for individuals with autism. We present the qualitative results from our study participants’ interviews.

• Chapter 6 Aging and Engaging:

In this chapter, we present a communication skills training program for older adults – Aging and Engaging. This program was specially customized for the older population who have declined cognitive functioning. The program utilizes the feedback mechanism
presented in chapter 3. We present a pilot study with 25 older adults to show the acceptability of such a training program. We then conducted a longitudinal randomized control study with 18 participants over a period of 10 weeks. Through the analysis, we show the generalizability of the feedback.

• Chapter 7 SOPHIE - Standardized Online Patient for Human Interaction Education:

In this chapter, we present a communication skills training program for medical students and oncologists. The target of this program to improve the prognosis understanding of late-stage cancer patients. The program features a virtual agent acting as a standardized patient. The virtual patient converses with users and after the conversation, the program gives feedback that is associated with better prognosis understanding. We present a qualitative analysis of a study we conducted with resident physicians.

• Chapter 8 Limitations and Future Work:

In this chapter, we discuss the limitations of the thesis. Specifically, we discuss the technical limitations, the challenges we haven’t solved, the study limitations, and the generalizability of study findings. We also discuss how we plan to address the limitations in the future.

• Chapter 9 Conclusion:

In conclusion, we summarize our work, highlight the key contributions, and future research directions.
Chapter Two

Related Work

In this chapter, we discuss the background of the related topics. This thesis touches a few distinct topics. Specifically, we discuss the previous works on communication skills development feedback strategies, frameworks, conversational agents, and studies. Additionally, we discuss the applications of the feedback strategies in a different context. Finally, we discuss the areas for our research opportunity.

2.1 Real-time Feedback

In the past, researchers have shown the feasibility of real-time feedback in computer mediated communication skills training. These real-time feedback included, colorful displays (Nojavanasghari, Hughes, and Morency, 2017; Fiorella, Vogel-Walcutt, and Schatz, 2012), change in trainer behaviors (Tanaka, Negoro, et al., 2017; Cassell, Vilhjálmsson, and Bickmore, 2004), text (Terken and Sturm, 2010; Fiorella, Vogel-Walcutt, and Schatz, 2012), auditory, haptic feedback (Alaraj et al., 2015; Hu et al., 2002; Zhao et al., 2015) and game-like rewards (Patel et al., 2013). Although real-time feedback was shown to be effective, many of these work augmented post summary feedback in their system (Faucett, Lee, and Carter, 2017; Liu et al., 2016). Also, the challenge with implementing real-time feedback is, it may disrupt the learning process and distract users from their primary task (Cannon-Bowers, 2001).
2.1.1 Utilization of Verbal and Nonverbal Behaviors

Many work has focused on nonverbal cues and give feedback in real-time. Patel et al. (Patel et al., 2013) detected nonverbal cues which include, talk time, turn-taking cues, pitch, speech rate, and head and shoulder activity. From these nonverbal cues, they identified affiliation (reflecting interpersonal warmth and connection, trust, and rapport) and control (reflecting dominance, influence, and authority) using social signal processing technology. Their target population was clinicians. They used sun/moon and seesaw visual representation as real-time feedback. Specifically, the sun/moon represents the affiliation and seesaw represented the control of the conversation. In a different context, Faucett et al. (Faucett, Lee, and Carter, 2017) developed ReflectLive’s for clinicians to help improve communication skills in video telehealth. They detect speaking contribution, interruption, eye gaze, and face position. Provided real-time feedback on each detected item and summary feedback. In a study with 10 clinicians, the authors showed that participants improved their eye contact and self-awareness. This demonstrates that identifying the behaviors related to the target population is critical before the intervention. The use of a virtual world has also been shown effective in combination with real-time feedback on nonverbal cues. Barmaki et al. (Barmaki, 2016) used a virtual classroom environment for improving teaching. They detected gestures of users and gave real-time feedback using red and green colors on open and closed gestures. In a study with 12 participants, the feedback receiving group showed fewer closed gestures after the intervention. In the context of a job interview, Yu et al. (Yu et al., 2019) developed a dialogue system that can track user engagement. They used facial expression and audio to model the engagement in real-time. Feedback was given through the dialogues of conversation. In a study, the authors found that the group who received feedback improved their speaking skills.

Real-time feedback on verbal and nonverbal behaviors was also proven to be effective in group settings. DiMicco et al. (DiMicco et al., 2007) Used a hardware device in a group
meeting to sense the speaking contribution. Displayed the contribution in real-time using bar graphs, which improved the overall effectiveness of the discussion such as decision making. Terken et al. Terken and Sturm, 2010 used a table-top display for giving real-time feedback on speaking contribution and eye gaze. In a study with 82 participants, 58 improved their speaking contribution and little effect on eye contact.

2.1.2 Virtual Agent-based Interventions

Many work utilized virtual agents to conduct communication training with real-time feedback on verbal and nonverbal behaviors. Foster et al. (Foster et al., 2010) designed ECHOES as a multimodal learning environment intended for children with ASD. The system features a virtual agent that engages a child in a collaborative learning activity and provides feedback based on sensed features including gaze direction and gesture. In a subsequent work with ECHOES, Bernardini et al. (Bernardini, Porayska-Pomsta, and Sampath, 2014) designed a virtual agent as a credible social partner for children with ASD, that engages them in interactive learning activities. Milne et al. (Milne, Powers, and Leibbrandt, 2009) designed a virtual agent as an educational tool for children with ASD, which helps improve their conversational skills and ability to deal with bullying. Feedback was given in real-time through icons and supportive text. In a study with ten participants, the authors showed that participants who received the intervention gained higher conversational skills scores and more knowledge about bullying than the control participants. Hopkins et al. (Hopkins et al., 2011) developed a computer-based social skills training system – FaceSay, for children with ASD. FaceSay contains three different games with realistic avatars designed to teach children specific social skills. The system provided three interactive games promoting the eye gaze, facial expression recognition, and recognizing faces. In a study with 49 individuals, the authors demonstrated that the program helped the participants improve their emotion recognition ability and social interactions. Mower et al. (Mower et al., 2011) developed
Rachel, an embodied conversational agent designed to elicit and analyze naturalistic interactions. This tool was designed for children with autism to encourage their affective and social behavior. In the domain of public speaking Batrinca et al. (Batrinca et al., 2013) developed Cicero and explored the possibility of using interactive virtual audiences for public speaking training. The virtual agent changed their behavior in real-time to give feedback to the speaker. Hubal et al. (Hubal et al., 2000) developed a virtual standardized patient for patient interview practice. Their system incorporated a virtual agent acting as a patient and a human instructor for giving feedback. The feedback included exhibiting medical signs and symptoms with real-time, true-to-life physiological behavior of the agent. Kron et al. (Kron et al., 2017) presented a virtual patient framework (MPathic-VR) to train second-year medical students on how to break the bad news. The agent gave feedback by changing its’ behavior while responding to the students. In a randomized control study (N=421), the authors showed that the students using MPathic-VR achieved significantly higher composite scores on the advanced communication skills objective structured clinical examination (OSCE) than computer-based learning-trained students.

2.1.3 Virtual Environment and Virtual Reality

The virtual environment provides a unique opportunity to provide subtle real-time feedback by manipulating the virtual environment since modification is a low cost. Adery et al. Adery et al., 2018 developed Multimodal Adaptive Social Intervention in Virtual Reality (MASI-VR) for improving social functioning among individuals with schizophrenia. Participants exhibited a significant reduction in overall clinical symptoms, especially negative symptoms following 10 sessions of MASI-VR. The VR allowed the participants to play a game where they received real-time feedback through text. Park et al. (Park et al., 2011) used VR for practicing social skills and compared them with traditional training. In a 10 week-long study, with 91 participants, they show that the VR group had better communication
Ochs et al. Ochs et al., 2019 used virtual patients for training doctors’ communication skills. They compared three methods of training – VR headset, VR environment, and PC simulation. In a study with 22 participants, the authors found that VR environment significantly impacted the display on the perception of the user (sense of presence, sense of co-presence, and perception of the believability of the virtual patient. Hartholt et al. Hartholt, Mozgai, and Rizzo, 2019 presented a versatile system for practicing job interviews with virtual agents. The target population was people with a high level of anxiety. The system provided three modes of conversation-friendly, neutral, and unfriendly. Their system offered visual variation with the virtual environment.

2.1.4 Multiple Modalities

Many researchers have utilized multiple modalities such as visual and audio to deliver feedback effectively. Nojavanasghari et al. (Nojavanasghari, Hughes, and Morency, 2017) designed a system to mediate human-to-human interaction for children with autism. The aim of their work was to design a tool to help improve the social skills of children with autism by providing visual support for the children and real-time feedback to the interactor about the children’s affective states, using a recommendation system. Their system performs real-time social signal recognition and provides visual support on maintaining eye contact while interacting, initiating social interaction, understanding emotions and expressions, turn-taking, and greeting someone. Fiorella et al. (Fiorella, Vogel-Walcutt, and Schatz, 2012) explored two modalities of real-time feedback – visual feedback with spoken and printed text. They experimented in simulation-based training where participants had to make a complex decision-making task. In a study with 60 participants, they used a deployable virtual training environment simulator in which participants played as a military fire support team. The task was to make a decision for ‘call for fire’. Researchers show that the spoken feedback group showed higher order of cognitive skills overall suggesting audio support with
text is effective during real-time feedback. Rock et al. (Rock et al., 2009) used a bug in the ear method to deliver feedback in the context of teaching. Participants include 15 teachers enrolled in a field-based graduate special education teacher preparation program. They collected video-recordings of teacher observations and written reflections by teachers about their experiences. Quantitative results indicate that advanced online bug-in-ear technology is a practical and efficient way to provide immediate feedback to increase teachers’ rate of praise statements and their use of proven effective instructional practices.

2.2 Post Feedback

In addition to real-time feedback, many works have shown success in utilizing post feedback. Post feedback can be very useful where the users need to spend time on the feedback to fully understand. Also, in many cases, the target population, such as older adults, can not focus on real-time feedback (Getzmann, Edward J Golob, and Wascher, 2016). In this section, we discuss the works that utilize post feedback in a different context.

2.2.1 Training Individuals with ASD

Individuals with ASD show various degrees of effectiveness since each individuals’ requirement is different. This is why incorporating post feedback in training systems showed success. Tanaka et al. (Tanaka, Sakti, et al., 2015) used virtual agent-based conversational skills training for individuals with ASD. In the system, they detected pitch, volume amplitude, voice quality, words per minute, pause duration. The system provides summary feedback after a conversation with a virtual agent. The feedback contains user video, overall score, pitch variation, comparison with others in terms of z-score, comments with encouraging words, and areas to improve. In subsequent work, the authors used an embodied virtual agent for conversational training (Tanaka, Negoro, et al., 2017). The agents’ behaviors were modified
in real-time using the behaviors sensed and gave post-feedback on a smile and head pose. In a study with 18 participants with ASD, they showed improvements in nonverbal behavior when rated by external experts. Benton et al. (Benton et al., 2011) presented a methodology for incorporating children with ASD in the design process. They conducted a study with 20 participants with ASD aged between 11 and 14 and, using their design methodology, came up with ten design guidelines specific for game design and idea generation.

### 2.2.2 Healthcare

In healthcare training feedback has shown success. Liu et al. (Liu et al., 2016) developed the EQClinic platform which is a training tool for medical students and can provide summary reports about speaking contribution, volume, and pitch as well as facial expressions, head positioning/nodding, and hand-over-face. In a study with medical students, authors found that reviewing summaries of non-verbal communication behaviors collected by EQClinic improved student’s interview skills. DeVault et al. (DeVault et al., 2014) developed SimSensei - a virtual agent in the context of the healthcare decision support system. The goal of this system is different from 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 uses four classifiers to categorize the users’ speech, and hence to generate a relevant response. Peddle et al. (Peddle et al., 2019) developed a virtual patient (VP) in a VR platform. The aim of the program was to develop knowledge, skills, attitudes and practice of non-technical skills in undergraduate health professionals. The VPs in the program belong to the narrative category portraying a patient’s story as it evolves over time with a cause-and-effect approach to demonstrate the consequences of actions and decisions. In a study with second and third-year nursing students, authors found that interactions with VPs developed knowledge and skills across all categories of non-technical skills to varying degrees. Third-year students suggested
that interactions with VPs helped develop the knowledge and skills in the clinical setting. Web-SP (Zary et al., 2006), a web-based platform, designed to facilitate the development of virtual patients for healthcare education. It provides a general platform to design and develop virtual patients for specific healthcare cases. Angus et al. (Angus et al., 2012) developed a graphical visualization tool to model physician-patient dialogue to identify patterns of engagement between individuals including communication accommodation, engagement, and repetition. Kleinsmith et al. (Kleinsmith et al., 2015) developed a chat-based interactive virtual patient for early-stage medical students for practicing empathy. During the training, students can gather information regarding history of present illness, medical history, family history and social history. Additionally, during each session, the VPs delivered a statement of concern. These statements, termed empathetic opportunities, were designed to elicit an empathetic response from the user. In a study, medical students interacted with VP and standardized patients. The responses of the participants were then rated by coders and found that responses were more empathetic with virtual patients than standardized patients. The level of empathetic response positively correlated with response length. Bond et al. (Bond et al., 2019) used virtual agents to train and generate cases for history taking task among resident physicians. The system gives a score to the physicians after performing the history taking.

2.2.3 Older Adults

Researchers have focused on developing social companions for older adults which often included post summary feedback. Vardoulakis et al. (Vardoulakis et al., 2012), for example, designed a virtual agent companion. They used a Wizard-of-Oz technique to collect data from older adults who were socially isolated. Participants’ self-reported ratings showed that they were willing to use the device again. Ring et al. (Ring et al., 2013) developed a computer program with a virtual assistant that exhibits synchronous non-verbal behavior using
the Behavior Expression Animation Toolkit (BEAT) (Cassell, Vilhjálmsson, and Bickmore, 2004). Their program has two modes: passive and proactive. In proactive mode, the program can initiate a conversation after observing the user with motion sensors. In passive mode, users had to initiate the conversation by themselves. In a one-week study with 14 older adults, the proactive group reported a significantly greater decrease in loneliness than the passive group. Supportive text and audio were used in these systems targeting social isolation among older adults.

2.2.4 Job Interview

Johnson et al. (Johnson et al., 2019) used virtual agent-based training for negotiation tasks in the context of job interviews for students. After virtual negotiation the students receive feedback. The feedback described the extent of good outcomes achieved, and how they followed specific strategies to achieve these outcomes. The feedback also shows how the students were evaluated. The students were then provided specific actionable strategies for improvement in the future.

2.3 Thesis Contribution in Contrast to the Related Work

Feedback is an indispensable part of communication skills training. Feedback can be very subtle or can be very obvious. Both real-time and post feedback has been studied extensively in the past. However, there are still some areas where the feedback strategies were not fully effective. Here we discuss the limitations in the previous work and how we have pursued to solve them.

• Less emphasize on data-driven design. While designing a system that provides feedback, the design and generation of feedback lack the data-driven approach. Many work has focused on expert opinion (Barmaki, 2016), user experience (Benton et al.,
2011) or simply intuition (Tanaka, Sakti, et al., 2015) while designing feedback. In this thesis, we present a data-driven approach for generating feedback.

- **Lack of proof of concept.** In the past work, there is a lack of proof of concept while designing feedback. Past work has ignored testing the viability of their system. While designing the feedback we have focused on the proof of concept first and then applied the gathered knowledge.

- **Subsequent studies proving generalizability** Although past work has shown their effectiveness of feedback by conducting user studies, those studies are limited to artificial/lab settings and does not show generalizability. In this thesis, we have conducted several user studies which support the generalizability of the feedback.

- **Lack of autonomy and naturalistic conversation** In the past while training communication skills, the system did not incorporate a naturalistic environment. For example, while teaching social skills among teens with ASD researchers used virtual agents but the communication was done via text. Often the virtual agents were driven by wizard-of-oz. In this thesis, we discuss how we have designed a conversational agent to make the conversation natural.
Chapter Three

Feedback Design For Communication Skills Training

Our goal is to design feedback that can help improve communication skills. We have designed feedback which is effective while practicing communication skills. In this chapter we discuss how we have designed real-time and post feedback, the data collection, and algorithms for generating the feedback. Our designed interfaces includes a virtual agent, capable of having an open-ended conversation. The design of the feedback is wrapped around the virtual agent based conversation practice.

3.1 Real-time Feedback

Real-time feedback is useful where the user needs to adjust their behavior on the fly. However, real-time feedback can be distracting and put cognitive load on users (Cannon-Bowers, 2001; Getzmann, Edward J Golob, and Wascher, 2016; Kulyk, Wang, and Terken, 2005; Leshed et al., 2009). Which is why while designing real-time feedback we need to make sure that the feedback is intuitive and easy to reflect on.
3.1.1 Design

We have focused on giving real-time feedback on nonverbal behavior. We restricted ourselves to four nonverbal cues. We prioritized behaviors that were important based on prior literature in this feasibility study. We placed four icons at the bottom of the virtual agent interface representing eye contact, smile, volume, and body movement (see Fig. 3.1). At the beginning of the conversation, all icons are green. During the conversation the icons turn red, prompting the user to adjust the corresponding behavior. For example, when the eye icon turns red it means that the user is not making enough eye contact, if the body movement icon turns red it means user need to adjust his body movements (gesture and head nods), and so on. Icons reverts to green after the user makes the appropriate adjustment. This simple two color system is as unambiguous as a traffic light. It allows users to pause and adjust before continuing. Fig. 3.1 shows the interface with virtual agent and real-time feedback.

3.1.2 Viability of Design

We developed a virtual agent based interactive system called LISSA - Live Interactive Social Skills Assistance. LISSA features a virtual agent capable of having a conversation with users and provide both real-time and post feedback. The system was initially designed to be driven by wizard-of-oz. To test the viability of the feedback, we conducted a randomized control
study in the context of speed-dating. In this study, we recruited 47 male college students who were evaluated by 8 Research Assistants (RAs), psychology undergraduate students trained to rate speed-dating conversations. The study was completed through multiple sessions of 3-6 participants that were scheduled at a time. Each session had three parts - speed-dating, intervention, and post intervention speed-dating. At each speed-dating session the RAs rated their conversation using Conversational Skills Rating Scale (CSRS) (Spitzberg, 2014). During the intervention the control group (n=24) was given a readily available pamphlet and a YouTube video on how to improve their conversational skills. The treatment group (n=23) received two training sessions (approximately 10 minutes and 7 minutes) with LISSA with a 2 minutes of break in the middle. The analysis of the RAs’ rating on CSRS scale suggests that the participants in LISSA condition showed improvements on their use of eye contact and nodding of head. In addition to the behavioral improvements, the participants said that the system feedback was easy to understand. The overall system usability score was also higher than the acceptable threshold (Brooke et al., 1996). The details of the speed-dating study can be found in chapter 4: LISSA – Live Interactive Social Skills Assistance.

3.1.3 Data Collection

The intervention received during speed-dating study also served as a 46 video dataset (i.e., two from each of the 23 participants who completed the program) for machine learning to design the feedback modules. We were then able to extract the features from these videos, which included volume and pitch through Praat (Boersma, 2001), smile intensity through SHORE 1, eye gaze direction through OpenFace 2.0 Baltrusaitis et al., 2018, and estimates of body movement through pixel differences between consecutive frames. We extracted these features for each 10 milliseconds of the videos and then normalized the features within each encounter to account for different video recording settings (e.g., distance from microphone).

---

In order to get reliable labels from each video, we then recruited six RAs who were trained to label body movement, volume, smile, and eye contact. They then labeled all the videos using a web interface in which they selected one video at a time and played it while labeling each moment in real time through key presses. To maximize reliability, RAs only labelled two nonverbal cues at a time and then reviewed ratings periodically in a weekly meeting. Given the brief duration of non-verbal behaviors (requiring split second reactions to be seen), RAs were only able to achieve a modest average inter-rater correlation between their labels as a group (Average Inter-rater $r = 0.42$). However, during meetings it was found that if at least two raters labelled a given moment, the group often agreed with their response. Examining the average correlations between any given rater and the remainder of the group (i.e., at least one other rater labeling the same moment) demonstrated more impressive reliability (Average Inter-rater $r = 0.48$). Based on this finding, we classified any moment labelled by at least two raters (out of six) as a “Bad Moment” meriting feedback.

In order to build an automated dialogue manager, we collected all the transcripts from the interactions of participants and LISSA. We used Automatic Speech Recognizer (ASR) and human generated transcript to analyze the degree of error that ASR could make after the automation.

### 3.1.4 Algorithm

As our features were time series data, we chose to use temporal models to train our system. There were four types of feedback and each type has two classes, one for red icon, and one for green icon. We used a Hidden Markov Model to train our system using four features. For each feedback type, we trained a Hidden Markov model (HMM), $M_{(i,j)}$ where $j \in \{\text{red, green}\}$ using the Baum-Welch algorithm. We divided our data set to train and test 10-fold cross validation. During the testing phase the data came in as a video format. We thus divided the video stream in two second chunks and extracted the features. After the
Figure 3.2 Feedback generation algorithm using HMM.

feature extraction a sequence of features vectors were created, where each feature vector $\mathbf{v}$ is four dimensional. In order to generate feedback (choosing a red or green icon) we calculated the observation probability for both models. For example, to obtain the prediction label for smile, we calculated $P(\mathbf{v}_t, \ldots, \mathbf{v}_{(t+\tau)} | M_{i=\text{smile}, j=\text{red}})$ and $P(\mathbf{v}_t, \ldots, \mathbf{v}_{(t+\tau)} | M_{i=\text{smile}, j=\text{green}})$ using the forward-backward recursive algorithm. These two probabilities give the prediction for the icons going red or green. For each of type of feedback we have a threshold $\delta_i$ which is determined from the ROC curve, indicating how much probabilities should defer in order to select a label. Here, $\mathbf{v}_t, \ldots, \mathbf{v}_{(t+\tau)}$ can be overlapping observation sequences. In our model we set $\tau = 250$, which was found the best value in our test set. Fig. 3.2 shows the feedback generation algorithm. In addition to the feedback we designed a dialogue manager for automatically conducting conversations with virtual agent. The dialogue manager is designed to lead a human-like, responsive, autonomous conversation with the user. To lead a dialogue, the virtual agent follows plans and sub-plans that are dynamically customized and modified. Throughout the dialogue, user inputs are mapped to explicit, context independent “gist-clauses”, where the mapping uses the gist-clause representation of the preceding LISSA output as context. The user’s gist-clauses are then used to generate an appropriate reaction to the user’s input in the context of the preceding LISSA output. Both mappings (user inputs to gist-clauses, and gist-clauses to responses) use a flexible, robust hierarchical pattern.
transduction method.

3.2 Post Feedback

In this section, we present the post feedback strategies we have designed targeting different population including young adults with communication difficulties, older adults, and oncologists.

3.2.1 Summary Feedback

We have designed the summary feedback which summarizes the real-time feedback mentioned in 3.1: Real-time Feedback. The LISSA system allows users to access summaries of their real-time feedback after the end of a conversation. We designed an interface where users can see three charts summarizing the session (see Fig. 3.3). The “Reminders” represents the amount of time for which particular icon was red. The “Best streaks” represents the longest time users kept each icon green. Conversely, “Response Lag” represents the longest time it took to fix a particular behavior. Users can look up previous feedback and see improvements. In addition to the charts, we put supportive text to help users interpret the information.

3.2.2 Verbal and Nonverbal Feedback

In addition to the post summary feedback, we have designed post feedback on nonverbal and verbal cues where the real-time feedback is not present. This type of feedback is useful when the target population can not concentrate on real-time feedback due to declined cognitive functioning. Specifically, in this thesis, we have designed the verbal and nonverbal feedback for older adults (>65 years old).

In designing the feedback for older adults, we encountered several challenges. First, we needed to design the feedback that is simple and requires minimal cognitive function to un-
derstand. Second, we needed to generate feedback on the nonverbal and verbal behaviors of the users which can be useful for communication training. Third, feedback needs to be generalizable in real-world. To understand these problems better, we consulted with a focus group of 12 older adults, each of whom volunteers as companions for isolated peers. They have valuable experience with older adults who are at an elevated risk of having nonverbal behavioral difficulties. We described our concept and showed them a low-fidelity prototype. The focus group offered thoughtful advice on the virtual assistant’s appearance, feedback, and dialogue topics. In summary, the focus group made the following suggestions: the avatar should look like an older adult; the feedback needs to be less distracting (i.e., avoid real-time feedback and have feedback appear in sequence, not all at once); avoid too much negative feedback, as it might disengage users; and reinforce positives and demonstrate understanding. We then consulted with the expert advisory panel. This panel consists of a geriatrician, a geriatric neuropsychologist, and a gerontologist with expertise in interventions. The advisory panel made the following suggestions: include short assessment points throughout the intervention in which the virtual assistant can give participants feedback directly—as opposed to real-time feedback, delivered through text and images—and also foster transi-
tions; and the avatar’s face should be highly contrasted against the background so that even visually-impaired participants will have no difficulty seeing it. We tried to incorporate all suggestions.

While designing the feedback we divided the conversation into four phases to reduce cognitive load on users. Our system gives feedback after each phase. Given that increasing difficulty with divided attention is common with cognitive aging (Getzmann, Edward J. Golob, and Wascher, 2016), many older adults may find it difficult to focus on a conversation and feedback simultaneously (Kulyk, Wang, and Terken, 2006; Leshed et al., 2009). This may be especially true of older adults with social functioning impairments, like those for whom our program is designed (Cacioppo and Hawkley, 2009). Additionally, feedback across four stages allows users to adjust their behavior in successive conversations. Afterward, our system summarizes the feedback and provides recommendations for future practice as well as areas to emphasize later (e.g., “vary pitch while explaining something”).

We chose four different dimensions of conversational skills to give feedback: eye contact, volume, smile, and speech content. The attendant behaviors (smiling, volume modulation, making eye contact, etc.) have been shown to foster communication (Bowie et al., 2008; Hames, Hagan, and Joiner, 2013; Ali et al., 2015). The feedback was given using pictures, text, and voice. Fig. 3.4 shows our feedback interface. The top row shows an example of positive feedback; the bottom row, negative feedback. The feedback interface contains a picture and some text. We added text feedback because feedback offered through multiple channels is more likely to be effective than that offered through only one channel. We selected this approach in response to the expert panel and the focus group. For each skill dimension, our system offers positive and negative feedback. If the user makes eye contact with the avatar, for example, they receive positive feedback on eye contact. Each instance of positive feedback is accompanied by an animated green checkmark and a “ding” sound. The feedback text is read aloud to the user using a text-to-speech engine. To avoid repetition, the negative
Figure 3.4 An example of feedback interface for each conversation phase. Users can receive either positive or negative feedback for each of the four conversational skills cues. For example, a user can receive positive feedback on eye contact and speaking volume, and negative feedback on smiling and content.

Feedback text changes across each conversation. For example, if a user does not smile at all during the first conversation, the system will say, “You didn’t smile at all. Try smiling more, including with your eyes.” If the user does not improve in the second conversation, the system will say, “Like I mentioned before, try smiling more.” After all phases of the conversation, we generate an integrated feedback. This final feedback summarizes the feedback the user received after each phase. Fig. 3.5 shows the final feedback interface. The final feedback shows, at most, two dimensions in which the user received negative feedback. We decided to show only two cases, as too much negative feedback can lead users to disengage. The system suggests ways for users to improve. For example, if the user receives negative feedback on speech content, our system will say: “Practice casual conversation with people you encounter during your day, and keep the conversation focused on positive topics.”

To generate feedback from video and audio, the facial and prosodic features, including
smile intensity, volume, and eye gaze direction are extracted. A hidden Markov model–based classifier then classifies the patterns of nonverbal features into two categories: positive and negative (see section 3.1). We generate feedback based on the classified temporal patterns. On the audio we perform automated speech recognition, which we later use to perform a sentiment analysis (ratio of positive to negative words, for instance). After each phase of conversation, the transcript is uploaded to the server, where we look for negative and positive words in the transcript from a prepopulated list. The list of words was generated with direct input from two clinical psychologists who provide regular therapy to elderly patients.

We conducted two separate studies with the post feedback strategies. First study was conducted to test the applicability of the feedback and the second one was to test the generalizability. In order to conduct the studies we developed a communication skills training program for older adults – Aging and Engaging. Chapter 6: Aging and Engaging has the details of the studies.
3.3 Task Specific Feedback

In this section, we discuss generation of feedback for improving communication skills of oncologists. Specifically, we focus on help individuals with cancer to understand their prognosis. In patient-physician communication there are several behavioral paradigms that help improve information transfer. Since our target is very specific (i.e., improving prognosis understanding) the feedback design needs to be task specific. To do so, we need to identify the specific behaviors on which we can give feedback. The behaviors in doctor-patient communication can be categorized into two groups – behaviors to avoid and behaviors to cultivate. Among many behavioral paradigms, we have explored two patterns of behavior – *lecturing*, and the *sentiment trajectory* of conversation. We first present how we set about detecting these phenomena automatically and determining how they are associated with prognosis understanding. Then we explain our feedback design for these two behavioral patterns, applicable in conversation practice with a virtual conversational agent.

**Lecturing:** Lecturing occurs when the physician delivers a lot of information without giving the patient a chance to ask questions or to respond (Siminoff et al., 2000). In order to detect a lecturing event, we calculate the turn lengths (i.e., number of words spoken by the physician or the patient) in a fixed-size sliding window. When the number of words spoken by the physician is greater than a threshold, we classify it as a lecturing event. Fig. 3.6 shows the area where the lecturing event can occur. The thresholds are determined by maximizing the entropy of the outcome variables (i.e., prognosis misunderstanding). In this section, we present a method of detecting lecturing in dialogues. Later we present the relationship between the lecturing and prognosis misunderstanding.

**Sentiment Trajectory:** In the field of natural language processing, *sentiment* represents the classification of emotion in text data. In this work, we focus on positive language usage. We define sentiment trajectory as the change that occurs in physician positive sen-
timent over the course of the conversation. Findings from communication research suggest that the trajectory of affective communication features (e.g., sentiment) may be particularly important (ali2018and). However, the physician sentiment trajectory over a conversation has not been well-studied in the context of patients’ prognosis understanding. First we describe how we have defined and identified effective sentiment trajectories. Later we present the association between the trajectories and prognosis understanding.

3.3.1 Data Collection

We performed a post-hoc analysis of a study (Hoerger et al., 2013) involving 382 visits between cancer patients ($N = 382$) and their oncologists ($N = 38$). The data includes a transcript of the conversation, in addition to both patient and physician surveys associated with each visit. The survey also included questions to the physician and to the patient regarding the patient’s prognosis. The prognosis questions were a modified version of the SUPPORT prognosis measure (Weeks et al., 1998). Specifically, the prognosis question directed to the physicians was: "What do you believe are the chances that this patient will
Table 3.1 Study Data: Counts and Prognosis Survey Options

<table>
<thead>
<tr>
<th>Resp. #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>1</td>
<td>about 90%</td>
</tr>
<tr>
<td>2</td>
<td>about 75%</td>
</tr>
<tr>
<td>3</td>
<td>about 50-50</td>
</tr>
<tr>
<td>4</td>
<td>about 25%</td>
</tr>
<tr>
<td>5</td>
<td>about 10%</td>
</tr>
<tr>
<td>6</td>
<td>0%</td>
</tr>
<tr>
<td>X</td>
<td>don’t know</td>
</tr>
<tr>
<td>X</td>
<td>refuse to answer</td>
</tr>
</tbody>
</table>

*live for 2 years or more?*; the options provided for a response are shown in Table 3.1.

Patients were separately asked "What do you believe your doctor thinks are the chances that you will live for 2 years or more?", with the same options for a response. By comparing patient and physician responses, we derived a misunderstanding percentage. More specifically, when the absolute difference of the responses is greater than 1, the patient-physician prognostic understanding is defined as being misunderstood. Data in which either the physician or patient refused to answer were not used. The transcribed visits each involved a regularly scheduled visit between a late-stage (stage 3 or 4) cancer patient and their oncologist. Many of the visits included a family caregiver and/or other health care staff (e.g., nurse, second physician).

### 3.3.2 Lecturing

In this section, we describe our automated algorithm for calculating the LECT-UR Score (Lecturing Estimation through Counting Turns with an Unbalanced-length Ratio), a mea-
sure of lecturing-related conversational structure. The LECT-UR Score is based on Back, et al. Back et al., 2005’s definition of lecturing (i.e., when a Patient-physician transcript shows turns when the "physician delivers large chunks of information without giving the patient a chance to respond or ask questions") Back et al., 2005; Siminoff et al., 2000. The LECT-UR scoring technique was not "trained" on a set of subjective, manually labeled instances of human perceived "lecturing", but rather was designed as an objective algorithm to measure lecturing-related conversational structure, i.e., how often the physician turns were disproportionately longer than his/her patient’s turns over a window of conversational turns.

**Lecturing Detection (LECT-UR Score)**

As shown in Fig. 3.6 region 1, when both the physician and patient speak with short duration over a window of conversation, it is not counted as an instance of lecturing. Similarly, if, as shown in region 3, the patient is speaking with a long average turn length over the window, it is not labeled as lecturing. Only when the physician’s average turn length exceeds a threshold level over the patient’s average turn length, shown in Region 2, is the window labeled as an instance of lecturing.

This algorithm is expressed in the following equations:

\[
L = \sum_{k} \left( \sum_{i=kW \in D} \omega_i - \tau \right) \times \left( \tau - \sum_{i=kW \in P} \omega_i \right)
\]

\[
\mathbb{I}(x) = \begin{cases} 
0 : x < 0 \\
1 : x \geq 0 
\end{cases}
\]

where

- \(L\) : LECT-UR Score
- \(W\) : window length in number of turns
- \(\tau\) : turn length disparity threshold
Referring to equation 3.1, a value for the $\tau$ parameter must be determined. As $\tau$ approaches zero, the area of region 1 in Fig. 3.6 will also approach zero. Alternatively, if a very large value is used for $\tau$, every window will be classified as *not lecturing* since region 1 will cover the entire data space. In order to be useful, the LECT-UR score should have variability (i.e. if all data points have the same LECT-UR score, we will not learn much). Borrowing concepts from information theory, the amount of *information* in a signal can be measured by the signal’s *entropy*, where entropy is a measure of the amount of disorder or uncertainty Shannon, 2001. More specifically, for a given data set $X$, the definition of the entropy, $H(X)$, is:

$$H(X) = \sum_{i=1}^{n} P(x_i) \log_b \frac{1}{P(x_i)}$$

(3.2)

where $P(x_i)$ represents the probability of observing the $i^{th}$ data point. As the probability of an event $x_i$ approaches certainty (i.e. $P(x_i) \approx 1$), the information content approaches zero. Similarly, as the probability of an event $x_i$ approaches zero, the contribution of such events to the total information content in the data set approaches zero. Thus, in order to maximize the information contained in the LECT-UR score, the scores should be well distributed (i.e. maximizing the entropy).

In order to determine the optimal $\tau$ and $W$, we perform a grid search. For a given $\tau$ and $W$ we first calculate the LECT-UR score $L$ based on equation 3.1. We then applied the kernel density estimation method Parzen, 1962 to compute the probability density function $P(x)$. From the probability density function we then obtain the entropy of $L$ using equation

$\omega$ : words in the transcript
D : Doctor utterances
P : Patient utterances
Figure 3.7 Finding the Optimal Lecturing Threshold and window size Based on Entropy. a) Contour plot of Entropy, b) Heatmap of Entropy

3.2. In Figs. 3.7b and 3.7a, the entropy values for different values of $\tau$ and $W$ are shown. The maximal entropy occurs with $\tau = 103$ and $W = 20$. After calculating the LECT-UR score with the optimal parameters for each office visit transcript, we partition the data into high and low LECT-UR groups based on the median value. We then use the Z-score two-tailed population proportion test to see the difference in the percentage of prognosis misunderstanding.

Association between LECT-UR Score and Prognosis Understanding

As shown in Table 3.2, the High LECT-UR Score group has a larger percentage of prognosis misunderstanding than the Low LECT-UR Score group (83.6 vs. 72.3) with a corresponding p-value of 0.00058 and an estimated Cliff’s d effect size of 0.37 Cliff, 1993. Shown in Fig. 3.8 are the distributions of the Prognosis Misunderstanding for the high and low LECT-UR groups. Note that we have calculated the misunderstanding percentage based on the absolute difference between the survey question response. Here a difference of 5 and 6 were treated as misunderstood prognosis. In the high LECT-UR group, over 50% of the patients had a prognosis misunderstanding of 5 or 6 levels. A level of 5 or 6 represent the situations in which
there is a 90\% or more difference between the patient’s understanding of their physician’s two year survival estimate and their physician’s actual two-year survival estimate.

### 3.3.3 Sentiment Trajectory

To investigate the relevance of speaking with positive sentiment as part of an automated system, we utilized the VADER (Valence Aware Dictionary for sEntiment Reasoning) automatic text analysis tool (Hutto and Gilbert, 2014). VADER calculates sentiment through the use of a rule-based model that employs a sentiment lexicon (dictionary of words containing an associated valence measure). The sentiment lexicon used by VADER was produced from a human-labeled corpus in which humans rated sentiment in terms of the overall positive, neutral, or negative emotion associated with a given word in a phrase or sentence. The VADER positive sentiment feature is the result of a large number of human raters’ understanding of
positive and negative emotion associated with particular words. The VADER positive sentiment score was evaluated for each turn of the conversation. These physician and patient sentiment scores were used in two ways — 1) average analysis, and 2) sentiment trajectory.

In average sentiment analysis, the average sentiment scores for the physician were calculated for each transcript. The transcripts were split into two groups based on the median of the physician average sentiments (i.e. a High Sentiment group and a Low Sentiment group). The outcome measure (Prognosis Misunderstanding%) is then compared between the two groups using the z-score population proportion test.

For the second way of using physician and patient sentiment scores, we defined the sentiment trajectory as the time series of average physician positive sentiment over the segmented conversation. More specifically, we partitioned each conversation transcript into a number of non-overlapping segments (each segment having the same number of conversational turns) and calculated the physician’s average positive sentiment within each segment. Each conversation’s sentiment trajectory is represented as a multidimensional vector, each dimension corresponding to the average sentiment within a corresponding segment of the conversation.

**Sentiment Trajectory Detection**

We next determined whether distinct styles of physician sentiment trajectory existed among the conversations and investigated whether any of these physician styles demonstrated significant differences in any of the indicators of communication effectiveness. To determine whether distinct styles of sentiment trajectory exist in the physician sentiment among the transcripts, we applied the k-means clustering algorithm (Lloyd, 1982). The k-means algorithm groups the conversation trajectories into a number (k) of clusters (or groups) of trajectories based on their relative Euclidean distance. The number of clusters k was selected using the widely used Silhouette method (Rousseeuw, 1987), in which a grid search over a finite space of integer values for the k parameter is performed in order to find the
number of clusters which maximizes the Silhouette score (i.e. a combined measure of cohesion among data points within a given cluster and separation of data points among different clusters). In order to determine whether any of the resulting sentiment trajectory clusters had statistically significant differences in the outcome measures, we applied the inference test for population proportions pairwise between the groups.

To understand the effects of the demographic and confounding variables we performed a logistic regression analysis. Specifically, we applied logistic regression on gender, age, disease severity, average sentiment of the conversation, study site, study arm, and the conversation styles to predict the outcome measures. In analyzing confounding variables, there are mainly two approaches: 1) stratification and 2) multivariate methods (i.e., logistic regression aka logit). Since we have multiple potentially confounding variables, we used the multivariate method of logistic regression. Stratification would likely be problematic since we have a small sample size. Our outcome variable is binary (either you understand or don’t understand your prognosis), therefore instead of linear regression we use logistic regression, which estimates the probability of getting an outcome as a linear function of all of the input variables (including confounding variables as well as cluster membership).

After fitting data to logistic regression, we can compare the relative effect that each of the input variables has on predicting whether a given data point (conversation) results in a "Don’t understand prognosis" classification. After normalizing the inputs (i.e., scaling and shifting to have mean=0 and variance=1) we fit the model (using the hyper-parameter that provides the highest data likelihood) and hence find the model weights. We then investigate the weights of the logistic models and the prognosis misunderstanding percentage for each of the conversation styles.
Figure 3.9 Resulting Sentiment Trajectory Clusters for best K=3.

Association between Sentiment and Prognosis Understanding

The difference in the prognosis misunderstanding % between the high and low average positive sentiment groups did not show a significant difference. Out of the analyzed number of clusters (k = [2, 10]), the number of trajectory clusters that had the highest silhouette score was k=3. In addition, the BIC (Bayesian information criterion [schwarz1978estimating]) analysis also found that the optimal value for k is 3. Shown in Fig. 3.9 are the resulting three trajectory clusters: cluster A (red, n = 15); cluster B (orange, n = 58), and cluster C (blue, n = 191). It should be noted that the K-means clustering algorithm does not inherently attempt to produce clusters of equal sizes, but rather finds clustering groupings which minimize the within-group variation. Cluster A (Dynamic) is represented by a more dynamic shape with increases in positive sentiment at 25% into the conversation (segment 2), as well as at the end of the conversation (segment 7). In contrast, Clusters B (Medium) and C (low) have a mostly flat positive sentiment level throughout the conversation with approximate average VADER sentiment levels of 0.1 and 0.05 respectively.

Shown in Table 3.3 are the outcome measures for each of the three trajectory cluster groups along with pair-wise population percentage inference test p-values. As shown, the Prognosis Misunderstanding %, the low cluster (cluster C) showed the highest percentage
Table 3.3 Prognosis misunderstanding % in three sentiment trajectory clusters.

<table>
<thead>
<tr>
<th>Trajectory Group (size)</th>
<th>Pairwise statistical comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (15)</td>
<td>P_{AB} 0.34</td>
</tr>
<tr>
<td>B (58)</td>
<td>P_{BC} 0.04</td>
</tr>
<tr>
<td>C (191)</td>
<td>P_{AC} 0.06</td>
</tr>
</tbody>
</table>

with 67.9 % of the patients having a discordant understanding of their prognosis. The p-values for comparing the percentages between low and dynamic and low with medium clusters were 0.04 and 0.06 respectively.

Fig. 3.10 shows the logistic regression weights when predicting the Prognosis Misunderstanding %. The variables marked with a (*) had p<0.05. The more positive weights indicates higher chances of the particular outcome. In fig. 3.10 the higher positive value was assigned to severity and age by the trained model. In Fig. 3.10 the highest positive value was assigned to severity. This indicates that patients with higher severity level of disease are more likely to misunderstand their prognosis. Among all the clusters the dynamic cluster has the lowers (negative) value. This indicates that when physicians used the dynamic sentiment throughout the conversation the patients were less likely to misunderstand their prognosis. It should be noted that in both outcome measures the effect of average positive sentiment variable is very small.

Unlike the linear regression, with logistic regression there is not a simple way to adjust the output (i.e. “correct” the output) for the effect of confounding variables of each data point. This is because the actual outputs are binary, whereas the model output is a probability. Instead, we can compare the predicted model Prognosis Misunderstanding % for each cluster. When all confounding variables are set to have the average value over our data set, we compute the models’ predicted Prognosis Misunderstanding % for each cluster (see table 3.4).

The Wald test p-value of the logistic regression is also shown in table 3.4 (marked *
Figure 3.10 Logit model weights for predicting whether the prognosis misunderstood.

in Fig. 3.10 when \( p < 0.05 \). This again indicates that with confounding adjustment the dynamic style cluster has Prognosis Misunderstanding % among other clusters.

Table 3.4 Confounder-Adjusted Logit Model

<table>
<thead>
<tr>
<th>Group</th>
<th>PMU %</th>
<th>( \beta )</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>49.76</td>
<td>-0.294</td>
<td>0.033</td>
</tr>
<tr>
<td>B</td>
<td>70.74</td>
<td>-0.155</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>84.85</td>
<td>0.209</td>
<td>-</td>
</tr>
</tbody>
</table>
Chapter Four

LISSA – Live Interactive Social Skills Assistance

In this chapter, we present a virtual agent-based interface for communication skills development. We discuss how we have tested the real-time and post feedback strategies using the training program. We then discuss a randomized control study in the context of speed-dating to validate our feedback design.

4.1 Design of Interface

The LISSA - Live Interactive Social Skills Assistance, is an online virtual agent-based communication skills training program which focuses on real-time and post feedback on nonverbal behaviors. Users can have a conversation with a virtual agent and receive real-time feedback on four nonverbal behaviors – eye contact, speaking volume, smile, and head movement. After the conversation, the system shows post-feedback. Users can compare their performance with their past sessions.

We developed our system on a web-based platform. It allowed human wizards to control the dialogue and the feedback using a separate machine using TCP/IP protocol. Our system provides feedback in two different formats – real-time feedback and post-session feedback
(see 3.1). The user visits the LISSA webpage (available at http://tinyurl.com/lissadate) and presses the Start button in order to start a conversation. While having the conversation the system gives real-time feedback. At the end of the conversation, the user receives summary feedback. Instead of practicing social skills in front of a mirror, people may prefer speaking with a real person. This is why our system features a human-like virtual agent. We used an avatar from “Site Pal”\(^1\). The background and appearance were chosen to give an impression of a college student (see fig. 3.1).

### 4.2 Wizard-of-Oz Prototype

Our primary goal was to determine the effectiveness of real-time feedback in a conversational scenario. We used a Wizard-of-Oz design to control real-time feedback and dialogue. This allowed us to see the effect of our system on the users, highlight important points for refinement, and collect data for automation without automating all components. The system used two wizards, the dialogue operator and the feedback coach, responsible for dialogue and feedback respectively. Users were kept unaware of the wizards. However, the wizards could see and hear the user in real-time through a camera to perform their roles responsively. The Operator managed the virtual agent through a dedicated control panel. The controller has four types of input to control the virtual agent’s expressions, generate a core script of the avatar, generate quick replies, and a chatbox for free text. Conducting an open-ended, meaningful conversation with a user remains an open problem in computer science. To circumvent this problem, we opted to use a scripted strategy. The conversation scripts were designed to lead the user with questions open-ended enough to elicit long responses where users can practice skills, but specific enough to keep users on track. The system had two small-talk topics (Living in the city and Movies/TV) and two deeper topics (Dream Home

\(^1\)https://www.sitepal.com/
Keeping conversations on track. Multiple types of responses are pre-loaded to help lead users to the desired follow-up questions.

and Career plan). Keeping the conversation on the predefined topics helped to reduce the dialogue operator’s time typing responses (improving user experience). However, a user can respond in different ways even to the most leading stems (e.g., ‘Have you seen ‘Guardians of the Galaxy’?’). Therefore, the script included multiple options for keeping users on the topic (Fig. 4.1). Most deviations from the ideal course can be corrected with more flexible versions of the question or simple redirects. But irrelevant responses (e.g., “Are you into superheroes?”) may take the conversation far off course, forcing the Operator to type. In that case, quick replies from the evasive answer group can help the Operator regain control of the conversation. Feedback coaches providing real-time feedback were instructed in basic principles of behavior modification in a 90-minute session with the author and met once a month over the course of the study to maintain reliability. By turning each toggle button on and off, the coach was able to activate the corresponding icons on the user’s interface.
Figure 4.2 a) Researchers monitoring the participant from a remote control room. Feedback and dialogues are being generated by the wizards. b) A speed-dating scenario.

4.3 Speed-dating Study

To evaluate our system’s effect, we conducted a randomized control study in the context of speed-dating study. While other social skills, such as interview skills, can be readily assessed in a highly structured mock interview, how we behave in casual conversation is dependent on a wide variety of factors (e.g., topic choice, partner factors, speaker interest/motivation). To control these factors through a standardized conversation, would be highly artificial and may not generalize to other contexts. By having participants engage in serial 4-min interactions, speed-dating enables measurement of social skills in a scenario with high external validity while removing excess variability when ratings are averaged across all interactions. We removed additional variability by having participants engage in two rounds of speed-dating with the same partners, yielding both a baseline rating and meaningful estimates of change. Between these two rounds, participants were assigned to one of two 20-min skills training conditions – a control condition where they would access readily available self-help methods for social skills, and a LISSA condition where they engaged in two brief sessions with the system. While all participants would be expected to improve due to practice effects
and increased comfort, this design captures relative increases in performance in the LISSA condition.

4.3.1 Participants

For this study, we limited our recruitment to male undergraduate participants. Past research has suggested that men are less skilled at both using and decoding non-verbal behavior (Hall, 1990) and receive less benefit from computerized coaching systems (Hoque et al., 2013). Thus, limiting recruitment to men represents the most stringent test of LISSA’s effectiveness. We recruited 47 male undergraduate students from the University of Rochester. All participants were age 18 or older and native English speakers. Our staff included 8 female research assistants (RAs), also students from the University of Rochester. There were 3-5 RAs at each session to serve as speed-dating partners and raters.

4.3.2 Study Design

The study was completed through multiple sessions of 3 to 6 participants that were scheduled on weekend afternoons. Each session was divided into 3 parts. In the first round of speed-dating, participants went through a series of interactions with RAs. To reduce participant discomfort and improve reliability, RAs were trained to be encouraging, friendly, and inquisitive across interactions regardless of the participant’s performance. Each interaction lasted 4 minutes, after which each male participant rotated to a new table for the next interaction. Before beginning the next conversation, participants and RAs were both given one minute to rate their previous conversation, with RAs rating variables of interest while speed-daters completed a number of distraction questionnaires. This design (4-min interactions and 1-min rating), is sufficient to capture meaningful, robust differences in romantic and nonromantic attraction (Finkel, Eastwick, and Matthews, 2007).
After this initial speed-dating round, participants were taken to another room to be separated into two random groups. Participants assigned to the control condition (n = 24) were given time to watch a YouTube video ² and read free articles from a social skills book (Wendler, 2014). Participants assigned to the LISSA Condition (n = 23) had two skills-training sessions (roughly 10 and 7 minutes in length) with LISSA. After this, participants were then able to give feedback about their experiences. The third section of the study involved a second round of speed-dating in the same format as the first session. RAs were kept blind to participant conditions and were additionally instructed to avoid asking any questions about participants’ skills training.

4.3.3 Analysis

Conversational Skills

To test whether the participants improved their communication skills, we used the Conversational Skills Rating Scale (CSRS) (Spitzberg, 2014). All items were presented on a six-level Likert scale with higher scores indicating more effective use of that skill. We selected 5 item stems from the CSRS to represent behavioral targets of LISSA. Specifically, “Use of eye contact”, “Smiling”, “Vocal confidence”, “Use of gestures”, and “Nodding of the head in response to partner statements”. These items showed strong internal consistency in RA evaluations ($\alpha = 0.87$).

System Evaluation

After completing two interactions with LISSA, participants in the LISSA Condition completed a questionnaire evaluating their interaction with LISSA using a 5 point scale ranging from -2 (Strongly Disagree) to +2 (Strongly Agree). Participants additionally completed

²https://www.voicetube.com/videos/40554
the System Usability Scale (SUS) (Brooke et al., 1996), a commonly used 10 item scale that ranges from 0 to 100 (with higher scores indicating a more usable system).

4.3.4 Results

Effectiveness of LISSA

RA ratings for each individual CSRS item were averaged across all interactions in the first and second round of speed-dating for each participant to create a baseline and post-treatment estimate. These 5 item estimates were then averaged together to create an overall skill. Encounters were excluded from analysis if RAs reported having met the participants prior to the study or if any possibly biasing events occurred (e.g., the conversation went too long; participant revealed condition to the rater; participant behaved inappropriately). A series of t-tests on participant pre-session self-ratings and baseline RA ratings failed to reveal any significant differences between the two conditions, suggesting reasonable baseline equivalence. As baseline and post-session ratings were strongly correlated (r[46] = 0.74; p < 0.01), the effect of the intervention on post-treatment RA skill ratings was assessed using analysis of covariance (ANCOVA), adjusting for baseline ratings of that skill. Using this ANCOVA approach would thus provide an estimate of the impact of LISSA when accounting for initial differences in participants’ social skills (Dimitrov and Rumrill Jr, 2003; Dugard and Todman, 1995). Analysis of the 5-item composite indicated that RAs assigned individuals in the LISSA condition only marginally higher post-treatment ratings after adjusting for initial ratings (F[1,44] = 3.28; p = 0.08). Fig. 4.3 shows the adjusted differences between the LISSA and Control condition for each skill as a Cohen’s d, which scales each different in terms of the pooled standard deviation for that group. All effects were in the expected direction, with participants in the LISSA Condition demonstrating gains in all targeted skills in comparison to individuals randomized into the Control Condition. In particular,
participants in LISSA were rated as significantly higher with respect to nodding in response to partner statements (F[1,44] = 5.49; p = 0.02). Two other difference emerged on the skill of gesturing to emphasize what is being said (F[1,44] = 3.34; p = 0.08) and effective use of eye contact (F[1,44] = 3.89; p = 0.06), which, while marginally significant, were modest in effect size. Taken together, these effects suggest a promising efficacy of the LISSA system.

**Evaluation of Real-time Feedback**

As our evaluation questionnaire included separate items each assessing a unique feature of LISSA, they were evaluated separately. Responses to each item were evaluated using a one-sample t-test against a null hypothesis of 0 on the -2 to +2 scale (the “Neutral” option). Summaries for each can be found on Fig. 4.4. Users tended to find LISSA’s feedback useful (t[21] = 2.53; p = 0.02), accurate (t[20] = 2.75; p = 0.01), and consistent with feedback from others (t[21] = 0.86; p = 0.02). In addition, not only did users tend to deny finding the feedback difficult to interpret (t[21] = -3.15; p < 0.01) or distracting (t[21] = 3.73; p < 0.01), many open-ended responses actually highlighted the “immediacy” and “nearly instant” responses of the signals as strong advantages of the system.
Figure 4.4 System evaluation results. Green bars indicate significantly ($p < .05$) favorable evaluations, red bars indicate significantly unfavorable evaluations, and grey bars indicate that mean evaluations were not significantly different from the neutral option for that item ($p >= .05$). Individual responses ranged from -2 (Strongly Disagree) to +2 (Strongly Agree), figure is expanded to highlight differences between means.

Evaluation of LISSA’s Conversation

Participants tended to find the conversation with LISSA very unrealistic, ($t[21] = -3.81; p < 0.01$). Open-ended criticism of the conversation seemed to center on the lack of varied, relevant responses to the unique material brought in by the user: “LISSA could be improved by having more fluid natural responses that are less monotonous.” “She should try and comment more on the content of what I say, and by that, I mean that she should ‘say’ the same things I just said, as a form of acknowledgment.” “The conversation didn’t feel like a conversation. It was far too scripted to feel realistic.” However, it is notable that users did not seem to have any concerns about the format of the conversation (with LISSA mainly asking questions and the respondent mainly answering them) and did not feel “put on the spot” by LISSA’s questions. Deepening LISSA’s responsiveness while maintaining the general structure of LISSA as an inquirer may greatly improve participants’ experiences. Taken together, the strength of the instant feedback system combined with the rigidity of the conversation script left participants with varied feelings about the overall usefulness and usability of LISSA. Of all the LISSA
evaluation questions, participants had the widest variety of responses to the question about their interest in using the system in the future (SD = 1.22). Similarly, the average SUS ratings of participants was 69 (SD = 10), suggesting that LISSA was at an average level of usability. In addition to a wide disagreement between participants’ evaluations, many participants’ ratings seemed to represent mixed feelings about the program’s strengths and weaknesses, with one participant writing: “It was a good way to practice body language. I think if the actual conversational abilities were more realistic then it would be greatly improved.”

4.4 Dialogue Module

To elicit the behavior a user might display in front of a human conversational partner, our system must simulate a real conversation. In many dialogue systems, the content of the dialogue is central to the purpose of the system; in LISSA, what is said is unimportant as long as the user acts as if he or she is participating in a real conversation. Where a user and an agent collaborate towards a shared goal in a discourse, a system can often afford to be rigid in its adherence to a preconceived linear or branching plan. LISSA’s relative openness in this respect poses a challenge.

4.4.1 Dialogue Schemas

The general strategy for managing the dialogue is through a structure called a schema. A schema contains the knowledge needed by LISSA to carry out a high-level task in the conversation, such as making small talk about general topics, asking about college courses the (presumed) fellow student has taken or discussing problems the user might be having at school. It consists of two kinds of knowledge structures: the dialogue plan, a dynamically modifiable outline of the events LISSA expects to occur in the dialogue; and two types of transduction trees: the first kind matches a succession of patterns to an input, and hence
generates an English "gist clause" as an interpretation; the second kind matches patterns to
gist clauses in the context of LISSA’s prior output, to generate an appropriate response, or
select a sub-schema. A schema may be invoked with arguments that are needed to execute
it but cannot be specified in advance, such as the name of the user. Our conception of
a dialogue schema is similar in spirit to a Schankian script, and an extension of the more
general concept of event schemas.

Invoking a schema causes the agent to update its overall dialogue plan and begin se-
quentially executing the set of actions suggested by that schema. The plan contained in a
schema outlines how LISSA expects to carry out the relevant discourse, specifying actions
(in easily understood predicate format) such as listening to the user and asking a general
question. The plan may be modified according to LISSA’s interpretation of the user’s input.
For instance, if the user asks for a question to be repeated, LISSA will insert actions into
the plan to accommodate that request.

Schemas are intended to be hierarchically structured, so that during the execution of the
dialogue plan provided by a schema, LISSA may invoke a subschema in response to some
input by the user. For instance, if LISSA asks about the user’s favorite free-time activity
while executing the top-level plan, and LISSA interprets the user’s response as signaling
interest in movies, LISSA will invoke a subschema to ask the user about their favorite kinds
of movies. This hierarchical structure allows LISSA to accommodate spontaneity in the
conversation while maintaining comprehension and recall.

4.4.2 Receiving Input

The user’s input is gathered using a laptop microphone with the Nuance speech recognizer.
We found that this speech recognizer achieves good performance for this application. A
significant challenge in developing this system was the automation of turn-taking during
the conversation. For many trials, we chose not to allow the system to handle turn-taking
automatically; instead, a human would operate a remote control letting LISSA know when it
should speak, and when it should listen. To automate this function would require addressing
the open problem of determining when a speaker intends to end their dialogue turn and
the precise time at which it would be appropriate to begin speaking in response. When
we attempted to automate turn-taking, we encountered technical issues and found that
the system’s poor timing could cause significant frustration among research participants.
Because of this, we chose to leave automated turn-taking as an area of future inquiry.

4.4.3 Understanding input

When LISSA receives input from the user, it arrives at a reaction through a two-stage
template transduction process. Associated with each dialogue schema are sets of rules cor-
responding to each question LISSA may ask the user. When the user provides input in
response to a question, LISSA first produces one or more high-level interpretations of the
user’s input that we call “gist clauses”. This is an ad hoc, an English-like format conveying
as much of the meaning content of the user’s input as we can extract using the transduc-
tion rules corresponding to the relevant question in the dialogue. These rules work by first
attaching features to certain words in the user’s input, such as GOODPRED for words like
“happy,” and ACADEMIC-SUBJECT in addition to the sub-feature SOCIAL-SCIENCE for
“linguistics,” then using pattern matching to arrive at the best interpretation among a set of
predefined gist clauses. At least one gist clause will always be produced, and more than one
will be produced if LISSA finds multiple units of salient content in the user’s input, such as
a direct answer followed by a question from the user.
4.4.4 Reacting to input

After producing gist clauses from the user’s input, LISSA will attempt to react appropriately. Gist clauses that are indicated as questions will cause LISSA to give a prepared answer. Gist clauses marked as responses to questions will undergo a second stage of template transduction, producing a concrete action for LISSA to execute. For instance, if LISSA asks for the user’s name, and the user says “Sam”, the gist clause derived from the user’s input will cause LISSA to say “I’m glad to meet you, Sam”.

This second stage of template transduction may also trigger the invocation of a subschema in response to the user’s input. Additionally, every gist clause is stored in LISSA’s memory, so that, for instance, if the user tells LISSA about their favorite class in school, LISSA won’t ask for this information later in the conversation. Fig. 4.5 shows an illustrative example.
Chapter Five

LISSA Autism Study

In this chapter, we present the design of LISSA focusing on design lessons learned from trials with teens with autism spectrum disorder. The initial design of the interface was guided by an expert UX designer, psychologists, and a pediatrician. Here we focus on adapting the system design to teenagers with autism spectrum disorder (ASD). We conducted an initial study that informs further development of a system, evaluation with intervention studies, and subsequent deployment broadly. The study, whose lessons we explore here, involved nine teenagers with ASD. Figure 5.1 shows two teenagers with ASD interacting with LISSA. With the help of professionals in developmental and behavioral pediatrics, we recruited teens with high functioning ASD for preliminary interaction with LISSA and interviewed them about the experience. Through a thematic analysis of the interview transcripts we then identified several key design guidelines, including the following: 1) Users should be fully briefed on the purpose of an interface, and its capabilities and limitations. 2) The interface should incorporate positive acknowledgment of behavioral changes. 3) Realistic appearance of a virtual agent and responsiveness are significant factors in engaging users. 4) Conversation personalization would help the users engage, and thus have them benefit from the interaction.
5.1 Study Design

From 2015 to 2018, we have been recruiting participants for our studies through the developmental/behavioral pediatric research center at the University of Rochester medical center. As already mentioned, our experimental sessions have been conducted with nine teenage participants (i.e., between 13 and 18 years old). The participants were all diagnosed with some form of autism. We specifically, selected these nine participants based on their ability to read and have conversation. The research center, through which we recruited them made sure that all participants have high functioning autism. The goal of this study was to learn what aspects of LISSA are useful, and what adjustments need to be made in order to make LISSA useful iterative conversation training tool for teens with ASD. The study sessions with the teens (and parents) were scheduled on separate days. Each participant first interacted with LISSA for five minutes, then took a break for two minutes, followed by a second conversation with LISSA for another four minutes. For dialogue topics, we picked the ones that are common in casual conversations such as "getting to know each other", "living in the current city", "crazy room", "city I want to move to in future", "free time", and "movies". During the conversations, the participants received real-time feedback through the flashing icons (see section 3.1). After each conversation, the participants received the post-session summary feedback (see section 3.2). We conducted an interview with both the parent and the participant right after the LISSA session. The interview included survey questions on
LISSA’s usability and open-ended discussion. The interview was audio recorded and then transcribed by professional transcribers.

### 5.2 Survey Results

We presented 12 statements to the participants and asked them to specify their opinion ('strongly disagree' to 'strongly agree'). The questions were aimed to understand the usefulness of the feedback, the dialogues, and overall user experience. The questions were inspired by the well-established system usability questionnaire (Brooke, 2013). Figure 5.2 shows the specific questions and percentage of participants’ answers in each category. The questions marked with a star (*) were answered significantly more ($p < 0.05$) with options 'agree' or 'strongly agree' compared to other options. A single sampled non-parametric significance test (Mann and Whitney, 1947) with Bonferroni correction (Armstrong, 2014) against the option 'neutral' was performed.

Participants felt that they were being understood by LISSA and expressed that they could continue the conversation and pay attention to the icons without any trouble. This indicates
the applicability of real-time feedback for teens with high functioning autism. Additionally, participants felt that the feedback they received from LISSA was useful. During our interview session participants expanded on this perceived usefulness. The feedback was consistent and in accord with what their therapist said in the past. For instance, one participant had issues with his posture (i.e., slouching) and he received feedback through the 'body movement' icon. During the interview, the participant mentioned this and said that his parents often ask him to sit straight.

As can be seen in the figure 5.2, participants had mixed opinions about several questions, such as whether the conversational experience felt real, and whether LISSA’s movements seemed natural. Three participants responded positively about the latter question, four responded negatively, and two were neutral. In the interview session, several mentioned that the lip movement and eye gaze were unnatural. Additionally, LISSA was not responding immediately. This was due to the fact that LISSA processes the dialogues and facial features in real-time and the processing takes place on a remote server. In our future versions of LISSA, we will make it more responsive by performing most of the computing locally.

5.3 Qualitative Analysis

We performed a thematic analysis (Guest, MacQueen, and Namey, 2012) on the interview transcripts. In the past, thematic analysis was successfully used for user-centered design (McCurdie et al., 2012) and rapid online interface prototyping (Kinzie et al., 2002). Additionally, thematic analysis was used for identifying the design guidelines for developing computer and phone-based technology to help improve the social skills of the children with ASD (Mintz, 2013; Mintz et al., 2012). In our analysis, three researchers performed thematic analysis and then the themes were merged to produce the final analysis report. As a basis for a qualitative thematic analysis of the interview transcripts, we considered indi-
vidual interviews from the perspective of the following 14 labels: usefulness, perceptiveness, related systems, accuracy, familiarity, curiosity, realism, speed, appearance, improvements, social, multitasking, uncertainty, and adult identity. As a result, six themes (closely related to some of these labels) stood out to us as relevant to summarizing the experiences of the participants. These are elaborated in the following. The first theme briefly summarizes positively perceived aspects of LISSA, while the rest dwell mostly on aspects where further developments are desirable. We believe it is at least as important to focus on weaknesses as on strengths, as a guide to further development.

Utility for practicing conversation in private

Participants generally found LISSA useful for practicing conversations. Additionally, they thought that the program was not hard to navigate. One participant said,

"Wasn’t that hard to use. It could definitely be used for somebody who really needs help with conversations or for somebody who is not really social or for somebody who is not really the kind of person to be talkative."

Participants liked the fact that the feedback was not coming from a human and they could use it in their private space without being observed. When we asked if they prefer human or a computer for giving feedback, some were ambivalent, some preferred the computer ("I would rather have that (LISSA) for feedback,"), and some were skeptical about LISSA.

Some participants with favorable reactions added that the automatic facial feature detection and the accuracy of the feedback were the main reason for endorsing LISSA for conversational skills training. One participant said,

"The fact that it was able to actually detect the facial features and everything being so accurate, I would consider that is good enough to actually train on."
The caregivers, as well, liked the fact that LISSA allows users to have a conversation with a virtual agent instead of (for example) a stranger online. They felt that LISSA was realistic and appreciated that it provided quite complex, and private interactions.

**Self-awareness**

When conversing with LISSA, participants were often evaluating the experience in relation to real, social settings, for example, whether the feedback they received was truly appropriate. One participant noted that smiling broadly enough for LISSA to notice might be perceived as inappropriate in a public setting:

"If they can make it so that she can respond a little quicker and that she is able to pick up the smile little better because I was smiling a little bit she just didn't think it was good enough and if I go too big then people are gonna think I am creepy."

Occasionally the self-aware evaluation of LISSA took the form of push-back against the presumption that they needed to improve their behavior, implicit in LISSA’s feedback. One participant indicated that any inadequacy in their behavior with LISSA was due to the awkwardness of interacting with a virtual agent:

"Well, I am actually social. I do make good eye contact with other people but it’s just kinda awkward you know."

Another teenager commented similarly:

"I do try to be social with other people and I am good with that it’s just that this was just a little new for me.

From this it appears that some participants would need some time to become more familiar with LISSA. After multiple, unobserved interactions they might well feel more comfortable
with the system. The comments also suggest that users might be more accepting of LISSA if the purpose of the system were more fully explained – that is, it is simply a tool for users who would like to improve technical aspects of their conversational ability, if they feel they could improve in that area.

Realism of the virtual human

Participants focused quite intently on the realism and precision of the LISSA persona. There were debates on how realistic the eyes, lips, face, voice, and overall movement were. Participants suggested various improvements, such as more flowing speech, immediate responses, and faster blinks:

"The other part I didn’t like about her is because it took her a while to respond to my statements. So it was kind of confusing and irritating."

"I would say blinking eyes was definitely a little bit slow, because if I blink it looks almost instant. But for that it was half a second total, it seems quite slow."

"Nice sounding voice, nice face but the moving the lips thing was kind of a little creepy. Mainly because kind of it's little too computer-ish maybe."

When participants were asked about the usefulness of LISSA, their judgments were dependent on realism. For the other interview questions, participants had varying responses, but there always seemed to be an underlying fixation on realism.

"Well, it was good that she asked me what I liked to do. But since it wasn’t really real it just seemed all awkward for me. I would make it better by making the quality High Definition and High quality better with talking to her for real."
Multi-tasking and feedback

When they were asked about multitasking between the conversations with LISSA and looking at the icons, there were different opinions: some found the icons helpful, and others wanted more explanation and bigger icons.

"Maybe the icons at the bottom make it a little like little bigger and maybe make it a little stand out a little bit more."

And when asked about if they preferred feedback from a computer or human, the participants had mixed opinions. For example, one participant said,

"The feedback I received from LISSA was useful. Well it was kind of a bit choppy and kind of pretty much prefer something a bit more realistic."

Interpreting feedback through LISSA’s behavior while conducting a conversation may be overwhelming for many participants. Perhaps if LISSA herself provided the feedback verbally during the conversations, rather than relying entirely on icons for feedback, it would feel more like a real interaction to users. Regardless of their preferences in feedback delivery, all participants expressed that the feedback was consistent with what other people had said to them in the past face-to-face training sessions. One participant said,

"They (feedback) were actually kind of handy and appeared to be pretty accurate."

An important observation was that users broadly agreed on the need for real-time positive feedback. While the flashing red icons noticeably indicated the need for improvement, the reversion to static green icons was not sufficiently noticeable as positive feedback. This indicates both the need for changes in icon functioning and, once again, the desirability of verbal feedback (especially acknowledgement and praise) from LISSA.
Related systems

At the end of the interviews, participants tended to compare LISSA to other conversational agents. They were often familiar with Siri and Alexa, and discussed the standard set by these virtual assistants in terms of speed and knowledgeability. They liked that LISSA could detect behavioral features, but hoped that LISSA could better understand and recognize them in the future. For example, a participant said,

"In its current state yes I might mess around with it some but I don’t believe I would use it as an actual social skills training and jump into actual conversations just yet."

It seemed that they had high expectations of LISSA which in some cases led to impatience during the conversation. One participant said they would feel more comfortable talking with systems like Siri, that they believe understands them.

"That’s like my first time ever talking to a computer, well except for Siri that’s different though. That one is fine."

The desire for adult identity

One teenager, after being urged by a caregiver to be honest, admitted that they probably wouldn’t choose to interact with LISSA, unless perhaps LISSA "popped up on their computer". Part of the reason seemed to be that LISSA straddles the boundary between fantasy and reality; and while fantasy is fine for a kid, for real conversations they prefer actual humans:

"I don’t know if it was like a real person. Cause I do like fantasy things but I am also a kid who but when it comes to talking to people I like talking to real people."
These participants were ambivalent about the future use of LISSA because they were inclined to regard it as a conversational tool for children. One participant said their schedule was too busy, but others could benefit from LISSA.

"Just so you know I am already 17 years old, I am growing up and some of these little kids’ things I have outgrown but not all of them."

In one conversation, a participant felt momentarily uncomfortable with LISSA’s comment about a "crazy room," (asking them to speculate what kind of crazy room they would enjoy) and said they wanted to feel like an average adult:

"Well, pretty much just the crazy room cause well I wanna be what your average adult is. Basically responsible, kind, but also a bit unique."

These comments again suggest the need to make LISSA’s purpose clearer to users. It is not intended to be a surrogate human, but rather a tool for repetitive, private practice of conversational behavior. Also, being sensitive in the choice of terminology is important, and perhaps asking for a creative response is inappropriate for some participants.

5.4 Discussion

A majority of participants found that LISSA provided useful feedback and might well be helpful for practicing their conversational skills. As we noted, they also liked the fact that LISSA would allow them to converse in private. Another point of interest is that the participants preferred real-time feedback to post-session feedback. Our interviews with users helped to shed light on how the interface design could be further improved. Our qualitative analysis of these interviews provide evidence for the soundness of our design so far and grounds for optimism about our further development plans. The analysis also allowed us to formulate several important guidelines for the design of LISSA-like interfaces for conversation practice.
Appropriate Prior Briefing of Users about LISSA’s Purpose

Our experience with users made clear that users’ assessment of LISSA’s behavior and potential utility for conversation practice depended very much on their expectations. They generally acknowledged that LISSA seemed to understand them and responded appropriately to inputs, yet was not genuinely human-like. The perceived shortcomings concerned LISSA’s physical behavior, accuracy of perception of user behavior, and depth of knowledge. In part, this perception arose from comparisons with commercial systems such as Alexa and Siri, which have been optimized for smooth functioning in targeted information retrieval and other assistive functions.

These reactions indicate the need for fuller preparation of users about LISSA’s purpose: It is not a surrogate human, and it is not an app for access to useful knowledge or personal assistance. It is simply a tool for repeatedly practicing casual conversation for those who feel they could improve in that area. While LISSA has a range of verbal reactions to users depending on their particular inputs, and provides nonverbal feedback as a function of the user’s behavior, the conversations are bound to be shallow, and to become more repetitive with multiple uses. Further, LISSA’s physical behavior is not the focus; it is merely intended to be sufficiently human-like to make a casual conversation possible. All this should be made clear to potential users – along with a comment that there is no assumption that all users with ASD are lacking in the skills that LISSA is intended to help with. Such preparatory information could be provided both in advance of actual use of LISSA, and as part of LISSA’s opening remarks (which already include "I might sound a bit choppy, but ... ").

The post-session interviews of users should likewise focus on relevant aspects of LISSA’s functioning. For instance, interview questions might include disclaimers, as in, "We know that LISSA doesn’t smile and blink very naturally, but was the content of her responses to you reasonable and natural?". In general, it is evidently important to prepare users not only for the capabilities of an interactive system, but also its limitations.
Positive Acknowledgment of Behavior Change

As noted in the previous section, the participants wanted to be made aware of positive changes. Perhaps flashing green could be used for behavioral improvements. Better yet, the virtual agent could say, for example, "You have good eye contact now". The efficacy of positive feedback and acknowledgment has been observed in past research (U.S. Department of Health and Human Services, 2014; Scott-Van Zeeland et al., 2010; Kasari et al., 1988) and our experience further confirms the desirability of positive acknowledgments for interventions aimed at social skills development.

Realistic appearance of a virtual agent and responsiveness

Notwithstanding disclaimers about LISSA’s physical behavior, the issue deserves further attention. A possible reaction to users’ comments about insufficiently realistic smiling, eye blinking, and reaction speed might be to back away from the "uncanny valley" (e.g., Chattopadhyay et al. Chattopadhyay and MacDorman, 2016) by using a more cartoon-like avatar. However, this would risk reducing LISSA to a toy in the eyes of potential users. Instead, we interpret the users’ comments as urging further development of the avatar towards greater realism. This is consistent with their age - they are approaching adulthood, and prefer a realistic to a childish avatar. In fact their comments suggest that the more life-like the character, the more likely they are to take it seriously. Realistic appearance of virtual characters has also been shown to be effective in other scenarios such as negotiation, tactical questioning etc. (Kenny et al., 2007; Tambe et al., 1995). The most important areas for improvement seem to be smiling and eye blinking. (Smiling is of course well-known to be very important in communication.) For example, smiling needs to be consistent with current feedback (one participant commented on co-occurrence of a smile by LISSA with negative feedback). Furthermore, smiles could be used directly to indicate improvements, or in support of positive verbal or icon feedback.
Conversation personalization would help the teenagers engage for longer terms.

Some participants expressed enthusiasm about home-use of LISSA. However, they varied in their opinions about the choices of topics. For example, while the "crazy room" topic struck a chord with some (e.g., they would fill it with video games), others objected to it, terming it childish. In addition, participants thought that it would be useful if LISSA could talk about topics of their own choosing. For example, one participant was very interested in computer programming and wanted to talk about it more. The LISSA program, at its current stage, is designed for initiating conversational topics, and treating a specialized topic like computer programming seriously would be a major challenge. However, adding further mundane topics is quite feasible, and in fact we have added many more in a related application currently under development. Thus we could personalize interactions to a considerable degree by having LISSA choose topics dynamically, skipping those that the user seems indifferent to. Also, choices could be made sensitive to the user's age or maturity. An immature 13-year-old may have quite different interests from a mature 18-year-old. (Some of our newly developed topics pertain only to seniors, just as some of LISSA's topics for teens, such as bullying at school, pertain only to school-age users.)

Another opportunity for personalization lies in the verbosity or otherwise of the user. Our experiments showed that while some users provide expansive responses to LISSA, others respond tersely. As the goal of the system is to help users improve their communication skills, the system could gauge users' verbosity and provide helpful feedback where appropriate. For instance, LISSA might encourage laconic users elaborate their answers, or conversely, provide gentle suggestions about curtailing rambling or off-topic inputs. Assessing users' verbosity throughout the conversation can also help improve turn-taking behavior. Users who tend towards longer answers should probably be allowed slightly longer silences before the turn is seized from them. The same applies to hesitant, slower speakers. Certainly humans adapt to such individual differences. This is an important research area – to our knowledge, no
available automatic turn-handling methods take into account individual speakers’ verbosity or rate of speech.

One of the most important observations we made about teens with ASD in comparison with (neurotypical) college students, independently of verbosity, was that the teens refrained from asking any reciprocal or other questions of LISSA (e.g., after telling LISSA about their favorite movie, asking "and what’s your favorite movie?"). Whether this is due to less willingness to treat LISSA as human-like, or to limitations in social intuition, it is an area where verbal feedback by LISSA could be particularly useful; for instance LISSA might say, "This would be a good point to ask me about my favorite movie. Would you like to try?".

5.5 Limitations and Future Work

The current version of LISSA was not designed for immediate use in a randomized control intervention study. Rather, it is an exploratory system, which will enable a randomized control study after modification and enhancement based on the lessons learned from the trials with the initial set of teens with ASD.

LISSA’s dialogue manager was adapted from the initial version for college students to the anticipated needs of teens with ASD, with advice from experts. It worked well, but the experiments have shown where improvements are most desirable, for example in topical adaptation to the user, inclusion of direct helpful hints in the verbal reactions to the user, and allowance for different turn-taking styles. Similarly, LISSA’s nonverbal feedback system was trained on the data collected from college students, and although the participants perceived the feedback as useful, our experimental results indicate ways in which flashing icons, sensitivity to user smiles, and reaction speed could be improved. Also, while the post-session interviews indicated that the users liked the appearance and voice of the avatar, they saw a need for improvements in the naturalness of the avatar’s behavior (especially smiles and
blinking of eyelids). In our future work, we will design a customizable interface based on the knowledge we gathered through this study.

Data collected from teenagers with ASD using future versions of the system will help us to further improve the sensitivity and responsiveness of the system. In the current system, the dialogue and the feedback modules are independent. It clearly would be useful to tie the nonverbal feedback to the dialogue content, and to supplement nonverbal feedback signals with direct verbal ones. In the future, we will make the feedback dialogue aware.
Chapter Six

Aging and Engaging

In this chapter, we present a conversational skills training program for older adults called "Aging and Engaging". Initially, it was designed for older adults who visit senior centers. Many older adults who visit senior centers are not able to have positive and meaningful social interactions, precisely because they lack the conversational skills to do so. In this chapter, we present a pilot study which was conducted test the acceptability of the program. We then present a longitudinal randomized control study to test the generalizability. We present how we used the feedback to develop the online communication program and how the participants perceived it.

Figure 6.1 Aging and Engaging virtual conversational agent
6.1 Interface

We designed an online communication skills training program. The program has two main interfaces. One is the conversational interface and the other one is the feedback interface. The conversational interface features an older woman from SitePal (Fig. 6.1). We consulted with a focus group of 12 older adults. The appearance of the virtual agent was suggested by the focus group. Our interface allows users to converse with the virtual agent by pressing a single "Start" button. After conversation the users can receive feedback on three nonverbal behaviors – eye contact, smiling, speaking volume—and one verbal: valence of speech content (sentiment). These four behaviors were selected because they are empirically linked to poor social functioning ([struchen2011examining]; Youngren and Lewinsohn, 1980). During the conversation with the virtual agent, the system uploads the recorded audio and video to a server for processing. After the conversation, the system uploads the transcript of the conversation to the server to generate feedback. The system automatically takes users to the feedback page. See 3.2.2: Verbal and Nonverbal Feedback for more details on the feedback. Fig. 6.2 gives an overview of the system we used.

**Figure 6.2** The overall system. Users initiate a conversation, which is driven by a dialogue controller. Conversation audio and video are processed in a server and feedback is generated. Users receive the feedback one by one and move to the next conversation phase. After four rounds of conversation, users receive final feedback summarizing the previous
6.2 Pilot Study

To assess the acceptability of the interface to its intended audience, we conducted a study with 25 participants who were 65 years old or older. The average age of the participants was 68.5. Of those 25, six were single, eight were divorced, one was widowed, and ten lived with their spouses. Participants were recruited from a hospital-based geriatric mental health clinic (n=5) (where difficulties with nonverbal behaviors are common) as well as through print advertisements (n=20). All participants were native English speakers. Our participants varied on self-perceived social skills, as we believed more diverse feedback would be most useful at this stage. However, the majority of our participants (67%) were below the 50th percentile on a population-normed measure of social skills (PROMIS Self-Efficacy to Manage Social Interactions). The study took place in a private room in the medical center. Participants first consented to the study with a trained assessor. We emphasized that participation would have no impact on services received from the Medical Center if any, and our research staff had no connection with the clinic. After the consent process, we explained how the interface works. Participants then interacted with the virtual agent by pressing the start button on the interface. During the interaction, the participants went through four conversation phases. After each phase, the system redirected participants to a feedback page. After four rounds of conversation and feedback, the participant received the final feedback (Fig. 3.5), which summarized the feedback previously given and suggested areas for improvement. Our research staff did not intervene beyond helping participants open the interface in the web browser. After the interaction, we asked participants a series of questions about the interface, their experience, their demographics, self-perceived social skills (PROMIS) (Hahn et al., 2010; Gruber-Baldini et al., 2017) any depression or anxiety (Riley, Pilkonis, and D Cella, 2011), their social connectedness (Van Orden et al., 2012), and their use of computers or other electronics. The study served to collect data since there
is no dataset available from any previous studies that target the desired behaviors in older adults. We did not conduct this study as a system intervention. We plan to conduct such an intervention study in the future to evaluate the effectiveness of the system.

6.3 Results

6.3.1 Survey Results

We showed the participants four statements and asked them to rate each between one and five, where one indicates strong agreement, and five indicates strong disagreement. The statements and their average ratings are shown in Table 1. On average, participants self-reported the interface as easy to use, with a value of 1.96 (SD=1.02). Our intuitive design choices, with minimal button-pressing and voice-assisted feedback, might have made our interface easy to use. Participants also disagreed with the statement about the system taking too long to use, rating it a 4.04 (SD=0.97). We wanted to know more about the opinions of the participants who agreed with the statement, “I feel disconnected from other people” (n=12). Fig. 6.3 shows no significant difference in ratings of the interface between participants who felt disconnected and those who did not. There was also no difference in the ratings provided by those who felt like outsiders at social events (n=12) and those who felt lonely (n=20). The participants were further divided based on a self-reported measure of perceived social skills (in Fig. 6.4). We did not find any significant difference in the ratings between the two groups. These results are important because they show the interface is acceptable to people who have (or are at risk for) difficulties with social communication (Struchen et al., 2011).

In our study, 19 participants possessed a home computer or a laptop—we called them the “computer user group”—and six participants had no access to computers—the “computer non-user group.” These two groups’ ratings are also shown in Table 1. The computer non-
user group thought the program was easier to use than the computer user group. This may be because the computer user group had higher expectations in terms of feedback speed and system responsiveness. The computer non-user group tended toward not wanting to use the program in their home. But these differences are not statistically significant. The lack of difference between computer users and computer non-users is important because it shows that our system is acceptable to even those who have limited experience with technology.

We also looked at the average ratings of the statements after grouping participants based on the presence (or absence) of clinically-significant depression, anxiety, and social isolation, as these are characteristics that often coexist with communication difficulties in later life. Table 2 shows the average ratings. In it, the “yes” column contains the ratings of those who reported elevated levels of depression, anxiety, or social isolation using population-based norms. In these three categories, we did not find any significant differences in ratings between the groups, which, while inconclusive given our sample size, nonetheless suggests that those who feel isolated, depressed, or anxious are not less likely to find the program acceptable. This is important, as those who report these characteristics are most likely to have deficits that could improve the program.

To see how the system performed in terms of feedback accuracy, we compared the system-generated feedback against that of human judges. Two trained research assistants watched each of the interactions individually and decided whether to offer positive or negative feedback. We calculated the accuracy of our system by comparing these determinations as ground truths. Table 3 shows the average accuracy of the system when compared to the decisions of the two human judges. The inter-rater reliability (Cohen’s Kappa) of the two human judges for each feedback dimension is also shown in Table 6.1. We found that our system-generated feedback accorded most with human judges on eye contact. Some of our participants were not sure where to look to receive positive feedback. Others sat facing the light, which reflected on their eyeglasses, making it difficult to accurately determine eye gaze
direction. Some dialogue topics became negative even when participants tried to keep them positive. Our system only looked at the ratio of positive words to negative words, not more complex syntactic structures. There were instances in which participants described a positive phenomenon with negative words and thus received negative feedback on their speech content.

Table 6.1 Table 3: Accuracy of feedback compared to human judges

<table>
<thead>
<tr>
<th>Statements</th>
<th>Accuracy</th>
<th>True Positive Rate</th>
<th>False Positive Rate</th>
<th>Cohen’s kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye Contact</td>
<td>79.60</td>
<td>76.06</td>
<td>3.50</td>
<td>0.84</td>
</tr>
<tr>
<td>Volume</td>
<td>72.00</td>
<td>58.50</td>
<td>14.25</td>
<td>0.79</td>
</tr>
<tr>
<td>Smile</td>
<td>66.00</td>
<td>60.25</td>
<td>15.35</td>
<td>0.62</td>
</tr>
<tr>
<td>Content</td>
<td>68.80</td>
<td>67.91</td>
<td>17.36</td>
<td>0.67</td>
</tr>
</tbody>
</table>
6.3.2 Feature Analysis

We looked at how participants were performing in each phase of their conversation. We took the percentage of participants who received positive feedback according to our human annotators (see Figure 6a). We saw a statistically significant difference between the first phase of conversation and the last (fourth) phase of conversation ($p<0.05$) on a smile, volume, and content. This suggests that participants were able to reflect on the feedback and change those behaviors more likely to elicit positive feedback. Among the four dimensions, the percentage of participants receiving positive feedback on eye contact changed the least ($\Delta = 0.31\%$), whereas smile changed the most ($\delta = 83.15\%$). We also looked at the average speaking time of the participants during each phase of the conversation (Figure 6b). Participants’ response times increased in subsequent conversations. Average response time for each question during the first conversation was 24 seconds (SD=4.5), whereas the average response time for the final conversation was 40 seconds (SD=4.9) ($p <0.05$). This suggests that, as the conversation progressed, the participants become more comfortable with the interface and revealed more information about themselves.
6.3.3 Interview Results

In order to get additional feedback to help us improve our system, we included a qualitative interview with our participants (this interview was added midway into data collection, thus we only have feedback from n=15 subjects). We took a phenomenology approach to design the interview guide, which was developed by one of our study investigators. Feedback was obtained from an additional researcher and the interview guide was edited accordingly. The interview guide included questions focused on participants’ experience during the time they engaged with the program and suggestions for improving the program.

Interviews were audio-recorded with participants’ permission and transcribed. We then conducted a thematic analysis using principles of grounded theory (including the constant comparative method) to analyze the data (Corbin and Strauss, 1990). Coding was done by one investigator and then reviewed by another investigator (we used open coding) (Corbin and Strauss, 1990). Discrepancies between the two raters did not arise as the themes were relatively straightforward and unambiguous. We added user codes with the quotations.

Accuracy of the feedback

All participants commented on their perceptions of the accuracy of the feedback, as they reflected on their performance and suggested areas for improvement. Many participants perceived that the feedback was accurate (n=5). “Every single thing she said I should try to be better at was absolutely correct.” (Subject 2, 69 y/o female, 45th percentile on social skills). Insight into their behaviors and its relation to feedback accuracy was also raised by several participants (n=3). “I’m pretty much aware of all those things I don’t do well. I don’t smile a lot, I don’t vary the pitch of my voice. I have a very soft voice.” (Subject 14, 74 y/o female, 39th percentile for social skills).

However, an equal number of participants thought the feedback was inaccurate (n=5). This may be the case because the training data was collected from college students, rather
than older adults. A key objective of the current study was to obtain data from older participants to better train the system for our subsequent efficacy study. One way in which our older participants differed from many of the younger participants was that most wore eyeglasses, which made it difficult for the program to recognize eye gaze direction. Further, some of the participants were not sure where to look (at the webcam or into the virtual assistant’s eyes) in order to get positive eye contact feedback. At first, this frustrated them, but eventually, they understood and looked at the assistant’s eye to get positive feedback. As another said, “I was frustrated by inaccurate feedback, but once it seemed to be in sync, it was fine.”(Subject 11, 70 y/o female, 18th percentile on social skills).

**Comfort with the interface**

Most participants (n=10) spoke about their level of comfort with the interface given that it was a novel experience for all of them. Some spoke about how they initially felt uncomfortable engaging with the program, but that this improved over the course of the session. “I got more comfortable as I went on because I wasn’t sure how it was going to work at first.” (Subject 2, 69 y/o female, 45th percentile on social skills). Another said, “I think the more it went on, the more comfortable you get with it.” (Subject 7, 69 y/o male, 55th percentile on social skills).

Many participants also spoke about the ways in which the program was easy or difficult to use (n=8). We carefully designed our system using feedback from our focus group. We had several key design concerns, which included participants’ ability to read the feedback, navigate through the interface, and hear properly. We minimized button pressing to make the program more intuitive. Said another: “The program is easy to use; I didn’t need training as I do with other new programs.” (Subject 11, 70 y/o female, 18th percentile on social skills).
Useful program for engaging

We asked our participants about how useful they thought the system was. Most participants thought the program could help improve conversational skills (n=12). Our interface allows these people to practice and possibly improve conversational skills without feeling judged by a human. Said one of our participants: “I know that I could probably improve by practicing. I’ve never had the opportunity to practice with something like this, which would be an impersonal coach. I started to think of her as a coach who wouldn’t react—like a psychiatrist, just listening and letting the person do all the talking.” (Subject 14, 74 y/o female, 39th percentile for social skills). Our participants also said the program could be a great teaching tool. Said another, “I think it’s a great teaching tool. The program is already helpful; it just needs tweaking.” (Subject 3, 67 y/o female, 32nd percentile on social skills). Our participants also viewed this program as a tool that can be helpful to those who are shy and withdrawn, learning the skills needed to connect and communicate with others. This program might also be helpful for those who find it difficult to start a conversation. “This would be good for people who have issues—who are so withdrawn that they can’t communicate. This would help them break out of their shell, and give them instructions on how to move forward.” (Subject 8, 65 y/o female, 70th percentile on social skills).

Content of the discussion

We carefully chose dialogue topics after discussing it with our focus group. Several participants (n=5) commented on the content of the dialogue. “I would’ve done all five conversation topics- they were great. There was nothing intrusive, it was very basic.” During the conversations, the virtual assistant asked some questions relevant to their selected dialogue topics. Our participants found that the questions the virtual assistant asked were common, natural, and comprehensive. “I thought the questions were so realistic. Those are the questions I grill people on all the time when I meet them. It was very natural.” (Subject 8, 65 y/o female,
70th percentile on social skills). Some of the participants felt that the conversation was one-sided, as the virtual assistant was not able to respond to all the questions the participant asked. “I gave answers but I didn’t feel like I had the chance to ask her questions. It felt strange... The conversation is too one-sided.” (Subject 2, 69 y/o female, 45th percentile on social skills)

**Suggestions to improve the interface**

Two participants addressed the fact that the conversations had no time limit, allowing participants to speak as long as they wanted. It was not clear to the participants how long the conversations would continue. One participant suggested using a timeline. “A timeline would be helpful—knowing how long each section is going to be, and how far along you are in the program.” (Subject 5, 83 y/o male, 61st percentile on social skills). Another participant brought up the issue of sensory impairment in later life. “At first I had a hard time hearing, but after she started talking, you kind of get what was going on. It could be part of the hearing problems I have, but once I got into it, I could understand her voice. As it went on it got better.” (Subject 7, 69 y/o male, 55th percentile on social skills). To further improve our program, we plan to consult with the geriatrician on our advisory board for features we could add to better accommodate age-related hearing and vision loss.

**6.3.4 Discussion**

Even though the agreement between the feedback generated by our algorithm and two labelers was fairly high, there is room for improvement. The feedback module was trained with videos of college students who are younger than our target population and have different facial, prosodic, and behavioral characteristics. We used a hidden Markov model to better understand temporal patterns to generate feedback. In the future, we will continue to improve our proposed machine learning model to generate more accurate feedback. Our
exploration will include other non-temporal models (i.e., support vector machine) and neural network-based classifiers (i.e., recurrent neural network). In this exploratory study, we collected 25 videos of interaction with older adults. We will improve our feedback module by using this new set of videos. Formulating feedback by counting positive and negative words has many limitations. Future work will involve more semantic analysis, using techniques from natural language processing.

The effectiveness of the system (i.e., improving social skills) was not assessed through the preliminary study. The objective of the study was to assess the acceptability of the system in older adults. This type of intervention is acceptable to younger adults. There are no evidence-based reasons to predict that older adults would react differently than younger adults to the prospect of using the program. Some of the end-users take part in the design of the system. Some of the older adults we included did have difficulties with social skills. We also included those without these difficulties because we believe a wider range of feedback is helpful. In the future, we will recruit participants with social skills deficits.

The study was a one-time visit by the participants. Thus, it was not necessary to remember the information from the previous sessions, and there was none. However, the system remembers the feedback it gave to the participant in subsequent conversations (in the same day) and synthesizes them to produce final feedback. In the future, we will enable the system to remember the past conversation by augmenting the login capability. Participants appreciated the dialogue we chose. The interactions, however, were one-sided, as the virtual assistant was not able to answer all the questions posed by our participants because we generated a limited number of responses. The transcripts collected from this study will allow us to generate more responses for the assistant. To automate the dialogue controller, we will build a topic-based dialogue module. The automated dialogue module can have conversations on particular topics and the responses will be chosen by pattern-matching. Spontaneous back-channeling was not implemented in our current prototype. In the future,
we will incorporate features, such as sharing a smile or nodding appropriately, which can lead to a more natural interaction. This remains an open research problem for us and the virtual assistant community. Our aim is to build an interface that will become a part of the users' lives. With the initial exploratory study and findings, our next step is to deploy a fully automated system online. We used the Wizard-of-Oz technique as our first step to build the automated dialogue system. This study gave us the valuable data necessary for an automated dialogue manager. It will allow us to run a longitudinal study allowing participants to use our system from home. It is important to investigate how the participants are able to utilize their practice with our interface to improve or maintain their relationships and maintain better health and quality of life. We will have older adults who demonstrate difficulties with social communication (as measured by low levels of social functioning and impairments on standardized ratings of communication behaviors) engage with the program several times and assess their communication behaviors at baseline, at the end of the two weeks of practice, and three months later. Doing so will allow us to measure learning and change over time, as well as maintenance of gains after finishing their sessions. Our study presented here is a necessary first step to this future study.

Further, later life brings with it numerous potentially stressful social encounters, including discussions about end-of-life care, for example. Even older adults with strong conversation skills will likely to have some difficulties with these conversations. Thus, a future direction with our system is to adapt it for assisting all older adults with challenging social situations common in later life, including end-of-life planning, discussions with physicians, discussions with adult children about giving up driving and moving to retirement communities.
6.4 Longitudinal Study

In the work reported in this paper, we focused on validating the program in the wild. Specifically, we wanted to see whether older adults can retain and generalize the skills in the real world after interacting with an AI-driven system. In addition, we wanted to see if the older adults would be able to use such a system on their own. To address these questions, we ran a longitudinal randomized control study. Fig.6.6 shows the study design. The treatment participants used the Aging and Engaging program for eight days from their home over the period of a month. The control participants also completed an online conversational skills development program over a month with eight exposures. To assess the change in participants’ conversational skills each participant participated in three role-play sessions before and after their conversational skills developing the program. During the role-play session, participants had face-to-face conversations with a trained research assistant. Another research assistant observed a face-to-face conversation. Both research assistants rated the participant on their conversational skills independently. We used the average of their ratings in our comparison. The research assistants were blind to the participants’ condition. We then compared the two groups’ performance in their role-play session. Our results indicate that the treatment participants improved their eye contact and facial expression. Additionally, the treatment participants improved overall conversational skills. In contrast, the control group did not show any improvement in the four targeted behaviors.

We made the following contributions -

1. We have introduced a fully automated online conversational skills-building program specifically for older adults. The system is available anytime, anywhere using a computer browser.

2. We designed the program targeting specific behaviors for developing conversational skills with longer exposure to validate the generalizability of the intervention and the
Figure 6.6 The study design. Participants first have a face-to-face conversation with a rater and then they are randomly assigned a condition - treatment or control. All participants in both conditions complete their online program (Aging and Engaging for treatment, an online pamphlet for control) in eight sessions. After the intervention, all participants have another round of role-play sessions. In both role-play session the rater rates the participant on their conversational skills.

scalability of the system.

3. Through a randomized control study consisting of eight independent practice sessions over a one month period, we show that the Aging and Engaging program helped individuals (65+ years old) to retain and generalize their conversational skills in face-to-face conversations.

As mentioned above, each interaction consists of three subsessions. During each subsession, the virtual agent leads a casual conversation on a topic, which usually consists of 3-5 speaking turns of each side. In order to make the interaction more natural and motivating, the virtual agent provides sporadic self-disclosures over the course of the conversations by talking about her life, beliefs, interests, etc. She introduces herself as a 65-year old widow who moved to the city a few years ago to live with her daughter and relates relevant information or experiences of hers. The virtual agent’s choice of character, questions, comments and expressions were meticulously designed by a team of RAs who were in close contact with gerontologists with expertise in interventions. The gerontologists suggested 30 topics along
Table 6.2 Examples of conversation topics and their emotional intensity in Aging and Engaging Program

<table>
<thead>
<tr>
<th>Emotional intensity</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>Getting to know each other, Where are you from?, City where you live (I), Weather, Family and friends, Activities and hobbies, Plan for rest of the day, Household chores, Pets</td>
</tr>
<tr>
<td>Medium</td>
<td>Holidays and gathering, Cooking, Travel, Outdoor activities, Managing money, Home, Employment and retirement, Books and Newspapers, Dealing with technology</td>
</tr>
<tr>
<td>Hard</td>
<td>Life goals, Growing older, Spirituality, Tell me about your hopes and wishes</td>
</tr>
</tbody>
</table>

with several questions that can be asked under each topic. The focus group’s potential inputs and appropriate verbal reactions were discussed with the experts, which formed the basis on which the dialogue manager was designed so as to lead a natural, engaging conversation with users. To facilitate conversations that aren’t cognitively overwhelming to elderly patients, the gerontologists divided the topics into three categories based on their emotional intensity or degree of intimacy: easy, medium, and hard. The easier topics involve little personal disclosure, while the harder ones are more intimate and emotionally evocative. Table 6.2 shows some examples of topics in each category. The dialogue manager is designed so that users start with easier topics in early sessions and smoothly transfer to harder ones in later sessions. In a separate study, we learned that the participants tend to show higher self-disclosure and provide longer responses when conversing about hard topics Razavi et al., 2019.

To handle dialogue on each topic, the dialogue manager follows a predefined modifiable plan, which specifies a sequence of intended and expected interactions with the user. The
plan contains variables that are instantiated over the course of conversation based on the
user’s inputs and the avatar’s reaction. At each user turn, the user’s input is interpreted
in the context of the previous virtual agent’s input. A hierarchical pattern transduction
tree is used to extract one or more simple English sentences as the gist of the user’s input.
These gist-clauses are used by the dialogue manager to generate appropriate verbal reactions
by applying a reaction hierarchical pattern transduction method to the gist-clauses. More
details on the dialogue manager design and structure can be found in Razavi et al., 2019.

To run a longitudinal study we further customized our system. First, the program was
set to conduct the conversation over three topics per day. The system provided feedback
after each conversational topic and final feedback after the whole conversation. Starting
from the second day of the study, the system first reminded them what they talked about
on the previous day, what type of positive and negative feedback they received, and what
improvements they should focus on during the current conversation.

6.4.1 Study Design

We conducted a longitudinal randomized controlled study. The study was approved by the
Institutional Review Board of a research university. In this study, we recruited 18 partici-
pants from community advertisements and an outpatient geriatric psychiatry clinic. All the
participants were 60 years old or older (mean = 71.5, sd = 7.5). The study was described
to potential participants as examining “methods of improving conversation skills”. While
self-report is limited in accurately identifying individuals with difficulties in social commu-
nication, study resource limitations required some form of rapid screening to increase the
likelihood that participants enrolled in the study would demonstrate communication diffi-
culties. Thus, the inclusion criterion for the study was at least mild self-reported difficulties
on the social skills and communication sub-scales of the Autistic-Spectrum Quotient, a val-
idated self-report screening tool Baron-Cohen et al., 2001. Exclusion criteria were assessed
during a phone screen: diagnosis of dementia (based on self-report) and lack of an email address and/or access to the Internet in a private location (requirements of the intervention). Physical functioning and psychiatric disorders/symptoms were not exclusion criteria, with the exception of adequate hearing and vision to interact with a computer program (hearing aids and glasses were permitted).

Each participant was invited to the lab one at a time. The participants first filled out the consent form, demographic information, and completed a series of self-report questionnaires, used here to characterize the sample (see section 6.4.2). Next, the participant completed a performance-based assessment of social communication, the Social Skills Performance Assessment (SSPA, see section 6.4.2), which involves structured role-plays with an assessor. In the past, assessment using role-play has been used widely in the context of social skills (Torgrud and Holborn, 1992; Kazdin, Matson, and Esveldt-Dawson, 1984; Ratto et al., 2011). After the role-play session the participants were assigned either a treatment or control status randomly. The treatment participants used the Aging and Engaging program for 8 days to improve their conversational skills. The first day of the Aging and Engaging program was completed in the lab the same day they completed their baseline role-play session. For the next seven days, participants received an email with a link to the program. The participants followed the link to start the program. All treatment participants completed their next seven sessions from their home, using their home computer and webcam. The participants were instructed to complete their session within 3 days of receiving the email with the program link. A research assistant made phone calls and sent emails if they were not completing the sessions on time.

The control participants received an email every three days containing a link to a pamphlet (Wendler, 2014). The pamphlet contains articles and exercises on how to improve communication skills. This pamphlet was used in previous studies (Ali et al., 2015) for comparison of control participants with treatment participants. We had the control participants
read from a computer, to better match the conditions between the two groups. After completing the eight days of the conversational skills training, the participants came back to the lab for their follow-up session. During the follow-up session, the participants had another role-play session with a research assistant. The research assistants rated the participants the same way they did in the baseline sessions. In both the baseline and follow-up visit, the research assistants were blind about whether the participants were in the treatment or control group.

6.4.2 Methods

We performed several analysis to answer our research questions. In this section, we describe our method of analysis.

Participant characteristics

We first characterized our sample, in order to place our subsequent results in context and ensure that our participants were representative of the population the system was designed to help. Our participants completed the following self-report assessments, which have been validated with older adults.

1. PROMIS computerized adaptive tests were used to assess depressive and anxiety symptoms (Schalet et al., 2016), social support (Hahn et al., 2010), and two aspects of social functioning—satisfaction with one’s social roles and activities (‘social satisfaction’) and self-efficacy to manage social situations (‘social self-efficacy’) (Hahn et al., 2010). Scores from these measures are population-normed T-scores with a mean of 50 and a standard deviation of 10, with greater scores indicating ‘more’ of a given construct (e.g., greater social self-efficacy).

2. Cognitive performance was measured with the Montreal Cognitive Assessment (MoCA),
a brief screen for cognitive impairment with scores ranging from 0 to 30 and scores greater than 25 indicating intact cognitive performance (Fujiwara et al., 2010).

3. Functioning/disability was measured with the World Health Organization Disability Schedule 2.0 (WHODAS)(Üstün et al., 2010), which assesses the self-report of several domains of function—cognitive, physical, and social—that can be reduced by health problems (both physical and mental health). We used the scoring method that was developed using item-response theory and produces summary scores ranging from 0 to 100, with corresponding population percentiles (Üstün et al., 2010).

4. Participants provided a rating on self-perceived health (range from 1-5, very dissatisfied to very satisfied).

5. Participants also completed the UCLA Loneliness scale (Russell, 1996), which is a self-report measure of perceived social isolation, with higher scores indicating greater loneliness.

**Conversational Skills**

We wanted to see if the Aging and Engaging program helped the participants improve their conversational skills in face-to-face conversations with real humans. The Social Skills Performance Assessment (SSPA) is an observer-rated assessment of communication elicited through standardized role-plays (Patterson, Moscona, et al., 2001; Patterson, Goldman, et al., 2001). Participants’ behavior during the role-plays is coded along several dimensions. Patterson and colleagues (Patterson, Goldman, et al., 2001) developed the SSPA to assess social functioning in adults diagnosed with schizophrenia (and other severe mental illnesses) and displaying severe (versus more subtle) difficulties with communication skills, including difficulties taking turns in conversation and not knowing how to initiate conversations. In contrast, our system is designed to help older adults with a wider range of difficulties with communication,
including more subtle, micro difficulties, such as eye contact. Thus, we made minor adaptations to the SSPA for our project and population in two ways: first, by altering the scenarios in the role plays to emphasize skills involved in joining new groups: instead of scenarios with a landlord that call for assertiveness, our participants engaged in two role-plays more suitable for individuals attempting to reduce social isolation — an interview for a volunteer position and joining a group at a senior center. Second, we used the coding scheme from the Social Interaction Test (Trower, Bryant, and Argyle, 2013) which places equal emphasis on nonverbal and verbal communication (whereas the SSPA emphasizes higher-order verbal communication skills such as assertiveness). Third, we added a dimension of social skill that is relevant to older adults attempting to overcome social isolation, namely, using a great deal of negative content in their choice of what to speak about (e.g., physical illnesses, stressors). Two independent trained research assistants rated participants on four skills that we targeted with our intervention, two dimensions of verbal skills—speaking volume and negative content—and two dimensions of non-verbal skills that we targeted with our intervention—facial expressiveness, and eye contact. Each dimension was coded on a five-point scale, with five indicating no impairment/high skill and zero indicating a high degree of impairment (lack of skill) that impacted the conversation. Each item included examples of presentations that correspond to each score for that item to increase reliability between raters. All participants interacted with the same person for the role-plays — a doctoral-level clinical psychologist (research assistant) who used standardized prompts and responses. The SSPA involved three role plays: 1) an unscored practice to orient the participant to the process, 2) introducing oneself to a group of people at a senior center; and 3) interviewing for a volunteer position. Each participant was rated by two independent raters, including a doctoral-level clinical psychologist who was blind to all information gathered in the assessment prior to the role play and blind to which condition the participant was randomized to (for follow-up assessments). For each dimension, we averaged the scores from the two raters. The specific questions are
presented as rating item and the response options were given with descriptions. The question items and the response options are shown in Table 6.3.

The responses from the two research assistants were averaged for each participant. Thus, we have two ratings for each of the participants – baseline and follow-up ratings. We then compare the ratings between the control and treatment groups for each of the rating items. In addition to the individual rating items, we have calculated a composite conversational skills ratings by taking the mean of the rating items. We performed a paired two-sample t-test for both the treatment group and control group to see the difference in the baseline and follow-up ratings. Since we performed multiple significance tests, we applied a Bonferroni correction (Weisstein, 2004) to all of the p-values, by multiplying them by the number of tests.

Case Study

To better understand how participants interacted with the Aging and Engaging program and how the program helped them, we performed two case studies. To identify a potential candidate for a case study we first looked at the change in participants’ composite skills score. From the treatment group, we selected the participant who improved the most and the participant who improved the least as our candidates.

6.4.3 Results

In this section, we present the results of our analysis we mentioned in section 6.4.2.

Participant characteristics

The characteristics of the participants are shown in Table 6.4. The average age was 71 years (range of 60 to 86 years). The majority (68.4%) were female, identified as white (84.2%), and reported having graduated from college (79.0%). Fewer than half reported being mar-
Table 6.3 Conversational skills rating questions and response options.

<table>
<thead>
<tr>
<th>Rating Item</th>
<th>Response Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye Gaze</td>
<td>1. Completely avoids looking OR Stares continually</td>
</tr>
<tr>
<td></td>
<td>2. Abnormally infrequent looking OR Abnormally frequent looking</td>
</tr>
<tr>
<td></td>
<td>3. Looks too little OR Looks too much</td>
</tr>
<tr>
<td></td>
<td>4. Tends to avoid looking OR Tends to look too much</td>
</tr>
<tr>
<td></td>
<td>5. Normal gaze frequency and pattern</td>
</tr>
<tr>
<td>Volume of speaking voice</td>
<td>1. Inaudible OR Extremely loud</td>
</tr>
<tr>
<td></td>
<td>2. Abnormally quiet OR Abnormally loud</td>
</tr>
<tr>
<td></td>
<td>3. Too quiet OR Too loud</td>
</tr>
<tr>
<td></td>
<td>4. Quiet OR Rather loud</td>
</tr>
<tr>
<td></td>
<td>5. Normal volume</td>
</tr>
<tr>
<td>Facial Expression</td>
<td>1. Totally blank face OR Frequent strongly negative expressions</td>
</tr>
<tr>
<td></td>
<td>2. Abnormally inexpressive OR Abnormally strong negative expressions</td>
</tr>
<tr>
<td></td>
<td>3. Often inexpressive OR Frequent negative expressions</td>
</tr>
<tr>
<td></td>
<td>4. Tends to be inexpressive OR Mildly negative expressions</td>
</tr>
<tr>
<td></td>
<td>5. Normal range of emotional expressions</td>
</tr>
<tr>
<td>Content</td>
<td>1. Almost all negative content, very negative impression</td>
</tr>
<tr>
<td></td>
<td>2. Overly focused on negative content, negative impression</td>
</tr>
<tr>
<td></td>
<td>3. Some negative content but negative impression</td>
</tr>
<tr>
<td></td>
<td>4. Some negative content but no negative impression</td>
</tr>
<tr>
<td></td>
<td>5. No negative content</td>
</tr>
</tbody>
</table>
ried (42.1%), and close to half reported living alone (47.4%). Average scores for depressive and anxiety symptoms were in the non-symptomatic range, although there was significant variability, with some participants reporting mild and moderate symptomatology. There was also significant variability in functioning/disability, with the average WHODAS score (8.87) corresponding to a population-normed 70th percentile (indicating significant disability), with a range from no reported impairment to the 88th percentile for impairment. Participants reported average scores for social support (T-score=50.33), loneliness (mean score=43.59), and satisfaction with their social roles and activities (social satisfaction T-score=50.23). However, as would be expected given our inclusion criteria regarding self-reported difficulties with social communication, the average score for self-efficacy in managing social situations (social self-efficacy T-score=45.71) indicates mild impairment. These characteristics indicate that our study sample demonstrated difficulties with social functioning, as well as age-associated cognitive, sensory, and physical declines and we also included older adults with cognitive functioning falling in the range of mild cognitive impairment. This indicates that our results can meaningfully generalize to the population of older adults who could benefit from our program.

**Conversational Skills**

Next, we compare the change in conversational skills between the face-to-face role-playing sessions before and after the experiment. Table 6.5 shows the average rating of the research assistants on the four different conversational skill cues and a composite rating for the treatment group and control group. Among the four cues, eye gaze and facial expression had significantly higher ratings in the follow-up session than the baseline for the treatment group. It is important to note that the facial expression rating can be considered as a proxy for a smile, since we did not have an explicit rating for the smile. The overall composite rating also increased in the treatment group (3.25 to 4.25). However, none of the ratings
in the control group changed significantly. To check if the baseline ratings of the treatment and control groups were different or not, we performed an unpaired t-test on each of the four cues independently. None of the tests suggested a significant difference, which indicates that the treatment and control groups’ baseline conversational skills were similar.

Case Study

Fig. 6.7 shows the change in the composite score (baseline vs. follow-up). From here we see that in the treatment group only one participant (ID 3) did not improve. The participant with ID 6 improved the most. We selected them as our candidates for a case study.

Case Study 1: Participant A, ID 6

Participant A is a 69-year-old white female who is divorced and lives alone. She has a Master’s degree and is retired. She doesn’t report any mental health problems, anxiety, or depression. She takes a prescription medication for arthritis and reports mild difficulties with short-term memory which are consistent with her performance on cognitive tests; she also demonstrates some difficulties with the cognitive control domain of executive functioning. She reports mild social anxiety, including anxiety in new social situations and difficulties with chit chat and starting and maintaining conversations. She reports average levels of
social support but low levels of companionship. Additionally, she reports being very socially engaged, including volunteering every week, participating in social groups and clubs several times per week, and getting together with friends and family several times per week.

In this longitudinal study, she showed improved ratings in her speaking volume, facial expression, and eye contact. She already had a perfect baseline score in conversation content, which is why there was no improvement in the follow-up session.

During her baseline role play, she spoke softly, displayed very little expression in her face, exhibited frequent pauses as she spoke, and avoided eye contact. One rater noted,

“She seated at 90-degree angle at the table due to the table shape. Very awkward, closed, stiff, quiet, and uncomfortable socially.”

During her home sessions with the Aging and Engaging program, she received eight negative feedbacks (out of 24) on eye contact. This negative feedback came mostly on days six and seven. On the last day, she received positive feedback. This indicates that she tried to maintain her eye contact at the end of the program. She received ten negative feedbacks in total on her speaking volume. However, in this case, most of the negative feedback came on day 1 and day 3. Also, we noticed that when she received negative feedback she had long conversations. When she started getting positive feedback we noticed that her conversations were shorter. This indicates that she may have had trouble maintaining a good speaking volume for long periods of time. When she received negative feedback she started talking less, but she kept her speaking volume high. She received 16 negative feedbacks on her smile. There was no clear pattern of her smile feedback. We can assume that she might have worked on her smile throughout her home sessions. Surprisingly, she never received any negative feedback on conversation content. We investigated the contents of her conversation with the virtual agent and found that she was very positive in her speech.

At her follow-up role play, the participant continued to speak softly and to pause frequently, but her face was more expressive, including more frequent smiling, and she made
an appropriate amount of eye contact. She received a perfect score for eye contact, smile, and content. Although her speaking volume rating improved, it was not perfect. In her follow-up interview, she reported that the feedback she received from the program was useful, accurate, and consistent with the feedback she had received from others in the past. She reported that she became “more aware of her communication habits” from using the program. Subject A stated that the feedback from the program (both written and verbal) was helpful, the questions asked by the virtual agent were “relatable and answerable,” and the program was “simple and straightforward” to use. She disliked that the virtual agent was “unnatural” in her responses and that the agent would interrupt her during the conversations. She suggested that our program could be improved by “having humans instead of computers because it may be more motivating.”

**Case study 2: Participant B, ID 3**

Participant B is an 84-year-old white male who is married and lives with his wife from his second marriage; he is retired and has a doctorate in education. He is a Veteran and served in the U.S. army. He reports a history of depression but is not currently taking medication for depression and his scores on tests of depression and anxiety are in the normal range. Similar to participant A, he reports mild difficulties with short-term memory, which are consistent with his performance on cognitive tests. He also demonstrates difficulties with the cognitive control domain of executive functioning. Unlike participant A, however, he also demonstrates significant difficulties with other domains of executive functioning, including working memory, as well as more severe difficulties with cognitive control. On a self-report measure of social difficulties, he agrees with the following items: “I find it hard to make new friends,” “When I talk, it isn’t always easy for others to get in a word edgewise,” and “People often tell me that I keep going on and on about the same thing”. He reported being dissatisfied with his social relationships and difficulties making friends.

From the baseline to the follow-up session his rating improved a little on eye contact.
However, his conversation content rating declined, and ratings of speaking volume and facial expression remained the same. During his baseline role play, he spoke quickly and for extended amounts of time. He spoke over his conversation partner in the role play and spoke for such extended periods of time that the interviewer conducting the role play needed to cut him off. He perseverated on certain topics—including some with negative sentiment, demonstrated difficulty focusing, and did not ask the other person questions. One of the raters noted,

“(He) would perseverate on a specific topic; did not really ask me any questions and would talk over me (I had to interrupt at times to get him back on track or redirect the conversation or close conversation).”

The other rater noted,

“... a bit impulsive in speech - cut off interviewer a few times, spoke over the interviewer, spoke a bit fast, and talks for a bit too long at times and needed to be cut off.”

Additionally, both of the raters noticed that he had difficulty in keeping his focus and maintaining eye contact while listening.

During the home sessions, the participant received a lot of negative feedback on eye contact (20 negative feedbacks out of 24). At one point he said, “I noticed when I’m talking with you that I tend to look up and that’s not unusual for people in their thinking”. We found that usually he takes longer turns than the virtual agent and since he talks for a longer period of time, he does not make eye contact while talking. His smile and speaking volume feedback received a mix of positive and negative feedback. We did not find any interesting pattern on his smile and speaking volume feedbacks. On conversation content, he received 18 negative feedbacks out of 24. While looking at the content of the conversation, we found that he was somewhat negative about the things he was discussing with the virtual
agent. For example, while conversing on the topic of pets he said, “In my family, there are a number of cat lovers that I’m very aware of. I think cats are really dangerous but most people don’t know it because they (cats) carry a little parasite and when they scratch you it enters your blood and eventually sometimes gets to people’s brain. Most people are just not aware that they are a dangerous animal in many ways”.

In his follow-up role play, he continued to speak very quickly, speak too long, and to interrupt his conversation partner. Additionally, he showed a negative sentiment throughout the role-play session. One of the raters noted,

“Really turned off by the arrogance and the opinionated/condescending nature of the content from the subject. Very quick transitions between role play. The subject seemed in a hurry to be finished with it.”

In his follow-up interview, participant B reported that the feedback he received from the program regarding negative content was “helpful and maps onto his personal experience.” Regarding the other aspects that the program suggested he could improve, Subject B reported that he “could’ve used some convincing of what needed to be improved” in those domains and suggested that we could improve the program by including in-person feedback with a therapist the first time someone uses the program “to be able to have some discussion around it or be able to review the tape.” Similar to participant A, he reported that the program could be improved by including “more human contact.”

These examples suggest that our program, in its current form, maybe most useful for individuals with a certain profile of communication difficulties characterized by too much inhibition—shyness, anxiety, poor eye contact, and low expressiveness—versus those with too little behavioral inhibition—talking over others, speaking loudly, inappropriate negative comments. The latter profile may involve cognitive impairment, particularly executive functioning difficulties, as a contributing factor to difficulties communicating effectively.
6.4.4 Discussion

In this paper, we sought to answer the question of whether an online conversational skills development tool can help improve a person’s skills in face-to-face conversation, particularly in the case of older patients at risk of isolation. We found that the participants who used the program for eight days were rated as more skillful communicators than control participants in the follow-up role-play sessions. In the role-play sessions, the conversation partners were assigned roles that mimic potential real-world scenarios closely. Because the role-playing took place in a controlled environment, conversational skills could be assessed by an expert observing the face-to-face interactions. This allowed us to determine whether the program helped to improve conversational skills in face-to-face conversations, including on topics falling outside those covered by the program.

Among four verbal and nonverbal cues, eye contact and facial expression showed significant changes from baseline to follow-up ratings. This is consistent with previous work, which has found that these two behavioral cues are easy for users to improve Ali et al., 2015. Most participants received positive feedback on eye contact. This suggests that their eye contact skills were moderate to begin with and were improved through positive feedback. However, this was not the case for smiling, on which more than half of the participants received negative feedback. The results showed that frequently reminding users to smile more leads to improvements in smiling by the time of the follow-up role-play session.

The speaking volume of the participants was generally already high during the initial role-play session, which leaves a little room for improvement. The conversation content ratings did not improve either. The program consistently encouraged users to take a positive attitude during the conversation. This was difficult to comply with for some of the topics. For example, one of the conversation topics was 'friends and family'. Some participants were lonely and discussed not being able to see their families. However, it is possible to discuss even challenging topics in a more positive manner, which may be especially effective when meeting
new people. For example, one participant who received feedback about negative sentiment in a prior topic stated that she is disappointed that she does not have any grandchildren—an indicator of negative sentiment—but she then turned the conversation positive by making a joke that she is fortunate to have a 'grand-dog' living nearby that she enjoys taking for a walk. In future revisions of the program, it may be useful to give participants examples such as this one so they can better understand how to shift their negative statements to more positive ones. Currently, the program places more emphasis on helping participants notice what types of improvements they can make and less on specific ways to improve.

Our results, while promising, should be placed in the context of several limitations. One set of limitations involves how we selected our sample. There are no validated screening instruments for communication problems in older adults that can be administered over the phone or online in order to select a sample with difficulties in communication. Communication problems are not routinely assessed in mental health clinics either. Thus, our study relied on a self-report measure of perceived difficulties in social situations to select a sample. The result was that our sample was not highly impaired on communication at baseline and thus did not have a great deal of room for improvement. Developing methods to screen for individuals who may benefit from this program will be a necessary component of our work moving forward. Second, behaviors that are effective for communication are context-dependent, meaning that behaviors can be effective in one context (e.g., a doctor's office) but less effective in another (e.g., meeting a new friend). Further, individuals may have a harder time with communication in some settings versus others. For example, discussing a cancer diagnosis with a doctor will be more challenging for most people compared to making small talk over lunch with friends. Identifying contexts that are most appropriate for testing the efficacy of our program and that are most relevant for individuals to use as practice in developing their skills is another necessary direction for our work going forward. Third, because assessing social communication behaviors is resource-intensive, our sample size was
relatively small. Given promising results, however, an additional component of our future work will be testing the program with a larger sample size, including identifying any modifications that are necessary to improve efficacy. Fourth, we did not find improvements in two of the behaviors we targeted, speaking volume and negative content in a conversation. These null results may be due to limited difficulties with these behaviors at baseline, but may also indicate a need to improve the program with regard to these behaviors, perhaps through providing examples of how to make these changes. Another possible modification, suggested by our subjects, could be including a discussion with a human therapist or coach at the beginning of the program about what behaviors they could improve in and how to do so in order to increase motivation for using the program and maximize benefit.

Including only a single meeting with a therapist as an option could balance the benefit of including a human helper with the increased accessibility and reach of the intervention provided by offering it online without the need for an appointment with a therapist. Finally, it would have been desirable to conduct additional follow-up interviews several months after the subjects completed the program, and to perform assessments outside of our laboratory; but this was precluded by resource limitations. Thus we were unable to assess whether the changes in the participants’ behaviors attributable to the program were sustained after completing the interventions, generalized to settings in their day-to-day life, or whether these changes positively impacted their relationships, health, and well-being; after all, it probably takes time for behavior changes such as increased eye contact to impact relationship quality. Despite these limitations, this study has value because it demonstrates the feasibility of this line of work and produced a signal for efficacy that supports a subsequent, more resource-intensive study.

Regarding possible modifications, our results suggest that different intervention strategies may be needed based upon the type of difficulties individuals have with communication, such as being too inhibited versus not inhibited enough. These different patterns of com-
munication difficulties may be associated with different etiologies (causes) of the problems, including cognitive impairment or frontal lobe dysfunction resulting in decreased executive function and cognitive control that manifests as we saw with Subject B who spoke loudly, interrupted others, and discussed inappropriate topics. In contrast, Subject A spoke softly, avoided eye contact, and smiled infrequently, which could be caused by mild social anxiety. Other presentations may include Parkinson’s Disease, which impacts motor function and dampens expressivity. Other presentations could be context-dependent; for example, an older adult who presents as sad and unsure of himself and has difficulty joining a group for lunch may be showing the effects of recently losing a spouse and moving into a retirement community.

The promising results of this study indicating that our program is acceptable to our population of interest and improves eye contact and smiling. This indicates additional refinement and testing of the program could be useful.

Improving social communication could have numerous benefits in later life (and across one’s lifespan). For older adults who have difficulty communicating, it is challenging to maintain positive relationships and build new ones. Some older adults may have struggled with communication over their lifetime, while others may struggle when drifting towards depression in later life or when confronted with new challenges, such as social role transitions (e.g., retirement, bereavement), relocation (e.g., moving to a senior living community), and changes in health. One pertinent example is Parkinson’s Disease, whose consequent impairments in motor functioning are associated with difficulties in social functioning. Other older adults may have strong communication skills in most domains but may have some difficulty in especially challenging situations, such as end-of-life discussions with physicians and family. Difficulties with communication may have serious consequences: Medically serious suicide attempts among adults are associated with difficulties communicating emotions (Szanto et al., 2012). Thus, problems with social communication may inhibit efforts to seek
social support and timely help for mental health concerns.

Our study has several strengths, including a focus on a relatively understudied contributor to mental health in later life—social communication—and development and testing of a novel intervention for improving social communication. Social communication can be impaired in later life, especially in conjunction with mental disorders, and alleviating this impairment can provide one pathway to improving social function, mental health, and well-being in later life. Thus there are potential benefits in the availability of an evidence-based, low-intensity intervention for communication that can be delivered in the home with minimal support from interventionists.
Table 6.4 Participant characteristics. Notes: \(^a\) 1 participant identified as black, 1 as multiracial, and 1 declined to report on race. \(^b\) 7 participants reported being divorced, 1 legally separated, 3 widowed, and 0 as never married. \(^c\) PROMIS scales are computerized adaptive tests with population normed T-scores with a mean of 50 and standard deviation of 10. \(^d\) The full title for the ‘social satisfaction’ scale is ‘satisfaction with social roles and activities.’ \(^e\) The full title for the ‘social self-efficacy’ scale is ‘self-efficacy to manage social interactions’.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Mean (std) or n (%)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>71.47 (7.51)</td>
<td>60 – 86</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>13 (68.4%)</td>
<td></td>
</tr>
<tr>
<td>Race (White) (^a)</td>
<td>16 (84.2%)</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>8 (42.1%)</td>
<td></td>
</tr>
<tr>
<td>Living alone</td>
<td>9 (47.4%)</td>
<td></td>
</tr>
<tr>
<td>Education (college or greater)</td>
<td>15 (79.0%)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychiatric symptoms, Functioning</th>
<th>Mean (std) or n (%)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive performance (MOCA)</td>
<td>25.63 (2.56)</td>
<td>20 – 29</td>
</tr>
<tr>
<td>Depressive symptoms (PROMIS)(^c)</td>
<td>49.57 (8.87)</td>
<td>34.20 – 65.80</td>
</tr>
<tr>
<td>Anxiety symptoms (PROMIS)(^c)</td>
<td>50.18 (10.23)</td>
<td>32.90 – 65.40</td>
</tr>
<tr>
<td>Disability/functioning (WHODAS)</td>
<td>8.87 (7.98)</td>
<td>0 – 29.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social relationships</th>
<th>Mean (std) or n (%)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social satisfaction (PROMIS)(^c,d)</td>
<td>50.23 (8.61)</td>
<td>35.10 – 64.60</td>
</tr>
<tr>
<td>Social self-efficacy (PROMIS)(^c,e)</td>
<td>45.71 (7.77)</td>
<td>33.00 – 59.00</td>
</tr>
<tr>
<td>Emotional support (PROMIS)(^c)</td>
<td>50.33 (7.66)</td>
<td>37.80 – 66.20</td>
</tr>
<tr>
<td>Loneliness (UCLA)</td>
<td>43.59 (11.82)</td>
<td>23 – 60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social communication</th>
<th>Mean (std) or n (%)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal impairments</td>
<td>1.21 (1.03)</td>
<td>0 – 4</td>
</tr>
<tr>
<td>Nonverbal impairments</td>
<td>2.21 (1.78)</td>
<td>0 – 6</td>
</tr>
<tr>
<td>Total impairments</td>
<td>3.42 (2.27)</td>
<td>0 – 8</td>
</tr>
</tbody>
</table>
Table 6.5 Average rating of the conversational skills in role-play session.

<table>
<thead>
<tr>
<th>Rating Items</th>
<th>treatment baseline mean (sd)</th>
<th>treatment follow-up mean (sd)</th>
<th>p</th>
<th>control baseline mean (sd)</th>
<th>control follow-up mean (sd)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye Gaze</td>
<td>4.70 (0.35)</td>
<td>4.94 (0.17)</td>
<td><strong>0.02</strong></td>
<td>4.17 (0.79)</td>
<td>4.72 (0.36)</td>
<td><strong>0.06</strong></td>
</tr>
<tr>
<td>Volume</td>
<td>4.70 (0.42)</td>
<td>4.78 (0.51)</td>
<td><strong>0.15</strong></td>
<td>4.50 (0.61)</td>
<td>4.83 (0.35)</td>
<td><strong>0.12</strong></td>
</tr>
<tr>
<td>Facial Expression</td>
<td><strong>4.50 (0.47)</strong></td>
<td><strong>4.94 (0.17)</strong></td>
<td><strong>0.01</strong></td>
<td><strong>4.61 (0.42)</strong></td>
<td><strong>4.61 (0.49)</strong></td>
<td><strong>0.48</strong></td>
</tr>
<tr>
<td>Content</td>
<td>4.28 (0.62)</td>
<td>4.50 (0.61)</td>
<td><strong>0.29</strong></td>
<td>4.61 (0.42)</td>
<td>4.28 (0.57)</td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td>Composite</td>
<td><strong>4.56 (0.27)</strong></td>
<td><strong>4.81 (0.16)</strong></td>
<td><strong>0.02</strong></td>
<td><strong>4.47 (0.41)</strong></td>
<td><strong>4.61 (0.31)</strong></td>
<td><strong>0.21</strong></td>
</tr>
</tbody>
</table>
Chapter Seven

SOPHIE - Standardized Online Patient for Human Interaction Education

In this chapter, we present an online virtual agent-based communication skills training program for improving communication between oncologists and their patients. Similar to LISSA and Aging and Engaging, SOPHIE features a virtual agent who presents herself as a late-stage cancer patient. The doctor can have a conversation with the virtual agent and after the conversation, they can receive feedback on several communication paradigms. In order to build this program, we have used a data set collected from the VOICE (Values and Options in Cancer Care) study (Hoerger et al., 2013). The data set consists of conversation transcripts between oncologists and their cancer patients. We first identified on which aspects of the communication we should give feedback based on a metric for prognosis understanding. From the data, we developed machine learning models to generate feedback on two specific communication paradigms. Our goal was to improve the prognosis understanding among critically ill cancer patients. In this chapter, we discuss the design of SOPHIE. Additionally, we present a pilot study with eight clinicians to find the usability of the system.
Figure 7.1 A physician conversing with SOPHIE
7.1 Design of SOPHIE

Our aim is to develop a virtual standardized patient for practicing communication skills. In medical education, students practice with a standardized patient – an actor/actress pretends to have a medical condition. Students interact with the standardized patients and are later they receive feedback on their interaction. Our goal is to allow the medical students to practice their communication skills with a virtual agent with multiple repetitions in their own environment. Fig. 7.1 shows a physician conversing with SOPHIE.

7.1.1 Scenario

We have developed a prototype of the SOPHIE program, which allows individuals to have a conversation with a virtual agent concerning prognosis and treatment options. SOPHIE presents herself as a late-stage cancer patient. For a pilot study, we selected a particular case for the virtual patient, inspired by a case from another study (Shields et al., 2019; Elias et al., 2017). We have used the SPIKES protocol to guide the conversation (Baile et al., 2000). The SPIKES protocol was developed to train physicians to deliver bad news. This protocol has shown success in increasing confidence among oncologists in delivering bad news. The SPIKES protocol has six steps – 1) setting up the interview, 2) assessing patients’ perception, 3) obtaining patients’ invitation, 4) giving knowledge and information to the patient, 5) addressing the patient’s emotion with empathetic responses, and 6) strategy and summary. With SOPHIE, at the beginning of the conversation (SPIKES step 1) SOPHIE introduces herself and mentions that she has lung cancer. Then SOPHIE raises the topic of her sleep pattern at night and asks if she needs to change her pain medication, allowing the physician to assess her perception (SPIKES step 2). She states that her current pain medication, Lortab, is not working anymore. After discussing the pain medication, SOPHIE turns attention to her test results, giving the physician a chance to obtain her invitation to
talk about difficult topics (SPIKES step 3), before asking more specifically about prognosis if the physician did not already address it, allowing the physician to give information to the patient (SPIKES step 4). SOPHIE then asks about what her options are, allowing the physician to give empathetic responses (SPIKES step 5). Finally, she follows up by discussing whether chemotherapy remains an option, whether she should focus on comfort care, what the side effects of chemotherapy are (if mentioned), and how to break the news to her family, allowing for the physician to give strategy & summary (SPIKES step 6).

While designing the scenario, we have kept a few things in mind which are important in end-of-life discussion.

- SOPHIE presents herself as if she already seeing a doctor and now her medication is not working. She already knows that she has cancer but she is not certain about how long she has left.
- SOPHIE gives an opportunity to the user to discuss her treatment options but asking about chemotherapy.
- SOPHIE gives an opportunity to discuss her prognosis.
- SOPHIE presents an opportunity for an empathetic response.

This type of discussion promotes understanding the patient, gather information from the patient, discuss critical information, and responding with empathy.

7.1.2 Feedback Interface

In the beginning, users start the conversation by pressing the "start recording" button. Users can then proceed to converse with SOPHIE, and when the conversation is over the program takes the user to the feedback page. The feedback interface is shown in Fig. 7.2. On the left side of the feedback interface, we show the conversation transcript. The red
marked speech is considered too long for the patients, i.e., it is classified as lecturing. On the right side of the feedback, we show the speech rate of the user, the number of questions the user asked, turn-taking, and the sentiment trajectory. Past literature has established that conversational speech rate is important in enabling patients to understand their prognosis. Also, asking questions of the patient is important for ensuring that the patient understands what is being said (Back et al., 2005). The turn-taking annotation shows the length of each turn by SOPHIE and the user. The example was chosen to illustrate the lecturing style of conversation; the detection of lecturing style was explained in section 3.3.2. The feedback shows the sentiment trajectory of both SOPHIE and the user. Additionally, the feedback shows a suggested sentiment trajectory for the user. The feedback page displays explanations of individual items when users hover their mouse on them.

**Figure 7.2** Feedback interface of SOPHIE. On the left side, the conversation transcript is shown. On the right (from top to bottom) speech rate, and the number of questions are shown. Turn-taking shows the turn length and at the bottom, the sentiment trajectory of both physician and SOPHIE are shown with the ideal/suggested sentiment trajectory.
7.2 Pilot Study

To further assess acceptability and usability, we conducted a pilot study with eight practicing clinicians (fellows, residents, and nurse practitioners) from a University Medical Center. Our goal was to gather more information about their experience with SOPHIE, any limitations, and how we could improve the system. The study was performed with one participant at a time on the zoom communication platform. Each day, we asked the invited participant to have a conversation with SOPHIE and to look at the feedback.

After conversing and receiving the feedback we interviewed the participants. The aim of the interview was to understand the accuracy and usefulness of the feedback, the appropriateness of the conversation, and suggestions for new features. We have performed a thematic analysis on the interview transcripts; our findings follow below.

Medical History

All the participants mentioned that a brief medical history should be presented before starting the conversation with SOPHIE. One participant said,

“I think some kind of medical record would be extremely helpful. I thought I don’t have any information to say to her.”

The participants mentioned that in a regular standardized patient visit, they are given a medical record before they go into the room. They suggested the same scenario should be replicated for SOPHIE. In our program, SOPHIE starts the conversation by mentioning her increasing pain. The participants felt that this was abrupt and there should be a transition to this serious topic. They also mentioned that the way SOPHIE initiated the presentation of her symptoms was unusual. In most cases, patients do not actively start the conversation. Rather, the physician looks at the patient’s medical record and then starts asking about any new symptoms. In the future, we expect to modify the dialogues so that
SOPHIE appears more passive and lets the users ask questions, though completely user-driven conversation remains beyond the state of the art.

**Topics of Conversation**

Participants (four out of eight) mentioned that SOPHIE jumped between topics and did not allow full coverage of a given topic. For example, SOPHIE begins talking about her pain medication, but the participants often asked questions about the current dosage and about other pain medication she had taken in the past. Since SOPHIE’s limited dialogue repertoire falls short of covering those questions, she starts talking about her current medication (i.e., Lortab) and then about her test results. One participant said,

“**The dialogue didn’t match with the questions I was asking. When she mentioned pain and I was trying to find more about the pain in order to help her with her question. But the answers that I gave her to her questions did not really fit and she just jumped to the next topic so I jumped with it but that was a little bit jarring to me.**”

Although SOPHIE changed the conversation topics abruptly, the questions she asked were found to be realistic. Five participants felt that SOPHIE was able to express her concerns and make them feel the seriousness of the situation. One participant added,

“**I think the topics were absolutely realistic. All the questions she asked were appropriate.**”

**Feedback on Speech Rate**

Participants (seven out of eight) mentioned that the speech rate feedback was easy to understand and very useful. One participant said,
"I know I tend to speak very fast, receiving feedback on my speech rate is going to be very useful."

Another participant mentioned that in normal practice there is no way of measuring the speech-rate. However, with SOPHIE we could provide information about how fast the physicians are speaking, which is useful.

“I think the feedback (speech-rate) was useful. I never had someone measure my speech rate before. Sometimes I try to be cognizant of speaking a little bit slower with the patients but it was nice to actually get some feedback like you are doing okay.”

One participant mentioned that it is important to speak more slowly when delivering bad news. She said,

“For me, it (speech-rate feedback) is useful because I know that I have a tendency to speak really fast. So especially when I am delivering bad news I try to be super cognizant.”

However, the participants also noted that SOPHIE’s speech rate was constant, making it difficult for them to adjust their speech rate depending on whether they are discussing serious issues or a casual topic. In the future, we plan to adjust SOPHIE’s speech rate based on the seriousness of the topic being discussed.

**Number of Questions Asked**

Seven out of eight participants expressed that feedback on the number of questions asked was very useful. One participant said,

“It was helpful to get the information about how many questions you have asked because I think a lot of the times we walk away from the conversation thinking
that we really invited the patients into the talk when maybe we didn’t and we did a lot of lecturing. So I think that was valuable feedback.”

In addition to the number of questions asked, participants suggested that it should be highlighted what type of questions were asked, for example, how many history-taking questions were asked and how many emotional questions were asked. Though they expressed mixed feelings, participants (six out of eight) stated that this feedback would encourage them to ask more questions in the future.

**Explanation of Sentiment**

The participants asked for more explanation on the sentiment trajectory. Seven participants mentioned that they did not understand the meaning of the sentiment values. They also said that sentiment feedback is hard to interpret and they often confused it with empathy. Four participants wanted to see an example sentence of positive and negative sentiment. The participants also mentioned that changing or adjusting sentiment while engaged in the conversation may add to the cognitive load. They suggested that instead of asking the user to be positive at certain moments we should just stress the importance of dynamically adjusting sentiment.

**Additional Feedback**

The participant also asked to add some additional feedback that they found useful in practice. Two participants said that there are few expressions of empathy in the dialogue and they should be highlighted in transcripts so that users could look back and understand how they responded to them. One participant suggested we should give feedback on the way users addressed concerns.

One participant said that the turn-taking feedback is useful, however, it does not show the total amount of time a person was speaking. The participant said,
"I tend to speak a lot, but I don’t want to make the patients feel that I am not listening. I want to know that I am giving a chance to ask questions."

He suggested the addition of a bar chart to the feedback page that indicates the total speech times for SOPHIE and the user.

Three participants suggested giving feedback on nonverbal behaviors, such as eye contact. One of them said,

“One of the things I think is important, and I have seen it in other clinicians, is eye contact. I think it’s super important when we are giving bad news or having difficult conversations. I have colleagues who tend not to look at the patients”.
Chapter Eight

Limitations and Future Work

In this chapter, we discuss the limitations of the work. We present the areas we can improve in the future and how we can address the limitations.

8.1 Nonverbal Behavior

In this thesis, we have focused on four nonverbal behavior, the verbal sentiment, and speech lengths to generate feedback. These verbal and nonverbal cues were found important in communication skills training, in the past. We also observed that, in our studies, participants were improving their communication skills, who received this feedback. However, there are many nonverbal cues that are important in face to face conversation. For example, backchanneling is an important factor in a face to face communication. Backchanneling is nonverbal feedback that listeners give to the speaker. In our thesis, the feedback focused on the speaker. Also, the virtual agents were not designed to demonstrate proper backchanneling. In the future, we aim to explore the backchanneling aspect of the conversation both from the perspective of the user and avatar.

Paying attention and engagement is another important behavior that we did not explore in this thesis. In the past, research has focused on engagement tracking in job interview scenarios (Yu et al., 2019). We plan to use engagement as a metric for communication skills
development in the future.

8.2 Feedback Modalities

We have focused on real-time feedback and post conversation feedback. For real-time feedback, we have only explored the idea of changing icon color. Although it showed promising improvements, in real-life conversation there is no icon that would change color. There are a few modalities on which feedback could be given. For example, the virtual agent could give direct feedback by saying, "you have made good eye contact so far, let’s focus on the smile now." We haven’t explored those possibilities yet. A more naturalistic feedback through speech or subtle behavioral change may be more effective. In the future, we will explore these modalities of feedback.

In the past, changing the virtual environment as a modality for providing feedback was successful (Alaraj et al., 2015; Kenny et al., 2007). In the future, we plan to explore changing the virtual agent’s environment as a mode of giving feedback. In addition to this feedback, we want to explore the content of the conversation and give feedback. We aim to detect the topic of the conversation and adjust the feedback accordingly. The topic of the conversation could be very serious, such as the death of a close friend, or how long someone has left to live. At that moment, giving feedback on smiling or being positive is not appropriate. We will perform the topic modeling on the conversation to identify the subtopics of the conversation and adjust the feedback. For example, when the conversation topic becomes serious we will not give feedback on smiling.
8.3 Study Limitations

Although we performed several studies to validate and evaluate the effect of the feedback, there are still open research questions that need to be answered. We have conducted a pilot study with teenagers with ASD, however, we haven’t conducted a randomized control study to see the effect of LISSA. The same can be said for the SOPHIE. In the future, we plan to run a randomized control trial to show the generalizability of the feedback.

In our communication skills development programs, such as LISSA, we used the feedback that was delivered in real-time and after a conversation. Also, the feedback was generated from the users’ verbal and nonverbal behavior. With the studies we performed, we can not say exactly, which type of feedback (real-time vs post feedback) actually works better than the other. We also can not say whether the verbal or nonverbal cues we should focus on. To address these questions we need to conduct randomized control studies by conditioning on the different modalities and behavioral cues.

In addition to the target population of our applications, we have not shown how the feedback can be used in other scenarios with a different population group. Our feedback could be applicable to many scenarios such as job interview training, customer care agent training, and public speaking.

8.4 Technical Limitations

In this thesis, we have used several machine learning techniques to generate feedback. The machine learning models were trained to understand the mental model of the experts, who labeled the data. However, like other machine learning models, our models are not fully accurate. This means the feedback could be wrong sometimes. We can not answer what happens when a user gets the wrong feedback. Providing false or deceptive feedback is another area of exploration (Costa et al., 2016). We have focused on a temporal machine
learning model – hidden Markov model. We did not explore any neural network-based models such as LSTM or recurrent neural networks. This is due to the fact that deep neural network models require a huge amount of data to work efficiently. However, in recent years there are applications where deep neural network was successful without having a large data set (Chen and McDuff, 2018). In the future, we plan to apply deep learning-based techniques to generate feedback in the hope of better accuracy.

We relied on several off-the-shelf software to sense the nonverbal and verbal cues. Although they perform very well they are not accurate all the time. Due to different lighting, background noise, and speaking pattern this software often fail to recognize. For example, the speech recognizer we used for SOPHIE was not trained on the late-stage cancer visit conversations. As a result, it often fails to recognize the medicine and treatment names. In the future, we to augment the medicine and treatment names with the speech recognizer so that it can recognize the words correctly.

Our finding of associations between trajectory styles and lecturing tendencies with prognosis understanding measures may not be causal. Our lecturing analysis was motivated from prior research that suggested that when a physician tends towards lecturing, it results in the patient not retaining as much of the information presented (Back et al., 2005). An alternative explanation could be that when physicians sense that patients do not understand, physicians are motivated to speak more, explain in greater detail, leading to a more lecturing-like structured conversation. Additionally, apparently passive patients may just lack understanding, which can result in poor engagement (i.e., patients may be too embarrassed or confused to ask for clarification), and this may result in conversations with a high LECT-UR score. In explaining the association of higher prognosis understanding with the dynamic sentiment trajectory style, we surmise that being dynamic keeps the patient more engaged and that ending on a positive note keeps the patient less depressed and more likely to remember the information just presented. However, again, an alternative anticausal
explanation could be that patients’ lack of prognosis understanding, and their physician’s perception of this, motivates the physician to speak in a calmer, less dynamic way (e.g., sentiment trajectory styles B or C). Additionally, the extent that the LECT-UR score correlates with human-annotated instances of "ground truth" lecturing should be investigated. However, it should be noted that despite any difference between the LECT-UR lecturing-like structure measure and human-labeled ground truth instances of lecturing, our results establish that the LECT-UR score serves as a useful metric in its association with patient prognosis misunderstanding.

Some limitations exist with regard to the bigger picture of SOPHIE-like virtual agents. Past research suggests that while conversing with a virtual agent or AI-driven conversational agent, humans tend to use shorter turns (Hill, Ford, and Farreras, 2015). This could be a limitation of using SOPHIE to train users to avoid lecturing since users might use shorter turns regardless of feedback. Our LECT-UR scoring method utilizes a window of consecutive turns that also includes the virtual agent’s turn. This allows the lecturing feedback to dynamically adapt to the conversation states and to the user’s behavior. We think that this can help circumvent the limitation posed by using feedback trained on the human-human conversation with a computerized dialogue system, though addressing this concern through a randomized study remains part of our planned future work.

8.5 Generalizability

We have performed a randomized control study to show the generalizability of the feedback. We show that the participants receiving feedback had performed better in a face to face conversations than those who did not receive feedback. However, this can not be said for the lecturing and sentiment trajectory feedback. This is due to the fact that we haven’t conducted a randomized control study to show the causality of the feedback. From our
analysis, we found that the not lecturing and use of certain sentiment trajectory is associated with better prognosis understanding. We can not conclusively claim that the lecturing and sentiment trajectory made a better prognosis understanding or they had better prognosis understanding which is why physicians chose to not lecture and followed a certain sentiment trajectory.
Chapter Nine

Conclusion

In this thesis, we have explored the feedback strategies to develop communication skills. We have designed feedback using expert opinions, and participatory design methods. We show how to adjust the feedback for different contexts and population groups to be more useful (addressing RQ1). Through user studies, we show that our feedback was able to help individuals demonstrate measurable improvements (addressing RQ2). We developed and trained several algorithms to generate feedback using the collected data set (addressing RQ3). In addition to the feasibility studies, we ran longitudinal studies to show the generalizability of our feedback in real-world scenarios (addressing RQ4).

In this thesis, we were able to explore different modalities of feedback and apply those in communication skills training programs. We have developed three communication training programs LISSA, Aging and Engaging, and SOPHIE. All of these programs were customized based on the target population group. The LISSA program was developed for those who may benefit from real-time feedback such as college students, the Aging and Engaging program was developed for older adults focusing on the post feedback, and SOPHIE was specially developed for patient-physician communication. Although we have touched a few applications we think that this thesis opens up possibilities for many other applications, such as training customer care agents, job interview training, group interaction training, etc.
The online communication training is particularly helpful when access to such training is not available. For example, in many low and mid-income countries such as countries in sub-Saharan Africa, the training for palliative care doesn’t exist Plas, Benjamens, and Kruijff, 2020. Our training programs could be a low-cost option for these communities where the training is not available at all. Another example of the use case could be training communication skills when physical distancing is required. During the period of social distancing to fight the COVID-19 many schools were shut down and many chose to deliver the lessons online. This affected nearly 50 million school students 1. In this scenario online training tools could be extremely useful for those who would suffer due to lack of social interactions.

References


Baio, Jon (2014). “Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2010”. In:


Bowie, Christopher R. et al. (2008). “Predicting Schizophrenia Patients’ Real-World Behavior with Specific Neuropsychological and Functional Capacity Measures”. In: Biological


Elias, Cezanne M et al. (2017). “The social and behavioral influences (SBI) study: study design and rationale for studying the effects of race and activation on cancer pain management”. In: BMC cancer 17.1, p. 575.


Fiorella, Logan, Jennifer J Vogel-Walcutt, and Sae Schatz (2012). “Applying the modality principle to real-time feedback and the acquisition of higher-order cognitive skills”. In: Educational Technology Research and Development 60.2, pp. 223–238.


Hahn, Elizabeth A et al. (2010). “Measuring social health in the patient-reported outcomes measurement information system (PROMIS): item bank development and testing”. In: Quality of Life Research 19.7, pp. 1035–1044.


Hoerger, Michael et al. (2013). “Values and options in cancer care (VOICE): study design and rationale for a patient-centered communication and decision-making intervention for physicians, patients with advanced cancer, and their caregivers”. In: BMC cancer 13.1, p. 188.

Hopkins, Ingrid Maria et al. (2011). “Avatar assistant: improving social skills in students with an ASD through a computer-based intervention”. In: Journal of autism and developmental disorders 41.11, pp. 1543–1555.


Kinzie, M B et al. (2002). “A user-centered model for web site design: needs assessment, user interface design, and rapid prototyping”. In: J Am Med Info Assoc. ISSN: 1067-5027. DOI: 10.1197/jamia.M0822.s.


Kron, Frederick W et al. (2017). “Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial”. In: Patient education and counseling 100.4, pp. 748–759.


McCurdie, Tara et al. (2012). “mHealth consumer apps: the case for user-centered design.” In: *Biomedical instrumentation & technology / Association for the Advancement of Medical Instrumentation*. ISSN: 08998205. DOI: 10.2345/0899-8205-46.s2.49.


Park, Kyung-Min et al. (2011). “A virtual reality application in role-plays of social skills training for schizophrenia: a randomized, controlled trial”. In: *Psychiatry research* 189.2, pp. 166–172.


Rock, Marcia L et al. (2009). “Can you hear me now? Evaluation of an online wireless technology to provide real-time feedback to special education teachers-in-training”. In: *Teacher Education and Special Education* 32.1, pp. 64–82.


Trower, Peter, Bridget Bryant, and Michael Argyle (2013). *Social skills and mental health (psychology revivals)*. Routledge.


Weisstein, Eric W (2004). “Bonferroni correction”. In:


