Physical Engagement As A Way To Increase Emotional Rapport In Interactions With Embodied Conversational Agents

Ivan Gris Sepulveda
University of Texas at El Paso, ivangris4@gmail.com

Follow this and additional works at: https://digitalcommons.utep.edu/open_etd
Part of the Computer Sciences Commons, and the Psychology Commons

Recommended Citation
https://digitalcommons.utep.edu/open_etd/1056
PHYSICAL ENGAGEMENT AS A WAY TO INCREASE EMOTIONAL RAPPORT IN INTERACTIONS WITH EMBODIED CONVERSATIONAL AGENTS

IVAN GRIS SEPULVEDA
Department of Computer Science

APPROVED:

__________________________________________
David Novick, Ph.D., Chair

__________________________________________
Nigel Ward, Ph.D.

__________________________________________
Olac Fuentes, Ph.D.

__________________________________________
Louis-Philippe Morency, Ph.D.

__________________________________________
Justine Cassell, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School
PHYSICAL ENGAGEMENT AS A WAY TO INCREASE EMOTIONAL RAPPORT IN INTERACTIONS WITH EMBODIED CONVERSATIONAL AGENTS

by

IVAN GRIS SEPULVEDA

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at El Paso
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science
THE UNIVERSITY OF TEXAS AT EL PASO
August 2015
Acknowledgements

Acknowledgments are boring (most of the time). I’ve read dissertations who acknowledge beer, a dog, a cat, or other inanimate objects, and others who list every friend they have on Facebook and Twitter with a generic “thanks for everything” or mention how they contributed with “their invaluable support”. I believe that you, the reader, and they, the people acknowledged, deserve something more. So instead of the traditional acknowledgment section, here’s a story that I hope will help me explain how grateful I am for all these people.

It was 2010 and I had just finished my M.S. in computer science. I had also decided against entering a Ph.D. When I looked at my options, it was either find a job and work for several years until the U.S. would accept me, if they ever did, or start my own thing. I had recently decided that academic life was over for me, but that I also didn’t want to work for years on anything just to become at least, and at last, a legal resident. Then, I met this curious and slightly hyperactive professor.

I was asking around, trying to figure out a way for a Mexican like me to open a business in the U.S. It turns out it’s even harder than getting a (dream) job that will sponsor you a visa. I had some ideas in mind, which I bounced with David Novick, my then professor for human-computer interaction. You have to understand, he is no ordinary professor. This is a former lawyer who then got a Ph.D. in computer science, is involved in entrepreneurship activities on campus, and lead the University’s improv acting group. Quite a character. I was unfamiliar with his work at the time, and he was busy with administrative tasks around campus rather than engaged in research.

I wanted to open a technical school that specialized in interactive technologies. He had an empty projection room with a sign in its door that said “Interactive Systems Group” and not much else. Something clicked. He was interested in virtual agents, and I was interested in computer graphics. When my startup idea (as 99% of all startup ideas) turned out to be a fiasco, we decided that, if I were to (hypothetically) do a Ph.D., embodied conversational agents would be a line of research we would both like to pursue. Although I wasn’t 100% convinced at the time, I figured it would at least buy me some time to make some career choices if it didn’t work out. In case you are wondering (spoilers) three years and a half later, this dissertation proves it did work. Why did it work? Because he had an interesting
principle: Whatever I did had to be fun. He would always ask me after one of our meetings “are you still having fun?”, and if by any reason I wasn’t, he would immediately propose a solution. And man, did I have fun.

I started working in an old Linux computer, reading research papers, getting a feel for the field. Everything was new and overwhelming. I was eager to get my hands on something more tangible. I would read all these stories about great researchers. Every time I had an idea it turned out that someone had already worked on it years before I even got out of high school. This was done mostly by important, unreachable professors who worked at ICT (USC Institute for Creative Technologies) or Carnegie Mellon with these fantastic agents and incredible technologies.

Why were they unreachable? During my first year at UTEP, coming from public schools in Mexico, everything was incredibly amazing. I would be impressed by what other people might not even notice. Ever since I started my bachelors, I would look into these unreachable places. I had no contacts there, but I was determined to do my master’s in either of them. I went as far as going with one of my best friends, Cesar, to conferences all over the country. We would ride a bus for 17 hours (UTEP is in Texas, right next to the border with Mexico, which meant it was far from anywhere else), sleep in the cheapest motel we could find, and go into these conferences. In there we would talk with recruiters, researchers or graduates from both places, and Cesar, always the realist yet encouraging, kept me going. Although he and the recruiters gave me a lot of hope, my financial possibilities and my nationality (international student tuition and being unable to apply for work permits while studying) won in the end, and I wasn’t able to study there. I did not give up though, and through the years I did my best to get an internship in both places. I never did.

Fast forward to where I am once again reading research papers, feeling dumb for taking too long to truly digest their content. I started working by myself in this lab. I had no idea how the very complex agent systems were made, and had very low expectations on what I could achieve. Then, people started joining the lab, many as volunteers, just because they thought it was a fun thing to do. First came Anuar Jauregui and Joel Quintana, along with two French cadets, Guillaume Adoneth and David Manuel, who interned for a summer. A little later, Mario Gutierrez showed up. With their help and plenty of
frustration episodes, we built our first agent, from scratch. Neither of us knew what we had truly accomplished. It obeyed commands, moved horribly, and had the most annoying synthesized voice you can possibly imagine, yet, it worked. With this, my proposal came along.

With the help of two other peculiar professors, Nigel Ward and Olac Fuentes, it began to take shape. Nigel attended each one of our weekly research group meetings and all of my presentations, where he would give me the hardest time by asking me the tough questions. I came to appreciate this (at a much later time). As it turns out his critique was largely responsible for my project being feasible, realistic, and no question I have ever been asked in a conference has been more difficult than those he asked followed by his traditional I’m-not-so-convinced glance. Olac, an expert in computer vision and machine learning gave me the smartest ideas to improve our gesture recognition, and as the graduate advisor of the department, some earnest post-graduate advice. One of the things I valued the most though was his confidence in my work and my progress, he really kept me motivated, even if it was just by making a small comment when we passed on the hallway, the little details added up; nothing he ever proposed was too difficult for me in his eyes.

When I was looking at who would be part of my committee, my mentor, David, asked me who would be my top two choices in the whole world. I had read the papers, and narrowed down the list to two people, Louis-Philippe Morency, the most cheerful French-Canadian researcher you will ever meet, who used to be at ICT back then, and Justine Cassell one of the first and most important researchers in embodied conversational agents, who was at Carnegie Mellon. David just said “well, you should ask them”. I did, again, with exceedingly low expectations. They were both researchers who were far away and had never heard of me, not to mention extremely busy. Imagine my surprise when they both said yes! At this time I knew I was truly a Ph.D. student, as I was starting to have academic role-models.

A little while after that I successfully defended my proposal. Sometime later, Mario graduated and jumped into the master’s program, and Diego Rivera, Carolina Camacho and Alex Rayon joined the team. I didn’t know it yet, but they would all become some of my closest friends. Instead, I was mostly freaked out. It was like starting again with new people. There was nothing to be afraid though. These guys turned out to be the most helpful, smartest, idealistic people I’ve ever had the pleasure of working
with. Together we spent an additional year and a half developing a much stronger agent. I ended up presenting my (or more accurately, our) work in front of many researchers I had read so much about, in conferences I had never dreamt of attending (when I was younger I would pay to be able to sneak in a subset of sessions and meet people).

As our lab became more popular, more people joined. Now, three years and a half later, Caro began her work as a Ph.D. student and became my left hand at the lab (I’m mostly left-handed) helping me lead and organize the now very large team. She also helped me improve and create additional modules for the gesture recognition tool, and pulled quite a few all-nighters while we worked on our agents, doing the voice and the motion capture for them (you should know she has voiced over 1000 dialogs for several agents), all of her own initiative. Although she does not see herself as a risk-taker, you should have seen her when our whole group went skiing. I am sure you would disagree as you fear for her life.

Alex started his master’s, but not before creating tons of scenery, scripts and helping design truly immersive interactions for our agents. It might not sound like much, but if you really must know how much he helped, take a look at Appendix A, it wouldn’t exist without his work. His personality would always light up the lab and fill it with energy, reminding us of the really exciting work we were doing when nothing seemed to compile.

Diego went to intern at ICT twice. He learned so much and helped improve our technical infrastructure to a point where we could have agents with features the rest of us hadn’t even thought about. He never wondered or asked if something was even possible, instead he would just show us it was after he had done it. He is the type of guy that will say yes to every challenge you throw at him, and silently work it out through the middle of the night so that it works by the next day (even if we break a couple of other things in the process). He was also the first to jump on board with Inmerssion when I pitched the idea to the team, our virtual agent startup.

Mario graduated as we neared the completion of our project, but he left us with a warm memory of him, and has kept in touch with the group, always wondering if we finally made the agent work. We did, and he is a big part of the reason why.
Then, even more wonderful people joined our lab. A friend from high school, Laura Rodriguez began her master’s here and helped me analyze tons of data (I found it a bit weird how much she enjoyed that process). One of the most joyful and helpful persons I’ve met, with her inquisitive mind I know she will become a great researcher. She also organized and executed pilot tests, usability test, and playability tests on the system before launch. As a result, one of her newly acquired skills was the ability to memorize and paraphrase all game script dialogs.

Tim Gonzales is like a ghost in the lab, but friendly. He would appear when you needed him the most and volunteer for anything he noticed we were falling behind on as a group. He would surprise me more than once with some very sharp observations. Unfortunately his project drifted away at some point from my main research focus, but, along with Alex, he provided some of the base work that made us look into personality in the first place.

During my last months of the Ph.D. I have enjoyed the help and company of even more incredible people. Jaqueline Brixey, who used to be a little scary at first, turned out to be very friendly once I got to know her. She volunteered to read the first draft of this dissertation, and is probably the best advisor-illegible-scribbles interpreter at the lab, fixing quite a few typos along the way.

Then there is Alfonso Peralta, a guy who most people underestimate. As one of the newest and less experienced members of the team, he is afraid to make mistakes, but given the chance (and maybe a slight push), I’m sure he will go far. He helped, along with everyone else to run 80 subjects in just two weeks, from recruiting, to running the experiments, which is not as easy as you might think.

Victoria Bravo, although she might appear a bit clumsy, did a great job at creating some of the dialog trees that allow our agent to talk. When she jumped at the opportunity to help with that, she truly had no clue about what to do, but that didn’t stop her, and it showed what a quick learner she is. Since then she has safeguarded our code against the monsters that destroy backups (by creating a repository) and helped Paola.

Paola Gallardo joined the team during my last few months there. With her cheerful personality and very contagious laugh, along with some serious (and I mean serious) debugging skills, she helped
refactor thousands of lines of code that took us years to write in just a few weeks, making our agents much more reliable and robust.

Finally Yahaira Reyes was one of the latest additions to the team. She has helped recreate our whole animation system, by introducing hundreds of new animations and controllers for speed and amplitude of gestures. She also did this much faster than what we expected.

Finally, although they didn’t work directly with me, some other great friends whom I’ve shared the lab with deserve at least an honorific mention: Alejandro Vega, Shreyas Karkhedkar and Steve Werner were the first guys I met at the lab and made my transition easier. Saif and Chelsey helped the whole team by giving constant feedback and bouncing ideas with us, and Igor, who during the brief time he was at our lab became excited about our project and offered to help or collaborate with us in the future.

As you can see, this was not a solo mission. At this point you are probably tired and wish I would conclude with a “thank you all” note (unless you haven’t been mentioned yet). You should know that if I ever complained, discussed, talked about, or mentioned my research in your presence, you have helped and I am thankful.

I would like to close this section with a few special cases of people who did not share the lab with me, but have helped me nonetheless throughout the years: Salvador Ruiz, for keeping me always excited, amazed and motivated. He’s always been available 24/7, and because of that, we have met the strangest people and discovered hidden places that have filled our memory with one-too-many funny stories and adventures. Laura Berrout for her help on early projects and finding everything I do exciting. Patricia Jara for giving me a wonderful godson and some great real-life perspective on what I am doing.

You also need to know about Marianna, the loveliest and prettiest girl I’ve ever met, with whom I have shared my life almost since I started the Ph.D. She has kept me sane and cheerful throughout the process. Even now as I write this paragraph at 4am and I realize that this has been our life for a few months now, she has never complained. Quite the contrary, she has made the best of whatever time we have, and constantly pushes me to finish, to do better, and to follow my dreams wherever they might lead. Marianna, I love you. Thank you.
Of course this wouldn’t be complete without mentioning my parents and young sister. Aside from the traditional support that families provide, they have actually listened to my theories and hypotheses countless of times, contributing to interesting discussions and helping me in many aspects that led to the development of this project. Most of all, they kept my hopes high and my feet on the ground, sharing and cheering every victory, and truly being involved in my life and my work.

Thanks to all of you, I have finally found an answer to a question that haunts most of us for our entire lives: *What am I going to do next?* Have fun. It has worked quite well so far.
# Table of Contents

Acknowledgements ........................................................................................................... iv

Table of Contents ............................................................................................................... xi

List of Tables ...................................................................................................................... xiv

List of Figures ..................................................................................................................... xv

Chapter 1: Introduction ...................................................................................................... 1

Chapter 2: Background and Related Work ...................................................................... 4

2.1 Human-Agent Interactions ......................................................................................... 4
2.1.1 Commercial Agents .............................................................................................. 4
2.1.2 Mediator Agents .................................................................................................. 5
2.1.3 Direct Agents ...................................................................................................... 7
2.1.4 Multiparty Agents ............................................................................................... 8

2.2 Human Rapport Building Behaviors ......................................................................... 9
2.2.1 First Impressions ............................................................................................... 10
2.2.2 Familiarity .......................................................................................................... 11
2.2.3 Relational Theories .......................................................................................... 13
2.2.4 Rapport .............................................................................................................. 19
2.2.5 Rapport across Time ......................................................................................... 21
2.2.5 Nonverbal Rapport ........................................................................................... 22
2.2.7 Mimicry ............................................................................................................. 23

2.3 ECAs Designed for Rapport ...................................................................................... 25
2.3.1 Social Presence and Awareness ........................................................................ 25
2.3.2 Anthropomorphism ............................................................................................ 26
2.3.3 Storytelling ........................................................................................................ 26

2.4 Virtual Rapport ......................................................................................................... 27
2.4.1 Rapport Building Agents ................................................................................... 30

2.5 Engagement ............................................................................................................. 32

2.6 Physical Engagement to Increase Rapport ............................................................... 33

Chapter 3: Experimental Design ..................................................................................... 35

3.1 Experiment ............................................................................................................... 35
3.2 Physical Connection Enhanced Agent ...................................................................... 37
### Chapter 3: Rapport in Human-Agent Interactions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>Verbal Behaviors</td>
<td>39</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Paralinguistic Behaviors</td>
<td>40</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Gesture and Non-Gesture Conditions</td>
<td>41</td>
</tr>
<tr>
<td>3.3</td>
<td>Measuring Rapport</td>
<td>42</td>
</tr>
</tbody>
</table>

### Chapter 4: The Agent

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Agent Development and Tools</td>
<td>43</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Agent Architecture</td>
<td>43</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Markup Language Definition</td>
<td>45</td>
</tr>
<tr>
<td>4.1.3</td>
<td>SAIBA, BML and FML</td>
<td>48</td>
</tr>
<tr>
<td>4.1.4</td>
<td>Speech and Language Processing</td>
<td>50</td>
</tr>
<tr>
<td>4.1.5</td>
<td>Gesture Recognition</td>
<td>52</td>
</tr>
<tr>
<td>4.1.6</td>
<td>Gesture Capture</td>
<td>54</td>
</tr>
<tr>
<td>4.1.7</td>
<td>Animations</td>
<td>55</td>
</tr>
<tr>
<td>4.1.8</td>
<td>Animation Principles</td>
<td>58</td>
</tr>
<tr>
<td>4.1.9</td>
<td>Animation Blend Tree</td>
<td>62</td>
</tr>
<tr>
<td>4.1.10</td>
<td>Virtual Environment Design</td>
<td>66</td>
</tr>
<tr>
<td>4.1.11</td>
<td>The Agent’s Voice</td>
<td>66</td>
</tr>
<tr>
<td>4.2</td>
<td>Agent Features and Design</td>
<td>67</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Interaction and Storytelling Design</td>
<td>67</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Agent Capabilities</td>
<td>71</td>
</tr>
</tbody>
</table>

### Chapter 5: Data Collection

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Verbal Annotation</td>
<td>72</td>
</tr>
<tr>
<td>5.2</td>
<td>Nonverbal Annotation</td>
<td>72</td>
</tr>
<tr>
<td>5.3</td>
<td>Demographics</td>
<td>75</td>
</tr>
<tr>
<td>5.4</td>
<td>Personality</td>
<td>75</td>
</tr>
</tbody>
</table>

### Chapter 6: Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Rapport Scores Evaluation</td>
<td>76</td>
</tr>
<tr>
<td>6.2</td>
<td>Gesture Evaluation</td>
<td>78</td>
</tr>
</tbody>
</table>

### Chapter 7: Conclusion

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Contributions</td>
<td>84</td>
</tr>
<tr>
<td>7.1.1</td>
<td>Physical Engagement as a Way to Increase Rapport</td>
<td>84</td>
</tr>
<tr>
<td>7.1.2</td>
<td>Multimodal Information in Human-ECA Interactions</td>
<td>85</td>
</tr>
</tbody>
</table>

xii
7.1.3 Behavioral Analysis of Virtual Rapport ........................................85
7.2 Limitations and Future work ..........................................................86
  7.2.1 Attitude Gestures in Virtual Rapport .........................................87
  7.2.2 Shared Interpersonal Knowledge in Virtual Rapport .................88
  7.2.3 Personality as an Indicator of Rapport ......................................88

Bibliography .........................................................................................89

Appendix A: Scene Scripts .....................................................................98
Appendix B: Interaction Results ...........................................................121
Appendix C: Demographics .................................................................123
Appendix D: Agent Scripting Language Schema ..................................126
Vita ........................................................................................................129
List of Tables

Table 1: Rapport Instrument. Subjects indicated agreement or disagreement on a five-point scale........42
Table 2: Agent scripting language definition ..........................................................................................48
Table 3: Data from voice analysis. ...........................................................................................................67
Table 4: Gesture IDs used to graph the interaction timelines. IDs are unique but do not follow a particular sequence. Missing ID numbers are due to animations or gestures that were not used for this experiment. The agent gestures table (left) contains gestures exclusive to the agent. ........................................74
Table 5: Summary of personality types based on the MBTI ......................................................................75
Table 6: Non-gesture and gesture condition average scores across all participants. The header row initial stands for the dimension of rapport the item evaluates (Understanding, Physical, and Coordination). A positive number on the final row indicates an overall preference for the gesture agent while a negative number indicates preference for the non-gesture one. ..................................................................................................77
Table 7: Total number of gestures per participant. The full table containing gestures per scene per participant can be found in Appendix B. .............................................................................................................78
Table 8: T-test results for different condition comparisons ........................................................................79
Table 9: Standard deviation and average gesture rate among different personalities and conditions, where G is the gesture condition, N is non-gesture condition, I stands for Introvert and E indicates an extrovert personality trait ........................................................................................................................................80
Table 10: Comparison between introverts and extroverts under the gesture condition. ..............................80
Table 11: Comparison between introverts and extroverts in the non-gesture condition ................................80
Table 12: Number of participants under each personality type, under both conditions and their cumulative gestures ........................................................................................................................................81
Table 13: Normalized average gestures in gestures per across all personality types for the gesture and non-gesture conditions ..................................................................................................................................................81
Table 14: Survey items as seen by participants and their respective ID. ........................................................121
Table 15: Non-gesture condition questionnaire results per participant without reverse coding. All values represent positive aspects of the agent ..............................................................................................................121
Table 16: Gesture condition questionnaire results per participant without reverse coding. All values represent positive aspects of the agent .................................................................................................122
Table 17: Demographics including gender, age, college level, and first, second and third languages. ..........123
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lisa is an example of a very simple commercial agent for National Rail.</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Example of state-of-the-art NPC interaction with multiple characters facing the player’s avatar as a result from previous choices. The image shows gameplay from The Witcher, a game characteristic for its storytelling and character interactions. These mediator agents are not threatening the user directly, but rather the virtual representation of it.</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Ada and Grace explore a complex multiparty setting, in which multiple agents interact between them, and with an audience (Swartout, et al.).</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>First version of the first ECA built at UTEP’s Immersion lab.</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>Second version of the agent with improved playability and usability, and a recorded voice.</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>Importance of the three components of rapport across time (Tickle-Degnen, Rosenthal, 1987)</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>Familiarity timeline expressed as a relational model.</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>Conventional presence, social presence, and virtual presence as components of awareness.</td>
<td>26</td>
</tr>
<tr>
<td>9</td>
<td>(Zhao, et al., 2014) Dyadic state, strategy and action repertoire</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>Paralinguistic rapport model</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>Experimental setup. The wall is a frontal projection, meaning that getting closer to it would create a shadow. There is a camera hidden on the tree to the left, but users are informed and consent to being recorded. The Kinect is located straight in front of the user at the bottom center.</td>
<td>36</td>
</tr>
<tr>
<td>12</td>
<td>Facial capture setup, unfortunately unusable in this setting.</td>
<td>37</td>
</tr>
<tr>
<td>13</td>
<td>High-level system architecture of the ECA system.</td>
<td>44</td>
</tr>
<tr>
<td>14</td>
<td>Detailed architecture of the software interfaces and feature handling.</td>
<td>45</td>
</tr>
<tr>
<td>15</td>
<td>An FML example which models the intentions and behaviors for the phrase “give him some gold”.</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>This illustrates the complexity and versatility of the language, but it also shows how it makes it impractical to use in a 40–60 minute interaction system, for which a greater degree of abstraction would be preferable.</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>A scene section using our markup-language. This script causes the agent to look thoughtful while asking a question (through a predefined audio file). It then pauses for a second to prevent the speech recognizer from listening to the playing audio file. Finally, a grammar is created inside a decide tag, which enables users to progress to the next episode or scene depending on their verbal expressions.</td>
<td>51</td>
</tr>
<tr>
<td>17</td>
<td>Graphs of episodes and transitions showing different styles of dialog paths.</td>
<td>52</td>
</tr>
<tr>
<td>18</td>
<td>In this scene the agent uses a combination of animations to extend its hand to the user. The agent then prompts the user to grab her hand through a decide tag. With only one option, it is a required gesture to progress through the interaction. After the user performs a “Grab Hand” the control flow decides on the next episode to execute. At the end, the scene changes. On the virtual side, a new environment is loaded.</td>
<td>56</td>
</tr>
<tr>
<td>19</td>
<td>A section of the animation tree detailing the permitted locomotion transitions for lower body movement. The labels of the animations differ in style (e.g., “Running” vs. “Jump”) because some animations are continuous loops (e.g., running, walking), and others are one-time movements (e.g., jump, stumble).</td>
<td>63</td>
</tr>
<tr>
<td>20</td>
<td>From left to right, video and depth noise from the participant, motion capture data, and agent with animation data incorporated. The top right displays a thumbnail view of the raw front and side video as processed with iPi Soft.</td>
<td>64</td>
</tr>
</tbody>
</table>
Figure 21: To the left, the list of layers affecting different body areas. These layers can be blended with any or all others with the exception of the special animation. To the right, the list of parameters that characterizes the current running animation.

Figure 13: Annotation of nonverbal behavior during the interactions. The red line represents the agent and the blue one, the user.

Figure 14: Gesture rates across conditions, where A represents gesture-extrovert, B non-gesture-extrovert, C non-gesture-introvert, D gesture-introvert.

Figure 23: Demographics survey.
Chapter 1: Introduction

One of the major goals in research on embodied conversational agents (ECAs) is to increase the believability and perceived trustworthiness of agents. This is often accomplished by creating human-like virtual characters that follow social conventions. This careful interaction design usually leads to other favorable characteristics in agent interfaces, such as increased user engagement. The term embodied conversational agent refers to a form of human-computer interaction, represented by intelligent agents that exist in a virtual environment and communicate through elaborate user interfaces. Graphically embodied agents can take almost any form, often human-like, and aim to unite gesture, facial expression and speech to enable face-to-face communication with users, providing a powerful means of human-computer interaction (Cassell, J., 2000).

Human-ECA communication is a preferred method of interaction, when properly designed, for certain applications. This is particularly true when agents play roles such as healthcare assistants or instructors, where the anthropomorphic representation of the agent increases the long-term perceived trustworthiness of the system (Bickmore et al., 2008).

One distinctive advantage of ECAs over traditional interfaces is that they enable users to communicate in parallel with the system; for example, the user can concurrently convey paralinguistic information through gesture and provide verbal instruction to the agent. Depending on the interface design, it may also enable the user to perform additional nominal tasks (Kipp et al., 2006) because the human and ECA are engaged in face-to-face interaction.

To improve the efficacy of the interaction between humans and ECAs, I focus on the development of virtual rapport through paralinguistic behaviors. Rapport is a complex sense of connection, affinity, and mutual understanding that most of us can attain when interacting with other people, but the process by which we build it is difficult to model. Moreover, rapport-building behaviors are not always successful, and certain actions cannot only hinder rapport, but even destroy it. This feeling of connectedness can vary in intensity and is developed at different rates for different people, and, because it is subjective and difficult to describe, it is also hard to measure. Although this feeling occurs in every normal human interaction to some extent, it has to be carefully designed to affect people
in human-agent interactions. This study attempts to integrate rapport in human-agent interactions and improve human-ECA interfaces by increasing engagement.

ECAs are already deployed in multiple platforms and devices, and most people who interact with computers on a regular basis have been exposed to them. From the infamous Microsoft Office assistants, to the “live” chats several companies offer on their websites, ECAs already provide commercial value for customer-service applications. Although these relatively simple applications might benefit from rapport-building behaviors, more advanced ECAs that make use of personality traits can affect us at a personal or emotional level. Agents that are capable of building rapport can perform their tasks more efficiently. For example, although several ECAs have been built for learning and education, rapport-enabled agents could eventually lead to the development of virtual characters that specialize in teaching a particular subject, can detect boredom, lack of attention, or misunderstandings all signs of lack of rapport between user and agent.

To do this requires agents that go beyond attempting to perform the right behaviors at the appropriate times. Rather, these agents are reactive, which means they adapt to the user. If we transpose this situation to a human-human interaction, a non-rapport-building relationship would be dominated by one of its participants, where one gives everything to “fit in” while the other’s behavior completely disregards the signals from their conversation partner. In contrast to a symbiotic relationship that cultivates rapport over time, this is effectively a simulation of an abusive relationship. With this argument I am not attempting to defend ECAs’ rights in a world of abusive users; instead, this example reflects that certain social tasks cannot be accomplished in this manner. In the instructor ECA example, an agent that is supposed to aid users learn can perform more effectively if it gathers information that helps it elicit attention or is capable of adapting its behavior if it can detect a lack of interest in the material being presented.

Creating agents that elicit nonverbal behaviors that are not required for task completion can be problematic. Most users will attempt to finish their tasks with the least amount of work, which is usually fine for agents that have a defined and measurable task goal but not so much for subjective, longer-term goals. So far, agents have been built to assist in short-term goals rather than lasting interactions
with a few exceptions. One of the main challenges in moving towards long-term interaction is to maintain a sense of believability and awareness, that is, creating the illusion that the agent is perceptive of the user’s behavior at all times, not only when input is required. This in turn means that all user actions, conscious and unconscious, would ideally be accounted for and properly reacted to.

The goal of this research is to improve the naturalness and effectiveness of ECA interactions by increasing rapport between agents and humans. I posit that eliciting domain-specific gestures will motivate users to contribute physically, and that increased physical activity by the user will improve human-ECA rapport.

In the following sections I explain one process by which we can design interactions and agents to reach this goal. First, I review the state of the art in ECA development and rapport. This includes the current models that attempt to define individual dimensions of rapport, and a new comprehensive framework that connects previous models to observable paralinguistic behaviors. I also present previous studies I have conducted related to familiarity and rapport, after which I present my claim and hypotheses. I explain the methodology and experimental setup, including the interaction design, with their respective results. I also explain our ECA system, with an emphasis on its speech and gesture recognition modules, and I provide an overview of its architecture. Finally, I conclude with a discussion of possible future lines of work and interesting and unexpected results.
Chapter 2: Background and Related Work

To understand human-ECA relationships, one must first understand human-human relationships. In the following subsections, I review different rapport definitions and combine them into a model that emphasizes two-way rapport interaction through nonverbal behavior. In addition, this section will explain several social traits that are rapport-enhancing or are a consequence of rapport, such as familiarity, and several relational models. Finally, I present key differences between human-human rapport and human-ECA rapport.

2.1 Human-Agent Interactions

To facilitate the study of ECAs and generalize some of the particular features of different agents, I classified them according to their representation, features, and purpose into four main categories: commercial, mediator, direct, and multiparty agents.

2.1.1 Commercial Agents

Commercial ECAs improve the customer service experience either by presenting the company’s information in a more attractive manner or providing automated customer support. The main advantage of this approach is the possibility of uninterrupted service and a reduced workload for the human operators. These ECAs do not need realistic behavior; conversations are usually scripted and take the form of “frequently asked questions” instead of dynamic, unscripted conversation. Their representations are usually human-like 2D images.
As simple as they are in comparison to other agent types, these ECAs have been shown to significantly reduce customer service expenses in large companies and one day may become the norm. For the time being, however, they require improvements in their particular domain knowledge; that is, improve in question understanding, analysis, and search, rather than engage in relationship-building behaviors, as they are mostly single-use purveyors of information.

2.1.2 Mediator Agents

Another thriving area that focuses on improving virtual agents is the $34.7 billion market of the videogame industry (Bond, 2008), where ECAs are the key element in enhancing storytelling and creating more immersive player interactions. In videogames, ECAs usually represent characters that the player encounters across the flow of the game, also known as non-playable characters (NPCs). These are usually articulated humanoid characters that represent the participants in in-game social interactions, regardless of whether they are agents under the control of the game or avatars being controlled by human players.

Because the player is represented as an avatar and ECAs interact with the player indirectly by responding to the avatar’s actions and not the player itself, these interactions are mediated. In other
words, the player controls a virtual character, and all interactions usually occur with that virtual character. This mediated type of interaction poses usability and playability challenges (Desurvire, 2004), and interaction mechanics are limited to the affordances of the avatar that the player controls. However, this sort of agent offers a range of possibilities that agents that interact directly with their human counterparts cannot yet achieve.

Figure 2: Example of state-of-the-art NPC interaction with multiple characters facing the player’s avatar as a result from previous choices. The image shows gameplay from The Witcher, a game characteristic for its storytelling and character interactions. These mediator agents are not threatening the user directly, but rather the virtual representation of it.

Despite all their limitations, there have been interesting developments in next-gen games, which at the time of this publication refer to videogames published to PS4 or Xbox One. These games have created remarkable in-game character interactions. For example, games like the Witcher (Figure 2) or Mass Effect include hundreds of multiple-choice dialogs with automated gesture generation for both parties, making it appear closer to a cinematic experience with certain degree of control than a game (Tomaszkiewicz, et al., 2015; Hudson, et al., 2012). Along those lines, games like the ones in the Fable series (Molyneux, et al., 2004) place great emphasis in the NPC’s personality and actions at the social
level by changing the character’s verbal and nonverbal responses depending to the avatar’s physical characteristics and previous actions.

Although some of these elements can be borrowed for a more personal style of interaction, where the characters speak directly to their human counterparts, it is exclusively the direct, dyadic interaction explained in the following subsection is the focus of this research.

2.1.3 Direct Agents

Direct agents are those that engage in a first-person interaction with the user. These agents are lifelike characters that frequently use cutting-edge technology to keep track of user’s real-world physical actions, sometimes including voice recognition, facial recognition, full-body recognition, gaze detection, haptics, or any combination of those. These agents can have a wide variety of purposes, but they are most frequently used in training scenarios or simulation. Although they are far from replacing their human counterparts in some aspects, this type of agent is an appropriate alternative for areas where there is a low availability of real people to perform such tasks.

These ECAs are highly complex and are usually developed to serve problems with a representative social impact. Additional examples include military training, computer-assisted speech and language tutors for hard-of-hearing and autistic children (Bosseler and Massaro 2003), and many other applications. For example, these sorts of ECAs have been employed as museum guides, instructors, and in cross-cultural social training. Public museums have employed ECAs as guides, where they engage with visitors in natural face-to-face communication while providing information about the exhibits (e.g., Swartout, et al., 2010). In classroom settings, ECAs with a variety of gestures and facial expressions have been implemented in small groups, and proven to increase attitudinal and procedural learning (e.g., Baylor and Kim, 2008).

Unfortunately, even when agents these agents show great potential, the long-term interaction component and the hardware and technical skills necessary to deploy ECAs in real-life settings limit their use. In addition, these agents often require Wizard-of-Oz components that are difficult to automate.

During the last few decades, ECA technology has overcome some of the basic obstacles of rich human-agent interaction, including speech recognition, text-to-speech, and character three-
 dimensionality and animation. Immersive applications, such as ECAs, are not only perceptual but highly interactive and require user action.

Interfaces with rich behavior, such as ECAs, present additional complexities for design and development. In particular, although the functionality provided by speech-enabled and kinetic aware interfaces is impressive, the interfaces used to interact with these agents and access their functionality are often inconsistent and imprecise (e.g., motion trackers jitter or misrecognized speech) (Wooldridge and Jennings, 1998).

Unfortunately, it is this emphasis in naturalness and expression in ECA design that restricts developers from the traditional user interaction metaphors so common in interface development (e.g., a pair of scissors represents the function cut in most word processing documents, but it can be confusing if such scissors appeared in the hand of an agent). Because of this, an ECA should be designed to interact without the need of most of the traditional interface elements --that is, humans should interact with ECAs as natural as possible and preferably without explicit traditional interface elements such as buttons, text boxes, or point-and-click items (unless they are part of the agent’s tasks and goals, for example, a training agent for new software). Instead, ECAs should detect and simulate complex behaviors that encompass a combination of non-verbal cues such as gaze, gestures and mimicry, and verbal feedback such as backchannels, in conjunction with context specific content to create a shared state of understanding between human and ECA. This multimodal interaction in everyday life seems effortless, but a closer look reveals that such interaction is indeed complex and comprises multiple levels of coordination, from high-level linguistic exchanges to low-level couplings of momentary bodily movements (Zhang, et al.).

2.1.4 Multiparty Agents

Before going into detail with rapport and its associated nonverbal behaviors, I should mention one final agent category, multiparty agents. These agents are an improved version of the direct agents. In addition to possessing realistic and natural behaviors and interacting in first person by directly sensing the user, they are also capable of interacting with other agents or with more than one user simultaneously. A notable example in this category are Ada and Grace (see Figure 3).
These agents are so complex that only a handful have been made, most of which are only partially automated and rely on a series of tricks to make them work, such as push-to-talk or Wizard-of-Oz (Traum and Rickel, 2002; Plüss et al., 2011). Although these ECAs are not explored in this present work, they are a likely source for future research.

2.2 Human Rapport Building Behaviors

One of the main goals of ECA development and research is to increase the believability and perceived trustworthiness of agents and the user's engagement with the system --in other words to create ECAs that follow social conventions, similar to those in natural interaction (Bickmore and Cassell, 2001). One way to increase rapport during human-ECA communication is the use of natural non-verbal behaviors.

To create and apply virtual rapport requires understanding inter-human rapport. Rapport is not an individual trait but rather a collective combination of qualities that emerge from each individual during interaction (Tickle-Degnen and Rosenthal, 1987). One generalized definition of rapport is the feeling of mutual understanding --the connection and harmony experienced when two people are engaged in conversation (Huang, et al., 2011), or as it is often informally described, the feeling of being “in sync”.
Everyone has experienced some level of rapport in their daily interactions. From the casual greeting between a supermarket employees, to the long midnight conversation with close friends, rapport is a critical component in any communicative action. It is this set of indicators that enables humans to drive the conversation, to make friends, to flirt, or simply to accomplish a task.

It can be argued that an interaction without rapport is simply not possible or not a true interaction, as even the slightest unintentional communication act carries some mutual understanding or requires at least a trace of attention to be perceived, both important components that define rapport, which will be extensively discussed in this chapter.

Given its importance, this complex behavior has been extensively studied, and many independent models evaluating stand-alone traits such as intimacy or familiarity have been presented as specific and constrained ways to analyze conversation, interaction, and relationships in general.

In the next section, I explain, evaluate, and connect those independent, in-depth models that define particular aspects of the interaction and explain how they relate to different elements of rapport. After analyzing how those models work in human interactions, I translate those elements into ECA behaviors.

2.2.1 First Impressions

Given the natural progression of any relationship, it seems logical to start analyzing rapport at the start of the interaction. Although behaviorally very few things happen, at the cognitive level most of our initial assumptions and evaluations of other people occur when we first meet. In fact, judgments of attractiveness, likeability, trustworthiness, competence, and aggressiveness can be defined in less than 200ms of exposure to someone’s face (Willis and Todorov, 2006).

These impressions may become strong factors for the rest of the relationship, and, unfortunately, appearance does matter and can influence rapport. Some older studies (Kelley, 1950) began exploring how we are affected by stereotypes. Although some first impressions are culturally determined, the important finding is that they are accompanied by the human expectations of the interaction.

These results are cautionary. As ECA developers, special attention needs to be placed in the graphical design of characters, from their ethnicity to their attire.
There is also an effect on first impressions when interaction is involved, even in human-ECA interactions (Caffaro, et al., 2012), where after only twelve seconds of observation of the ECA’s non-verbal cues (smile, gaze and proximity) subjects formed impressions and judged its personality (extraversion) and interpersonal attitude (hostility/friendliness).

Although first impressions are only the spark that ignites rapport, that is, a brief, fleeting moment that serves as the entry point to any interaction, it is not the main focus of this research. Regardless, first impressions are a valuable tool for rapport and were carefully accounted for in our ECA’s appearance, initial behavior and dialog design, described in detail in a later section.

2.2.2 Familiarity

Perception and interaction can change over time. This phenomenon is known as familiarity. Human-agent communication cannot yet achieve the naturalistic and spontaneous communication that humans do unconsciously; familiarity-enabled ECAs are a step towards a more naturalistic human-agent conversation. Familiarity is the sense of knowing someone, usually built across time and during more than a single conversation.

Familiarity is a key component for building confidence and trust in cooperative relationships (Luhmann, 2000). Previous studies performed at UTEP’s Immersion Laboratory helped clarify how interactions change across time; however, these studies focused solely in terms of non-verbal behavior (Novick and Gris, 2013).

When evaluating familiarity, we hypothesized that by increasing the ECA’s extraversion as the relationship progresses across time, participants should experience a stronger sense of physical connection. In particular, we studied the effect of differences in the amplitude of nonverbal behaviors by an ECA interacting with a human across two conversational sessions. In the first session, the ECA used nonverbal behaviors with a lower-amplitude baseline. Our independent variable was whether, in the second session, the ECA used same baseline amplitude nonverbals, indicating no increase in familiarity, or used higher-amplitude nonverbals to convey an increase in familiarity.
Figure 4: First version of the first ECA built at UTEP’s Immersion lab.

The sessions involved a conversation with a life-sized, front-projected ECA in UTEP’s Immersion Laboratory, where the ECA served as the narrator for an adventure game developed specifically for this study. The game, “Escape from the Castle of the Vampire King,” was inspired by early text-based adventure games such as Zork (Anderson & Galley, 1985) and Colossal Cave (Crowther and Woods, 1976). Subjects can move from room to room, pick up, drop, and use objects, and kill vampires.

We chose this application because such games are known to be engaging, and we wanted our human subjects to want to interact with the ECA. A text-based game was also helpful from a practical standpoint, in that it limited the amount of speech that needed to be recognized; the subject’s possible utterances were both simple and highly constrained by the game’s context.
This ECA was the first fully functional agent developed by our team. Through playability, usability, and pilot studies we redesigned the agent to minimize external distractions and to improve the dialog navigation of the game. We clearly noticed the effect and importance of first impressions, which led us to change the ECA’s appearance.

Figure 5: Second version of the agent with improved playability and usability, and a recorded voice.

This study also suggested a low emotional connection between humans and ECAs. One of the main reasons was due to the lack of emotion in the speech synthesizer; varied emotion would have increased engagement for critical parts of the game. This also caused a mismatch between what was being said and the animation that was representing that behavior. This is where relational models, which can be built based on verbal behavior, play an important role.

2.2.3 **Relational Theories**

Relational theories or frameworks fill the gap between (a) base behaviors such as first impressions and familiarity and (b) the different dimensions of rapport, which I discuss in section 3. Although, some of these behaviors do not fit directly into rapport models, the ECAs nonverbal behaviors
or representations can be easily related. What this means is that additional relational models are more specific to the type of situation, task, relationship, and agent but can still contribute to rapport and therefore should influence ECA design.

At first glance these might seem to be simply aspects of social relationships or interactions, but when combined they contribute extensively to rapport-building behaviors. These account for intimacy, reciprocity, affinity, and continuity.

**Intimacy**

Intimacy is an interpersonal process. In this particular context, I do not refer to intimate relationships in a couple setting, nor do I relate their nonverbal behaviors to physical contact, but rather the degree in which conversation participants disclose information about themselves (Reis and Shaver, 1988).

Although some agents’ goals involve creating romantic relationships, these are usually mediator ECAs (with unusual and rather bizarre exceptions of Japanese dating videogames). Intimacy within the scope of this research is when one person (or agent) expresses personally revealing feelings to another. It continues when the listener responds supportively and empathically.

This is by no means a new idea. In the mid-1950’s several studies examined the para-social relationships built between humans and movie characters or radio characters (e.g., Horton and Whol, 1956). These studies explained how people connect to these theatrical characters as they identify themselves with them. A properly crafted script reveals the character’s life and persona to the audience but it is not enough to disclose or to attempt to support such disclosures. For an interaction to become truly intimate the discloser must feel understood, validated, and cared for (Reis and Shaver, 1988). In movies or radio shows, the characters can make the viewer or listener laugh, cry or yell, but the character is always unaware of those reactions and therefore is unable to adjust its behavior depending on the audience response. Fortunately, this is an issue that can be addressed with virtual agents.

The question then is how people can create intimacy in human-agent interactions. Since intimacy is based on mutual disclosure, Adriana, our agent, works its way around this limitation by having a persona. Personas have a positive effect with regards to the affective dimensions of rapport (Lester et al.,
Creating a backstory for a character can help it reach a certain degree of intimacy. In other words, for the characters to be able to disclose personal information, it must first have personal information to reveal. This is easily observed in non-embodied agents like Apple’s SIRI, Samsung’s Galaxy or Amazon’s Echo, all of which respond to questions about their names, origins, likes, and desires, albeit in a limited way.

Nonverbally, intimacy is traditionally represented by proxemics: the closer you stand to someone, the closer the relationship is assumed to be (Herrera, et al., 2011). There is much more to intimacy and proxemics. More elaborate models account for interpersonal distance, lean, body orientation, and the physical plane where the interaction takes place (Andersen, et al., 2006), and although it has been shown to vary culturally (Baldassare and Feller, 1965), it is still an effective way to demonstrate trust. Other nonverbal cues that help intimacy-building activities are agreement gestures that contribute by acknowledging the revelation. However gestures that match supportive and empathically responses play a larger role and, unfortunately, are usually context-sensitive. This is also true for similar nonverbal cues. A couple’s frowning at each other in public, where social pressure would tend to suppress negative affect, may indicate stronger negativity than the same degree of frowning in the privacy of the home, where intimacy is expected to generate a more open expression of frustration (Tickle-Degnen, 2006)

Intimacy is a complex behavior related to immediacy, affection, receptivity, trust, and depth (Burgoon and Newton, 1991). Intimacy is strongly related to mutual understanding and affinity, and it is a key component in rapport-building activities.

Affinity

Affinity is the process in which people try to make others have positive feelings towards them; it can also be described as a sense of connection (Bell and Daly, 1984). Using intimacy we can then build affinity. In fact, affinity and intimacy are so closely related that previous research often use, the terms interchangeably or uses one as part of the definition of the other (e.g., Burgoon and Newton, 1991).

A common technique to connect with people is to use shared knowledge and common ground to find shared ideologies, facts, and experiences. Affinity makes use of these agreements and associates
them favorably. In other words, it is easier to generate a sense of mutual understanding, or to put yourself in someone else’s shoes when your backgrounds and experiences are similar.

The concepts of affinity, shared knowledge, and common ground, lead to two related questions. First, can you build affinity with someone you dislike? And second, can you build affinity with someone you have nothing in common?

Although there are no definitive answers to either question due to their subjective nature and the blurred lines that surround the scope of many of these behavioral definitions, I will provide a broad explanation of both as applicable in this research.

You can build affinity with someone you dislike, but it is difficult, and, arguably, by the time you have built affinity you no longer dislike them. The key here is that in a controlled environment, it is unlikely for someone to dislike everything about a person or character. One might be biased to dislike everything about someone if that person has directly and greatly affected the person he or she is interacting with in a strong negative way, but for the sake of this work and realistic agent implementations, I will not go in depth for this particular point. Hollywood makes use of this to create attractive villains. They might do terrible things, but their charisma sometimes backed up by a preview of the events that have them into this path (which is an expression of intimacy, even if it is not provided directly by the character, for example, when the character is not telling the audience or another character their story) can make them into (somewhat) likeable characters. Although people rarely will side with villains and hope for their success, viewers might be displeased if some characters are too early defeated or removed.

Second, building affinity with someone that has nothing in common with you is also possible. Assuming that as human beings we all share similar behaviors, including physiological needs, among other things that will usually provide several things in common (e.g. weather talk is usually an excellent ice-breaker when first meeting people). This is why it is particularly important to create anthropomorphic agents --that is, human-like characters that will help people relate and connect with them.
Of course, affinity-building agents should try to avoid or minimize the impact of presenting themselves to people that initially dislike them and should attempt to use of common knowledge to make themselves understood and reacted to in a positive way. In fact, the two situations described are extreme scenarios that hinder affinity-building behaviors, and as such, rarely encountered in real-life scenarios with virtual agents.

It is important to note that the absence of nonverbal vocal and physical cues denies users important information about their partners’ characteristics, emotions, and attitudes, resulting in less sociable, relational, understandable, and/or effective communication, and therefore hinder affinity-building behaviors (Walther, et al., 2005). Culnan & Markus (1987) proposed that relational information is derived from nonverbal cues such as voice quality and vocal inflections, physical appearance, bodily movements, and facial expressions; however, they failed to provide a model of the specific nonverbal behaviors related to these cues. Although some attempts have been made to list and categorize nonverbal expressions of affinity-building behaviors, the field lacks a comprehensive list. Moreover, some behaviors overlap greatly with other dimensions of rapport.

**Reciprocity**

In this next section, I shift from rapport-building relational models to rapport maintenance models, which are particularly useful for long term relationships. Reciprocity can be defined as the preference of similarity. This is basically the golden rule: one should treat others as one would like others to treat oneself (Cole and Teboul, 2004). This makes sense from a relational standpoint, as we prefer people with behaviors that we can understand. If there is an absence of reciprocity, there might not be enough common ground and rapport to establish a long-term relationship.

Reciprocity is a multifaceted concept. It is commonly used in game theory and financial education, and it also relates to fairness, cooperation, kindness and competition (Falk and Fischbacher, 2006). From a more human perspective, reciprocity means that people evaluate the kindness of an action and react to it. Furthermore, this kindness is evaluated not only by its consequences but also by the intention underlying the action (Falk and Fischbacher, 2006).
Reciprocity therefore helps establish models for intentions. One clear example, of which I will go in detail at a later section, is the functional markup language (FML), which is used to specify communicative intentions (e.g., beliefs and emotions) that go along with what the agent wants to say (Mancini et al., 2008). While these models exist, they remain somewhat vague representations for behaviors in XML files and are far from being a standard. This is mostly because, like many relational behaviors, they have no direct translation to nonverbal behavior, or the nonverbal behavior has to run in parallel to the verbal behavior, a mode of interaction for agents that is complex to automate.

Even when there is some evidence of universal gestures or facial expressions for emotions such as happiness, sadness, anger, fear, surprise, disgust or interest (Eckman, 1989), although it could be argued that no particular class of behavioral events, such as smiling, is associated with an invariant social meaning (Scherer and Ekman, 1982) there are no gestures associated directly with reciprocity (e.g., there no significant evidence towards universal gestures for expressing thanks).

Continuity

When a relationship is based in intimacy, affinity, and reciprocity, it is already on its way to a successful, synergistic, and lasting relationship; maintaining that relationship over a long period of time is the last piece of the puzzle.

Continuity is not a behavior by itself but rather a progressive pattern of interactions. At its core, continuity means that each conversation should end with the possibility of continuing the interaction at a later time (Fisher and Dreksel, 1983). In terms of ECAs, this means that the interaction should be designed in such a way as to allow continuous discovery and disclosure throughout the interaction timeline. Failure to do so decreases engagement over time. This is commonly observed in many commercial agents, where asking the same questions lead always to the exact same answer, whereas continuity would have the response promote future interaction or provide additional information. At the very least, repeated questioning should trigger misunderstanding-repair behaviors.

Even when continuity is not an actual behavior, it is also an important element not only for rapport but for any interaction, as it will enable recovery should affinity, intimacy, or reciprocity
behaviors fail. It is also important to note that continuity is the only way to achieve familiarity, as it requires multiple interactions that relay are built over previous knowledge of past communications.

Each of these relational models were proposed independently; therefore, each model by itself only explains a constrained fragment of rapport. The highlight of these models is that these interaction traits rely heavily on context and verbal disclosure, which are unfortunately particularly difficult to implement in ECAs. Fortunately, with a carefully constrained interaction, these traits can be at least partially implemented, and some can even be translated into nonverbal behaviors that serve a similar purpose.

2.2.4 Rapport

Rapport has been described as a relationship quality. It serves as a powerful force in the accomplishment of tasks that are challenging and require mutual commitment to accomplish (Tickle-Degnen and Gavett, 2003). Individuals in rapport are more likely to cooperate to accomplish tasks that would otherwise cannot be performed as efficiently without collaboration (Tickle-Degnen and Rosenthal, 1989).

Rapport involves the coupling of internal experience and nonverbal action into an adaptive perception-action system (Fridlund, 1997). Once in rapport, individuals form a physical and social interdependence. People develop, maintain and indicate their rapport through a series of interconnected signals and responses that are shaped by the context of the task and the environment surrounding them (Fridlund, 1997).

Rapport exists only in interaction. It is not a personality trait, although an individual may be particularly adept at developing rapport in certain situations. Individuals experience rapport as the result of a combination of qualities that emerge from each individual during interaction. This experience is expressed clearly when people say they "clicked" with each other, or felt the good interaction to be due to "chemistry." The interaction itself during the experience of rapport becomes an entity not easily divisible into characteristics that each party brings to the interaction (Tickle-Degnen and Rosenthal, 1990).
In addition, rapport implies the existence of a mutual responsiveness, such that every member of the group reacts immediately, spontaneously, and sympathetically to the sentiments and attitudes of every other member (Park and Burgess, 1924).

While several theoretical frameworks explain at least one or more key aspects in rapport, Tickle-Degnen and Rosenthal’s (1987) model is by far the most popular among social scientists. In this model, rapport comprises three dimensions: attentiveness, positivity, and coordination.

**Attentiveness**

Attentiveness at its most basic level ensures that the conversant’s focus is directed toward the other. It creates a focused and cohesive interaction. According to this model, as a consequence, participants experience a sense of mutual interest in what the other is saying or doing (Tickle-Degnen and Rosenthal, 1987). In other words, there cannot be rapport if the people engaged in conversation are not paying attention to one another. Not paying attention might allow people to understand the words in the interaction but not the meaning. This can not only fail to generate rapport but also destroy any sense of connection that was previously attained.

**Positivity**

Positivity is, as the name implies, a positive affect. It causes conversants to feel mutual friendliness and caring. It creates respect and beneficence towards another person. Positivity is closely related to the degree of involvement and attentiveness. Furthermore, positivity can be bi-directional, reaching negative values when the interaction turns into a verbal combat (Tickle-Degnen and Rosenthal, 1987).

**Coordination**

The third and final essential component of rapport is coordination between the participants. Balance, harmony, and being "in sync” are positive connotations of coordination. In addition, in an interpersonal context, coordination conveys an image of equilibrium, regularity, and predictability between the interactants. Coordination is usually expressed through gesture mimicry or posture
mirroring and verbal synchronization, making it one of the most difficult dimensions of rapport to design, measure and evaluate (Tickle-Degnen and Rosenthal, 1987).

2.2.5 Rapport across Time

Tickle-Degnen and Rosenthal’s model assumes that positivity becomes less necessary over time, while coordination increases in frequency and importance. This relationship is depicted in Figure 6. The rationale behind this relationship can be understood by thinking of how our own day-to-day interactions correlate with the earlier discussion of familiarity and first impressions.

![Figure 6: Importance of the three components of rapport across time (Tickle-Degnen, Rosenthal, 1987)](image)

When we first meet someone, our nonverbal behavior is careful and controlled. It is important to be viewed favorably by the other person and so we attempt to form a positive image for ourselves. Coordination at this stage plays a secondary role, as it is difficult to mirror what has not been seen or experienced (Tickle-Degnen and Rosenthal, 1987).

As the interaction progresses and the both conversants form an image of their conversational partner, positivity’s effect decreases and coordination naturally evolves as signaling of mutual understanding with mirroring as the nonverbal representation of being in sync. It is important to note
however that there is no evidence of a specific timeframe that delimits the change (Tickle-Degnen and Rosenthal, 1987).

Most studies performed in controlled environments with virtual agents account for only a single session, that is, a one-time interaction is assumed to be representative of multiple encounters. A few exceptions to this experimental design have shown that there are effects for long-term rapport that can only be measured across several sessions and much longer timespans, some of which extend to several months (Bickmore, 2010). The interconnectedness of the relational models to familiarity can be observed as an iterative process (Figure 7) in which the relationship evolves by refining the affinity process and increasing the intimacy disclosure, while maintaining the key characteristics and structure of each element of the relational model.

![Familiarity diagram](image)

**Figure 7**: Familiarity timeline expressed as a relational model.

### 2.2.5 Nonverbal Rapport

My research attempts to test and create a model that reflects not only the affective nature of rapport (e.g. how it feels) but also behavioral expression, with an emphasis on the elicitation of rapport.

Due to the nature of the experience of rapport, it is not easy to identify behavioral correlates of rapport. For example, furrowing the brow during an easy task can indicate confusion and be associated with feelings such as incompetence, which are not conducive to rapport, while the same gesture during a complex task may be indicative of concentration and challenge, which may indicate an effort consistent with rapport (Tamir et al., 2004).
Previous research has suggested that behaviors should be analyzed at a molecular level, that is, counts or durations of specific behaviors such as head nodding or eye contact (Harrigan and Rosenthal, 1986). However, this may not be the most appropriate level for investigating rapport. The discrete nonverbal acts in a molecular analysis are typically measured for each participant separately, neglecting the interactional nature of rapport. Variables that combine the discrete acts of both participants may be more appropriate for relating nonverbal behavior to rapport.

Tickle-Degnen and Rosenthal (1989) suggested evaluating the effect of eight classes of nonverbal behavior: directed gaze, smile, head nodding, forward lean, body orientation, posture mirroring, uncrossed arms and uncrossed legs. Some of these behaviors can be associated directly with a particular dimension of rapport (e.g., smiling and head nodding indicate acceptance, liking or approval, which correlates to positivity; direct gaze, lean and body orientation correlate with attention), where others such as uncrossed arms or legs are dependent on the pose and much harder to correlate with a particular dimension and depend on additional context (i.e., participants’ stance). Furthermore, I propose that some of these behaviors can be classified as eliciting, while others are merely representative of rapport, depending on who holds the floor at that point of the interaction.

2.2.7 Mimicry

Mimicry is one of the nonverbal representations of the coordination dimension of rapport, a process that has been studied for over 100 years (e.g., Lipp, 1907). Although there is no definitive guide to proper imitation, partially because it is a subconscious process that varies depending on the imitation target, it is one of the most active nonverbal components of rapport, as it usually involves imitation of stance, gestures, facial expressions, and verbal expressions (Chartrand, 2008), and is partially responsible for the similarities that form consistent stereotypes.

Most important, mimicry serves an important role in communication. Imitation has been considered as a facilitator of emotional decoding. Subjects who engaged in a previous interaction where facial mimicry occurred detected the emotions of their original facial reactions with an accuracy of 10 to 40% (Wallbot, 1995). This suggests that mimicry is a dynamic, nonverbal and temporary common ground that helps to create mutual understanding during an interaction. If the subjects of the interaction
are not consistent, that is, there is no continuity in their relationship, this common ground is forgotten as it no longer serves a purpose, since it cannot be interchangeably applied to other conversations with different participants.

Motor mimicry is just one aspect of imitation. Congruence, contagion, empathy and interactional synchrony combined are a solid attempt at explaining this behavior. These additional concepts ensure that complete mimicry is avoided, or that there is a meaning and relational purpose behind imitation, while other concepts like contagion explain imitation in multiparty settings. For the purpose of this study, mimicry is used solely as a behavioral representation of rapport, and the specifics of these models are outside the scope of this dissertation.

Implementing mimicry behaviors is not easy. Although there has been recent progress in the classification and analysis of multimodal mimicry (e.g., Sun, 2011), additional problems for real-time implementation of these behaviors remain. For any type of nonverbal similarity to converge, be congruent, and establish rapport, both partners must perceive the behaviors correctly and quickly. Mimicry itself is an action that can contain an endless supply of context-sensitive nonverbal behaviors. These complexities make it difficult to create models with integrated gesture libraries that can be procedurally accessed in real-time during an interaction. If this complex behavior cannot be modeled with high accuracy, what is left is a matching set of congruent reactions based on observation, translated into general gestures. This is arguably why people mimic agents in interactions, but agents are not capable of doing the same from their human counterparts.

Finally, if mimicry exists to the extreme, one might hypothesize that, given two people engaged in conversation, and the emotional expression of the first person would be fully imitated by its conversational partner, which then, would imitate it back creating an endless cycle of identical expressions. This is obviously not the case, but it is difficult to assess the optimal degree of imitation, which relates to the difficulty of implementing these behaviors. In the case of agents, if they completely mimic the behavior of the user, at some point the user will feel annoyed, or erroneously think he or she is in control of the virtual body rather than considering it as a separate entity.
2.3 ECAs Designed for Rapport

I now shift the focus of the discussion from the human-human models to the set of agents that have been successfully designed to engage in different types of rapport-building activities.

Rapport-building agents are based on deception. Agents, as we know them, are programmable characters, designed to behave in a scripted or perhaps procedural way. They are nonetheless designed. These programmed behaviors enable them to exhibit cues such as attentiveness or positivity when they don’t really “feel” it. These reactions are in reality the expressions of the developer’s desire to make humans interacting with the agent believe that the agent has these feelings.

Creating this illusion of rapport requires making the end users forget that they are interacting with a machine. We, as developers, make extensive use anthropomorphism and storytelling as tools to create cinematic-like yet interactive experiences for the users. Interactivity alone, however, is not enough to achieve realism in a first-person interaction; in addition, this requires social presence and awareness to make users believe that the agents are part of the real world.

2.3.1 Social Presence and Awareness

Social presence is the sense of “being with another.” People are typically interested in social presence because it may mediate the effects of other variables of central concern, such as attitudes towards the mediated others, features of the interface, persuasion, illusions of reality, learning and memory (Bailenson, 2001).

Awareness is how I refer to a virtual agent’s social presence- that is, how aware the agent is of its virtual and physical surroundings and the user. Unlike the physical environment, social communication in virtual environments might be built upon minimal or constrained social cues. Animated characters and even the computer interface itself can generate strong automatic social responses from minimal social cues. Social responses to computer characters, for example, are generated even though the user is quite aware that the computer is not an emotional or social agent but a machine (Reeves and Nass, 1996). In other words, the more human-like the user thinks is being perceived by the agent, the more human-like the human behaves. Traditional human-computer interaction principles suggest that users will try to accomplish tasks with a minimum effort. Rapport is not about minimum effort but rather
about understanding, bonding, collaborating and interacting with virtual characters, preferably in tasks that cannot be completed as effectively or efficiently without it. Awareness is then how I define the process by which agents’ sense humans and communicate their feeling of physical or conventional presence, social presence and virtual presence to the users (Figure 8).

![Awareness Diagram]

**Figure 8**: Conventional presence, social presence, and virtual presence as components of awareness

### 2.3.2 Anthropomorphism

ECAs can improve multimodal HCI because of their appearance and because they show human-like communicative behavior, users tend to ascribe human characteristics to ECAs, that is, to anthropomorphize them. Since interacting with another human being comes naturally to people, anthropomorphisation of agents in an interface can improve the process of communication (Beun et al., 2003).

Although rapport may not be exclusively a human behavior, and several animals exhibit social traits beyond those of instinct (i.e. guide dogs), a basic premise of this research is that rapport is exclusively a human state of interaction. As a human trait, anthropomorphic agents are preferred to establish, analyze and evaluate rapport. As a consequence, ECAs create a stronger sense of connection, understanding, and shared knowledge by exhibiting traits particular to human beings, which are more easily understood by the humans interacting with them.

### 2.3.3 Storytelling

Extensive literature explains or improves upon storytelling techniques (Delgado, 1989; Ochs and Capps, 2009). These techniques differ across their intended audiences and mediums. Storytelling
performs a different role in videogames than in movies. TV-shows rely on continuity and parallel, long-term subplots that may extend over several years.

Similarly, ECAs require a variation of traditional storytelling techniques to perform their functions effectively. Storytelling in ECAs will vary depending on the user’s task and the goals of the agent, if any. Every agent should be able to share at least partial details about its persona and give enough context about the virtual world for it to be understandable for users.

2.4 Virtual Rapport

Virtual rapport is the state of the interaction that is built in collaboration by virtual agents and humans, analogous to human-human rapport, IT creates a sense of mutual understanding and synchrony.

A similar model to Tickle-Degnen and Rosenthal’s three-dimensional rapport model describes virtual rapport for human-ECA interaction (Huang, et al., 2011), at a somewhat higher level of abstraction. According to this model, rapport can be divided across three dimensions:

1. Emotional Rapport: The sense of connection with the user.
3. Behavioral Rapport: Verbal properties, such as speech duration, pitch, etc.

This model, as with the other models of rapport, lacks the specifics for the nonverbal behavior; that is, it does not provide a direct translation to actions.

Zhao et al. (2014) made significant progress toward a computational model of rapport management. Their model (see Figure 7) offered a good starting point towards rapport-enabling behaviors by dividing the strategies and translating them into actions. But this is a cumbersome and exceedingly difficult task. While it is possible to define a strategy repertoire based on the dyadic state of the conversation, which in this case combines the social and task goals, the main problem is mapping the dimensions and the dynamic context to subsets of nonverbal action.
The combination of the definitions or models for familiarity, intimacy, affinity, reciprocity, continuity, and rapport, in conjunction with strategies for ECA and interaction design, which make use of anthropomorphism, storytelling, social presence and awareness, and mimicry forms the base for a relational rapport model. While the model does not fill in all the nonverbal behaviors missing from the Tickle-Degnen and Rosenthal definition of rapport at a “molecular” level, and provides only general guidance towards a molar approach, it connects and categorizes the relational factors of rapport and combines them with human-ECA rapport. This makes it possible to translate some verbal rapport-building behaviors into paralinguistic behaviors.
There are two key differences between this approach and previous attempts at building rapport between humans and ECAs. First, this is not a direct attempt to improve the users’ perception of rapport in virtual agents; instead, it manipulates the interaction through the virtual environment to give contextual references in a relatively long interaction (40 minutes to one hour) to increase engagement and users’ nonverbal rapport-building behaviors towards an agent. In other words, this research intends to have users react to the agent by exhibiting rapport-building behaviors towards the agent, as a consequence of the ECA’s apparent awareness. Second, while most research is interested in facial expression, this study, in an attempt to increase the awareness level, makes use of a full-body, gesture sensitive, standing representation of an agent. In addition, the agent and the user can together navigate and explore the 3D virtual space in a semi-constrained way.
This research aims to improve naturalness and effectiveness in ECA interfaces by increasing rapport between agents and humans. Recall that the more human the human thinks it is being perceived, the more human the human behaves (Zhang and Hill, 2000). For example, if the human realizes that not looking at a talking ECA in a computer screen is perceived as an inattentive or even rude behavior and the ECA reacts to it, then it is likely that the human will pay attention to the screen.

This assumption may lead to the belief that creating agents that are more aware of their human counterparts, that is, agents that perceive and react to every movement, will solve any problem; unfortunately, this is not always the case. It is unlikely that with our current technology we can implement a library with every possible human pose or gesture, and even more so to assign all possible human interpretations for each gesture.

It follows then that it is currently impossible to create agents with proper reactions to every possible interpretation of every possible gesture. This would essentially be a brute-force approach, which would be extensive and ineffective.

A smarter solution would use a limited set of rapport-building behaviors. The main problem with these nonverbal behaviors is that users seldom act with full-body gestures in front of ECAs, perhaps because. One possible reason is that users don’t believe that their body language matters - and for the most part they are correct. Out of the more than one thousand virtual agents listed by chatbots.org, only a handful are even slightly aware of human paralinguistic behavior (not including haptics in mobile devices).

Given that rapport is a two-way interaction intended to achieve goals not directly related to the conversants' nominal task, and that mutual understanding can occur even when the conversants disagree (Novick, 2000), it is not enough to have the ECA generate rapport-building behaviors by itself. Rather, the ECA should be eliciting similar behaviors in the human conversant. In other words, ECAs should be capable of eliciting human rapport signals through mutual expression of rapport-building actions.

2.4.1 Rapport Building Agents

Intentionally or not, many ECAs are capable to some extent of building rapport. An ECA capable of building rapport more fully would have two key characteristics:
1. Non-verbal behaviors to represent each of the three dimensions of rapport in the composite model (emotional connection, mutual understanding, and physical connection).

2. Adaptive behaviors across time, for example, should a model dimension be preferred over another after a history of events, or if a dimension decreases in importance after a longer period of interaction?

These abilities that humans can innately execute permit us to display a physical representation of rapport. Creating agents that are expressive, adaptive, and capable of producing these behaviors is the first step towards rapport building agents. Of course being able to move, adapt and talk are merely the agent’s features, after which the interaction design has to allow the agent to follow a model to build rapport. With such an ECA, users should not only be able to notice the difference between a rapport-enabled ECA and a non-rapport ECA, but they should also prefer interacting with the rapport-enabled ECA.

Likewise, an ECA capable of making humans express rapport, would have two agent behaviors that induce the user to attempt to generate rapport:

1. Common ground and available shared knowledge.

2. Non-verbal behavioral cues to indicate misunderstanding or request information.

In both cases, appropriate timing for displaying non-verbal rapport behaviors is required.

Creating an ECA with all four properties is challenging. In particular, it is difficult to create common ground and available shared knowledge while in parallel tracking and reacting to the user’s full-body gestures. If however, the common ground and shared knowledge between human and ECA could be translated into nonverbal behaviors, then the agents can manage more accurate responses while being less technically complex. For example, imagine a couple on their first date. They are sitting across the table watching each other, being careful (and probably clumsy) with their verbal and nonverbal behaviors. This is confusing and hard to process, and even for human beings this can easily turn into an awkward situation. However, actions speak louder than words, and eventually the boyfriend-to-be stands up, grabs her hand, and takes her to the dance floor. Then they dance. While they are dancing, most of the verbal behavior is gone and the subtle nonverbal behaviors are absent because they are concentrating...
on the task instead. In other words, dancing is the common ground and shared knowledge, appearing as a physical bound that increases rapport, both immediately and over time, even while performing the same task repeatedly. To increase rapport, I attempt to find, model, and replicate this in a human-ECA dance.

2.5 Engagement

Rapport is an ongoing process, and as such it has to be maintained by both conversants. The lack of rapport between human and ECA in the early stages of the interaction originates from the lack of familiarity with a partner’s behavioral cues for indicating misunderstanding or requesting information (Cassell et al., 2007). Rapport is something that typically develops over time as inhibitions break down and partners begin to form emotional bonds. Strangers rarely exhibit the characteristic positivity, mutual attention or nonverbal coordination seen amongst friends (Welji and Duncan, 2004). A user who scores an ECA low on rapport during initial interactions does not necessarily mean that the rapport behaviors are inappropriate but that the relationship is not mature enough to establish a strong rapport connection. Instead, the relationship is at an initial stage where the most important non-verbal behaviors are aimed at facilitating communication by addressing misunderstandings and getting to know the other conversation participant.

To maintain rapport, it is necessary to engage the user in the interaction; that is, retain user’s attention through long periods of time. In our own studies, we have explored engagement by measuring attention through gaze, while others have tried to index it by measuring speech fluency, that is, the total time the subject engaged in conversation and the total number of words the subjects used (Gratch, et al., 2007). We have also been able to increase engagement by improving voice quality and including contextual virtual backgrounds, which are further explained in Section 4.1. Other approaches examine engagement in virtual humans that interact with users daily, and integrate dynamic social storytelling to maintain user engagement and improve retention (Bickmore, et al., 2010). Their experiment tested user attitudes towards the agent when it presented the stories in first person (as its own history) compared to third person (as happening to humans that it knew). Interestingly, participants in the first person
condition reported enjoying their interactions with the agent significantly more and completed more conversations with the agent, compared to participants in the third person condition.

Our agent was then developed as a first person, automated, dynamic storytelling interaction that takes advantage of virtual backgrounds and recorded voice to create the most engaging experience possible. Unfortunately, due to the subjective nature of self-reported engagement, it was only measured by a multiple-choice questionnaire presented in Table 1.

2.6 Physical Engagement to Increase Rapport

I have presented rapport as a complex and extensive behavioral state that is difficult to model, measure and interpret. One of the reasons is the dynamic nature of interactions, where rapport is not stable, but rather built, maintained, or destroyed.

This becomes evident when looking at previous instruments of rapport. Puccinelli and Tickle-Degnen established a rapport instrument in 2004 that includes the following questions: To what extent do you think the individuals liked each other (affinity); To what extent do you think the individuals were aware of each other (awareness); To what extent do you think the individuals felt coordinated with each other (coordination); To what extent do you think the individuals felt the same (affinity); To what extent do you think the individuals understood each other (mutual understanding); To what extent do you think the individuals had a feeling of mutual agreement (affinity). This instrument covers some key aspects of rapport, however, it applies to an observer and not a participant of the interaction, and therefore does not include any type of engagement measure.

Other instruments, such as the ones developed by (Gratch et al., 2006) apply to human-agent interactions, however, they are dependent upon the interaction style. This survey included the following questions: Do you think he/she understood the story completely (understanding); Did you feel that you had a connection with another person (emotional); While you were telling the story, was seeing the avatar helpful or disturbing; Did you feel that you used some feedback from the other person when telling the story (engagement); Do you think the avatar portrayed the movements of another person accurately (physical). The rapport agent in that study was a listening agent that reacted (or not) to users telling it a story. This survey was expanded a year later to include more rapport-specific questions: I
think the listener and I established a rapport; I felt I was able to engage the listener with my story (engagement); I felt I had a connection with the listener (affinity); I think that the listener and I understood each other (understanding). Some items in this instrument measure a fraction of rapport, while others measure engagement with the agent, or the user’s perception of the agent. One of the problems with this measure is that it was designed for a listening agent rather than an interactive one.

Another complication is the subtlety of rapport; many studies indicate that the benefits of rapport fall outside of conscious awareness and that people often show measurable impacts on their observable behavior without reporting significant differences when introspecting upon their experience (Gratch, et al., 2007).

This suggests that rapport should be also measured through non-verbal user actions. The main problem with using nonverbal behaviors as a measure is that most users will not engage physically with the agent. Even if they do, it is difficult for agents to measure, analyze and react to it immediately to maintain believability, engagement, and rapport. To address this, I propose a study in which I analyze if physical engagement is a good way to improve the emotional dimension of rapport.
Chapter 3: Experimental Design

3.1 Experiment

Given synergy between paralinguistic and verbal rapport-building behaviors, the relational rapport model can be mapped to nonverbal behaviors only if there is a contextual awareness. That is, in order to target specific nonverbals of rapport, some verbal context is required to let the user know that he or she is being understood by the agent.

Doing so however, limits two factors of rapport that usually are built upon verbal behavior, which are emotional connection, and mutual understanding. For example, a simple posture, even when it can potentially carry some conscious meaning or be an attitudinal behavior, can be easily misinterpreted without any additional information (e.g., a crossed-arms posture can indicate a certain closure to external social influences but it can also reflect that a person is feeling cold).

I claim that inducing participants to engage in physical interaction through domain activities can increase rapport behaviors in the emotional connection dimension.

To test this claim, I designed an experiment with two conditions. In the first condition, participants interact with an ECA that elicits paralinguistic rapport-building behaviors from the user where these actions are required to complete the task. In the second condition, participants interacted with the same ECA, and although the dialog and agent’s gestures remain unchanged, it did not encourage users to perform physical tasks.

My first hypothesis is that users who interact with the physical agent will report higher emotional rapport scores than those who interact with the non-physical agent.

The second hypothesis is that users who interact physically with the agent will show greater engagement (produce more non-task gestures) than those who interact with the non-physical agent.

Over 80 participants were invited to interact with our ECA, 40 in each experimental condition. Of those we had a total of 64 usable subjects, 32 for each condition. Each subject interacted with the agent for a single session of 40 to 60 minutes. The length of the session is set to account for the initial phase of rapport, where neither agent nor human have had time to establish a relationship, and to give time for building common ground, so that it is possible to reference past events.
Figure 11: Experimental setup. The wall is a frontal projection, meaning that getting closer to it would create a shadow. There is a camera hidden on the tree to the left, but users are informed and consent to being recorded. The Kinect is located straight in front of the user at the bottom center.

Each session was recorded with a regular camcorder aimed at the user (see Figure 11). Unfortunately, our initial attempts at recording the interaction through a Kinect™ camera with the pose detector enabled, and another Kinect™ with facial tracking were unsuccessful (Figure 12), as they lowered the performance of the other modules (speech recognition, full-body gesture recognition, game engine virtual environment rendering, etc.) causing a lag that aimed the interaction. The Kinect set to gather facial expressions in real time was too far away, it had to be placed at an angle to avoid obstructing the projected screen or being in the range of the projector light, and the data were too noisy to be usable.
All participants played the same overall scenario described in the next section under one of the two different conditions designed to measure the context-based verbal behaviors, paralinguistic behaviors, and their combined effect on rapport-building.

### 3.2 Physical Connection Enhanced Agent

Previous studies in the Immersion Lab follow a bottom-up approach, where the observations made on natural (human-human) and virtual (human-ECA) dyadic conversations were used to develop the paralinguistic representations of new models. Observations focused on when do paralinguistic signals such as, turn taking, control acts or mutuality confirmation signals appear, and how they differ across time. Gaze, nods, and upper body gestures were annotated for both the natural interactions and the resulting virtual interactions.

For example, our first pilot study (Novick and Gris, 2014) examined familiar and non-familiar ECA behavior, and how this affects user interaction. We examined the effects of users’ perception of the agent’s familiarity levels based on the agent’s level of extroversion. Then we analyzed the effects of the user’s experience and the user’s behavioral changes with respect to the agent’s current state, and explored automated annotation methods for both, agent and user verbal and non-verbal behaviors. The interactions occurred while playing a verbal version of a text-based game. In this experiment, the
participants were exposed to an ECA for two half hour sessions at least one day apart. During the first session, the agent exhibited non-familiar behavior in an attempt to mimic the paralinguistic actions of people when they make acquaintance for the first time. In the second session, for subjects in the experimental condition, the agent changed to a more extroverted, fast-paced behavior in an attempt to enact familiarity solely in terms of extraversion; subjects in the control condition continued with the initial non-familiar behavior throughout the session.

The latest version of the agent, Adriana, leads the user through a series of activities and conversations while playing a game. We simulate a survival scenario where the user has to collaborate, cooperate, and build a relationship with the ECA to survive. This simulation is built with the intention to maximize rapport-building opportunities, as well as to take advantage of the non-verbal behaviors in a more immersive environment, where both the user and the agent can interact with the same objects in virtual space. The storyline provides the necessary flexibility and decision making, without creating a completely open environment where tasks would otherwise be difficult to set up and evaluate.

The scenario comprises 23 scenes, and an introduction, where each lasts approximately five to eight minutes depending on the player’s choices and interaction speed. The participants covered all scenes with Adriana with an average of 40 minutes of interaction. Although there was an implied one-hour limit to finish the interaction (the sign-in sheets had one hour slots), participants were told to take as much time as necessary. These times ranged between 30 minutes and 55 minutes.

Each chapter was carefully scripted to enhance an aspect of a rapport-building interaction. The script provides a guideline for the relationship to evolve. The rapport-enhancing script is explained more in depth throughout the agent features section and is provided in the appendix for reference.

The first scenes focused on eliciting and providing some basic personal information, and at learning the environment and interaction style. There is no text, heads-up display, buttons or notifications, and all interaction occurs verbally and non-verbally via full-body poses and gestures.

Through prototype refinement, and because of the complexity of implementing coordination gestures on behalf of the agent, non-task gestures are logged and plotted, but not reacted to. The next scenes focus on enabling cooperation and collaboration, first by having Adriana suggest survival ideas.
and show the player by example how he or she can help. In other words, the agent will teach the player survival techniques and explicitly mention and demonstrate different ways the player can help. The scenes that follow will focus on providing several opportunities for the player to help the agent. These opportunities range from simple tasks, to life-threatening situations. These scenes increase the magnitude of the dependability gradually until the user and the agent are potentially able to leave the island.

Throughout every scene, each player is offered the same decision-making opportunities and asked the same personal questions. However, scenario choices, non-verbal behaviors expressed, and the amount of personal information revealed depend on each player and can affect the behavior of the agent.

Timing and synchronicity between verbal and non-verbal cues and verbal disclosure play a major role in rapport-building techniques (Huang et al., 2011). To address these, the agent must be capable of performing a series of tasks. Thus we created a Jungle Survival game scenario, which includes a set of activities that through collaboration and cooperation engage the users in rapport-building activities.

3.2.1 Verbal Behaviors

In the Jungle application, verbal interactions focus on context and not on vocal properties. Context is provided in three different forms:

1. Required: The scenario script leads the agent to indicate verbally to the user how to interact and to give verbal confirmation of the choices made by the user to progress through the scenario. The required dialog is scripted and, although choice is present, it does not vary across people who make the same path decisions.

2. Scenario context: The scenario context drives the storytelling throughout the interaction. The human and ECA are placed in an immersive space. Throughout the sessions, participants are exposed at a number of situations explained verbally by the agent and visually by the environment. The agent is designed not to explain verbally things that are obvious or directly observable. In contrast with detailed task-oriented designs, the scenario context is there only to provide references about previous interactions with the agent. This can be accomplished in many different ways depending on the status of the human-agent relationship and the current scene. For example, during the first scenes the agent explains to the
user how to fish. Suppose the user was a slow learner. If so, the agent can mention an appropriate comment in a future scene, such as “come on, this can’t be any more difficult than spear fishing!” while facing a different task.

3. Personal context: The interaction script offers the illusion of free speech, while directing the human through a series of small-talk and social tasks in addition to the goal-oriented tasks. Some of these tasks include the eliciting and procuring of personal information. For this scenario, the information provided by the agent is based on a fictitious bibliographical background of a young woman stranded on an island. Adriana is capable of soliciting personal information and establishing small-talk in a carefully controlled environment, by asking the user several seemingly open ended questions based on adaptive questionnaires.

3.2.2 Paralinguistic Behaviors

Non-verbal gestures and interactions are all behaviors that require, display, or invite movement.

1. Required: The scenario in which the agent is situated requires the agent to move across the virtual landscape to advance through the storyline. From the agent’s perspective, it includes animations for gaze, limited only at essential objects located on the virtual side and the user, and the required natural movement through the virtual environment (such as jumping, walking, crawling, breathing, weight shifting, etc.)

2. Rapport-building: These behaviors are not necessary to advance through the scenes, but depending on the movement types, their frequency and amplitude across sessions can indicate different degrees of rapport. In other words, these behaviors carry additional meaning and can not only express but invite the person to interact or to respond directly. Examples include full-body movements that partially mimic the user’s movements, distance, orientation, arm position, leg position, mutual gaze, gaze at background objects, gaze at active objects, head nods, head shakes, head tilts, and smiles. It is important to note that these are not achieved through a Wizard-of-Oz type approach. To automate these behaviors, the system selects from the animation library reactions to behaviors and expects the users to perform based on the script. Unfortunately, this system is not capable of handling mimicry and gaze is hardcoded to motivate the agent to look either at the user or at objects with which she will interact.
3.2.3 Gesture and Non-Gesture Conditions

To measure the effects of gestures in rapport we built two versions of our agent, one which requires gesturing to perform some tasks, and another where the agent performs the tasks and requests feedback usually with a brief confirmatory dialog.

During the survival scenario, there are several key scenes in which the user is required to take physical action, or not, depending on the condition. If the user is under the gesture condition, the agent will offer her hand to help the user stand up. This is of course a metaphorical gesture as there is nothing physical to hold, however the stretched arm towards the agent is recognized, and the camera is shifted to indicate that the user “stood up” after the gesture. It is important to note that there is no user representation or avatar in the virtual world, and that the game is entirely in first-person view. Other key gestures in this condition are lighting a fire, building an SOS signal, catching objects, fishing, helping the agent stand up, and signaling a helicopter.

Under the non-gesture condition, users go through an almost identical dialog, but the agent takes the initiative and performs the physical action. For example, to light a fire under the gesture condition, the agent will ask if you know how to do it, then, regardless of the answer she will encourage the user and suggest that he or she is more than capable and should help. Under the non-gesture condition, the agent will ask the same initial question, and regardless of the answer, she will volunteer to light the fire.

In terms of dialog, only one or two words are changed per sentence during gesture elicitation, enough to make logical sense without altering the course of the conversation. For example, during the fishing sequence, the agent will say “you got one” when you successfully fish something, and “I got one” when she is performing the action. Users were not instructed as to what the appropriate gestures were or what the correct way to perform them, this to avoid any gesture bias and to maintain the immersion aspect of the game.

The full game script is provided in Appendix A and a more elaborate description of the game is provided in chapter 4.
3.3 Measuring Rapport

Our analysis compared the users’ non-task gesture input and user responses to our rapport instrument. As previously mentioned, self-reporting rapport is subjective and difficult to measure when introspecting one’s experience. Unfortunately there is not a standardized test to measure rapport. Previous studies that have attempted to measure rapport do so by measuring engagement, speech fluency (Gratch, et al., 2007), or social presence (Huang, et al., 2011). We included relevant questions from these surveys, and adapted and extended the survey of (Acosta and Ward, 2011) into an instrument of twelve Likert-scale questions, balanced for positive and negative responses that covered the three rapport factors in our model. Table 1 lists the questions in the rapport instrument. For all further analysis the questions were normalized so that a positive score always represents a positive effect.

Table 1: Rapport Instrument. Subjects indicated agreement or disagreement on a five-point scale

<table>
<thead>
<tr>
<th>Item</th>
<th>Rapport Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>The agent understood me</td>
<td>Understanding</td>
</tr>
<tr>
<td>The agent seemed unengaged</td>
<td>Understanding</td>
</tr>
<tr>
<td>The agent was excited</td>
<td>Physical</td>
</tr>
<tr>
<td>The agent's movements were not natural</td>
<td>Physical</td>
</tr>
<tr>
<td>The agent was friendly</td>
<td>Emotional</td>
</tr>
<tr>
<td>The agent was not paying attention to me</td>
<td>Understanding</td>
</tr>
<tr>
<td>The agent and I worked towards a common goal</td>
<td>Emotional</td>
</tr>
<tr>
<td>The agent and I did not seem to connect</td>
<td>Emotional</td>
</tr>
<tr>
<td>I sensed a physical connection with the agent</td>
<td>Physical</td>
</tr>
<tr>
<td>The agent’s gestures were not lively</td>
<td>Physical</td>
</tr>
<tr>
<td>I feel the agent trusts me</td>
<td>Emotional</td>
</tr>
<tr>
<td>I didn't understand the agent</td>
<td>Understanding</td>
</tr>
</tbody>
</table>
Chapter 4: The Agent

4.1 Agent Development and Tools

An ideal ECA would have a large pool of verbal and nonverbal behaviors from which to build human-ECA rapport. Such an agent would be capable of producing and recognizing full-body movements, proxemics, upper body orientation, lower body orientation, arm position, leg position, mutual gaze, gaze aversion, facial gestures, and head movements (e.g., nods, shakes, and tilts). In addition, it should be able to remember important conversational events, establish common ground, and attain a state of shared mutual knowledge with the human participant. Verbally, it should recognize natural language and communicate emotion making use of the prosodic and paralinguistic attributes of speech, while providing flexibility for turn-taking (e.g., interrupts, overlaps, and silences).

When designing our agent, one of the most important features we had in mind was full automation. Rapport is established and maintained over time. To do this, initially we considered long-term human-agent interactions that would have involved several sessions across several weeks, which would have been cumbersome and inefficient if done in a traditional Wizard of Oz (WoZ) style. There are several benefits for WoZ in terms of naturalness and believability of an interaction, for example, while a WoZ agent is always contextually aware, a fully automated system can be easily thrown off track by any unexpected user input, or by an expected user reaction but at an unexpected moment.

Our agent is fully automated. Although ideal agents are still technologically unattainable, in the following sections I explain our research group’s ECA’s architecture, along with its features, strengths and shortcomings.

4.1.1 Agent Architecture

Our research group developed agents independently of SAIBA, BML, and FML (Gris, et al., 2014), in part because we wanted to understand agent development from the ground up and in part because our research requires some less-common features, such as recognition of full-body gestures, blending animations, and agent portability and reuse. The implementation uses Unity 4, a Microsoft Kinect, and the Windows Speech SDK, interfaced and networked with each other and synchronized to
handle the agent’s complex behavior. We use the Unity 4 game engine to display our agents and Unity’s Mecanim system to create an extensive array of animations.

![Diagram](image)

**Figure 13.** High-level system architecture of the ECA system.

Animations are played by a state graph that follows user-specified parameters of when an animation should start, end, or blend with another animation. Multiple animations can be blended to obtain a completely different animation in real time and to give the user the impression that the agent never moves in exactly the same way twice. Animations are divided layers that can control different parts of the body, so multiple animations can be played at the same time and affect different limbs of the agent - for example simultaneously playing a blinking and talking animation on the face only, an explaining animation on the arms, and a walking animation from the hips down. The animations are played when the system decodes a message sent by the dialog tree that has the information about the specific animation to be played, the length of the dialog that the agent will say, and the position where the agent and the player should be. To describe interactive scenes, we developed a dialog interpreter that parses an XML document and links dialog states through conditionals. The XML file includes the responses anticipated by the systems at relevant parts of the scene. After the interpreter compiles the file, it builds a dialog tree that contains the relationships of the dialogs segments through the storyline.

The ECAs are built using a three-tiered architecture. At the bottom layer is the Kinect™ sensor, which supports RGB video, audio provided by an array of microphones, and a depth field based on infrared sensor information. The middle layer contains all the logic of the agent’s behavior and sensor
interpretation, including speech recognition, text-to-speech or audio playback, and gesture recognition. The top layer contains scripts in Unity3D Game Engine, which renders and animates the ECAs and contains the virtual environment (scenery) in which they appear.

Figure 14. Detailed architecture of the software interfaces and feature handling.

The agent is represented with her environment in the Interactive Systems Group Immersion Lab. She is projected at actual human scale on a wall.

4.1.2 Markup Language Definition

To conduct this research, I needed agents and systems that can maintain high levels of engagement with humans, in some cases, over multiple interaction sessions. These sessions can potentially extend to longer periods of time to examine long-term effects of the virtual agent’s behaviors. Our current ECA interacts with humans in a game called “Survival on Jungle Island.” In this
game, users interact with our agent across several scenes. Each scene comprises a collection of speech input, speech output, gesture input, gesture output, scenery, triggers, and decision points. Our prior system (Novick and Gris, 2014) was developed with procedural code, which did not lend itself to rapid extension to new game scenes. So to enable effective authoring of the scenes for the “Jungle” game, we adopted a declarative approach. We developed ECA middleware that parses, interprets, and executes XML files that define the scenes (Novick, et al., 2015). This section presents the XML coding scheme and its implementation and describes the functional back-end enabled by the scene scripts.

Authoring of scenes in which ECAs interact with humans currently has limited technological support. Scenes tend to be written directly as a computer program that specifies procedurally how the ECA should behave. As the number and complexity of scenes increases, writing the scenes becomes correspondingly more difficult and more prone to the sorts of problems associated with unstructured code. Accordingly, the research team and I developed a high-level script to write scenes less like writing a computer program and more like writing a script for a play.

Writing scenes for ECAs, however, is much more complicated than writing a play because the scenes have to account for the technical details of the agents’ inputs and outputs through speech and gesture and the technical details of the agent’s virtual world. Ideally, ECAs can interact with human beings in such a way that they require little to no human intervention to complete tasks and or to navigate conversations. But writing scenes for this is a daunting task, especially when the interaction space involves shared reality, where the virtual side of the inter-action contains elements with which both agent and human can interact). High-functionality ECA systems have to include gesture recognition for task-based processes, gesture recognition for general annotation, branching, 3D movement in virtual space, speech, and other sophisticated features.

Our current system, “Survival on Jungle Island,” is much more complex, though. Consequently, development of “Jungle Island” using the procedural representations and traditional software engineering techniques we used for the “Vampire King” game would have made writing the game’s many scenes unreasonably complicated and burdensome. This meant finding a much more efficient way
of developing scenes for ECA systems, a way that would enable us to write dozens of scenes with only modest technical effort beyond imagining the scenes’ substantive content.

The requirements of developing systems like “Jungle Island” led us to adopt a declarative approach to authoring scenes for human-ECA interaction. In our approach, developers write scenes as scripts, using a mark-up language represented as XML tags, and these scripts are interpreted by middleware to drive the real-time system.

Scripts using the mark-up language rely on twelve XML tags to identify agent behaviors and scene flow. A scene is a session of interaction, typically designed to take four to five minutes, that includes multiple human-ECA exchanges. Each exchange is called an episode, and the flow of the dialog within a scene is man-aged by choosing different episodes as a function of users’ responses.

In developing our ECA system, we aimed for a design that promoted the sensation of immersion in the virtual world and the naturalness of the interaction between human and ECA. This led us to adopt a robust approach that would provide scene developers with control of scene progression by adaptively enabling or disabling speech-recognition grammars for specific dialogs, gesture recognition libraries for specific gesture event, and control of what happens in the 3D scenes.

Authors specify scenes declaratively with XML tags such as scene (a collection of episodes), episode (a spoken and/or gestural interaction concluding with a control-flow decision), do (a gesture), speak (an utterance, specified via a string that can be output through a voice synthesizer, with an optional file name for a recorded version of the utterance), pause (for some number of seconds or fractions of seconds), decide (a conditional construct based on speech or gesture input from the human, go (to an episode), and nextscene. Table 4 presents the tags, their parameters, and their functions.
Table 2: Agent scripting language definition

<table>
<thead>
<tr>
<th>Tag</th>
<th>Parameters</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>scene</td>
<td>scene name</td>
<td>Name of the file. Corresponds to a 3D environment in Unity with the same name</td>
</tr>
<tr>
<td>episode</td>
<td>episode name</td>
<td>Name of the episode. Episode tags can contain any tag except scene</td>
</tr>
<tr>
<td>do</td>
<td>emotion, hands, locomotion and upper body</td>
<td>These tags specify the agent’s animations. You can specify an animation for the face (emotion), hands, lower body (locomotion) and upper body</td>
</tr>
<tr>
<td>speak</td>
<td>wav file and speech string</td>
<td>Contains either a .wav file name as a string, and/or the text of what the character will say</td>
</tr>
<tr>
<td>decide</td>
<td>speech id or gesture id</td>
<td>Makes a dialog branching decision based on user’s speech or gesture input</td>
</tr>
<tr>
<td>go</td>
<td>target episode</td>
<td>jumps to a new episode upon the conclusion of the current one or after a decide tag conditions have been met</td>
</tr>
<tr>
<td>pause</td>
<td>seconds</td>
<td>Pauses the interaction. Useful to avoid further input while the agent performs a lengthy animation.</td>
</tr>
<tr>
<td>createrule</td>
<td>grammar rule name</td>
<td>Sets the name of a grammar rule.</td>
</tr>
<tr>
<td>tag</td>
<td>grammar tag</td>
<td>Serves as the name of a grammar slot to which several items are mapped to. Each grammar rule can have any number of tags</td>
</tr>
<tr>
<td>items</td>
<td>grammar items</td>
<td>Items are the words users</td>
</tr>
<tr>
<td>ruleCall</td>
<td>grammar rule name</td>
<td>Calls a previously defined grammar. After tags and items were provided, grammars can be reused in any section through rule calls</td>
</tr>
<tr>
<td>nextscene</td>
<td>scene name</td>
<td>The name of the new scene to be loaded</td>
</tr>
</tbody>
</table>

The scene interpreter, implemented in C# for convenient integration with other modules of the ECA system, parses the XML script, creates scene, episode, and action objects corresponding to the script’s tagged elements, and sends the objects to the ECA’s run-time system for execution.

4.1.3 SAIBA, BML and FML

Our approach contrasts with that of some other standards for specifying ECAs and their interactions. In particular, our system differs from the SAIBA framework (Zwiers, et al., 2011) so that our ECAs can more easily integrate cognitive functions across what would be separate modules in a SAIBA-compliant system. SAIBA’s strong modularity may not be appropriate for interaction that is fine-grained or requires rich contextual information (Lee et al., 2008). Similarly, the strong theoretical frameworks of BML (Kopp et al., 2006) and FML (Heylen et al., 2008) tend to impose cognitive and operational models that may be over-detailed and limiting for ECAs intended to explore the frontiers of human-agent interaction in practical terms. In a sense, architectures such as BML and FML handle the
interaction from the agent’s point of view. In our case, we need additional control over the environment and its navigation for actions by both the user and the agent.

```xml
<utterance scene="map1" speaker="person1"><clause>
  <theme><turn type="take">
    <action><new> give </new></action>
    <object> him </object></turn></theme>
  <rtheme>
    <turn type="give" target="person2">
      <emphasis type="phrase">
        <reference type="visual" target="map:mine">
          <reference type="textual" source="person3">
            <object id="map:mine"> some

            <emphasis type="word">
              <new> gold </new>
            </emphasis>
          </object>
        </reference>
      </emphasis>
    </turn>
  </rtheme>
</utterance>
```

Figure 15: An FML example which models the intentions and behaviors for the phrase “give him some gold”. This illustrates the complexity and versatility of the language, but it also shows how it makes it impractical to use in a 40 – 60 minute interaction system, for which a greater degree of abstraction would be preferable.

Each scene is contained in a file. Each scene can have any number of episodes. Each episode can contain tags that enable the agent to speak, listen, decide, act, and then jump to other episodes. In addition, each scene corresponds to a physical, navigable 3D representation of the environment. By changing scenes, one can change the scenery, and even the agent, if required.

Inputs from humans to the ECA system come through speech and gesture. The author’s parameters for the decide tag determine how the system processes these inputs and chooses the next episode in the scene.

This approach has limitations of design and implementation. In terms of design, our approach is fundamentally limited by its orientation toward scene-driven interaction. For human-ECA systems that are more about spontaneous, less-constrained interaction than episodic storytelling, an FML/BML approach might be more appropriate. In terms of implementation, we found that rewriting scripts to conform to updates in the specification of the markup language or changes to our file-naming conventions turned out to be tedious. The team is currently working on improving this system by
including a tool for automatic updating of scripts to conform to changes in the scripting specification and a tool to update filename changes when you update the script. Some additional improvements include creating an XML compliant system, renaming some redundant or ambiguous tags, and including parameters for greater animation flexibility. The full schema can be found on Appendix E.

4.1.4 Speech and Language Processing

In this section I review the speech modules, in particular the *speak* and *decide* speech tags. I begin with the *speak* tag, which specifies the ECA’s spoken language output. To illustrate the scripting language’s handling of speech, Figure 16 presents an abridged episode from an early scene in the “Survival on Jungle Island” game. In this episode, as in all others, speak tags contain both a written and optionally an audio version of the dialog specified; for example, in the script below the file decideu1.wav is a recorded version of the utterance. Because authors can use text strings without speech files, they can quickly prototype and test scenes using synthetic speech before committing to recordings. When executing behaviors indicated with a speak tag, the system uses the audio file name to find the audio for playback and has the agent lip-sync to the audio through a dialog tree, implemented as a separate module.

The *decide* tag signals a branching condition, which determines the next episode. The agent asks a question and provides two or more possible answers. When the speech-recognition module recognizes one of the answers with high confidence, the system branches to a named *episode*, which is a parameter for the go tag. In the example in Figure 16, “grassy” and “rocky” refer to other episodes of the scene, to which control will pass depending on the human’s answer to the ECA’s question.
Figure 16: A scene section using our markup-language. This script causes the agent to look thoughtful while asking a question (through a predefined audio file). It then pauses for a second to prevent the speech recognizer from listening to the playing audio file. Finally, a grammar is created inside a decide tag, which enables users to progress to the next episode or scene depending on their verbal expressions.

To improve reusability, the decide tag can transfer control to any episode in the scene, enabling authors to create dialogs that loop until certain criteria are met. The grammar is a list of items and contains every word or phrase that can be recognized by the system and is not affected by the current state of the conversations. It is composed of rules, and each rule defines what words or phrases should be recognized during a particular dialog state, which is what we call an episode. As the control moves to a new episode, the rule likewise changes to enable a different set of words available for recognition. Grammars are automatically created based on the rules defined on the scripts and are saved as separate grxml files. The idea behind this rule-based grammar, in which the rules act as speech-recognition separators for episodes, is to enable only a small subset of the whole grammar to be active at any given time. This helps improve speech recognition by not engaging in grammar over-coverage.
By creating a rule, an author can refer to a speech-recognition grammar in future episodes without the need to declare the grammar again. In other words, if a rule has been declared, it can be called and reused again from the script by just specifying the name. If the rule is being used for the first time, its name, tags, and items must be declared. The only requirement is that the grammar must be used at least once before being reused.

The scene markup language enables authors of scripts to develop dialog paths that can vary depending on the nature of the application. In a Q&A application, the dialog path will revisit states multiple times and in any order, as shown in Figure 17a, as a fully connected graph. In a storytelling application, the dialog path will follow a decision-based progression, as shown in Figure 17b, as a dialog tree. The markup language enables both forms to be created by adjusting parameters and rearranging the episode order.

In the markup language, tags are flexible and new tags can be added; however, changes made to the structure will require the users to update their scripts. The language enforces consistency but does not provide it automatically.

4.1.5 Gesture Recognition

For gestures, the decide tag connects to the ECA system’s gesture-recognition module. In the episode in Figure 19 the agent waits for the user to extend his or her hand towards the screen. The gesture is then captured and processed by a Kinect and our gesture-recognition software. Similar to the
way the decide tag works for speech, this list can contain several gestures leading to different episodes, depending on the gesture performed by the human. Depending on the gesture that the user performs, the decide tag will transfer control to the appropriate episode. This is also useful to create activities that require a gesture response to progress. Gestures are specific to the application; in the case of the “Jungle Game”, they represent survival activities, such as lighting a fire, spear fishing, and signaling for help.

Recognition of body gestures has long challenged developers of interfaces for real-time interaction between humans and embodied conversational agents (ECAs). Full-body gesture recognition provides natural human-computer interaction in applications such as embodied conversational agents (ECAs). However, this approach to interaction remains difficult to achieve due to low recognition accuracy, distant sensor positioning, performance issues in real-time processing, unobtrusive interactive tracking solutions, and the expense of capturing motion for representation of gestures.

To speed the process of capturing human gestures for purposes of generating ECAs’ gestures and of recognizing the gestures of the ECAs’ human conversational partners, we developed a tool that is capable of recognizing full-body gestures in real time and can generate pose libraries for recognition across applications.

Although most applications are often controlled through a computer screen and a traditional keyboard and mouse setup, some ECAs, such as those developed in our lab are life-sized projections of virtual human characters. The more detailed the information about the user’s non-verbal actions are, the better the agent can interpret and more accurately react to them (Bailenson and Yee, 2005; Gratch, et al., 2006). We wanted to avoid intrusive systems, those that often require users to wear special suits or markups to be detected by a set of several cameras positioned across an empty room, which would affect the immersion of the interaction.

In most motion-capture software there is no further analysis of the gesture data after it has been recorded, which makes impractical the identification of gestures and reactions in real-time to these gestures.

To address these problems, the research group and I built a tool that suits specifically the full-body gesture recognition scenario while standing at a distance of six to eight feet away from the sensor.
The tool is capable of recognizing full-body gestures in real time and can generate pose libraries for recognition across applications. The tool can detect large sets of unique gestures, and users can create, export, and import these gesture sets. The tool can be used not only to interact with different applications but also to generate log files as spreadsheets that record the users’ behavior across time.

This application is connected to the markup language and interpreter middleware system to enable external applications to access pose libraries and gesture detection. In addition, the team created a user interface that enables users to build pose libraries based on screenshots of the desired poses and that has additional features aimed at improving accuracy through basic statistical analyses. Figure 19 shows a human performing a “hi-five” gesture that is recognized and interpreted by an ECA.

4.1.6 Gesture Capture

For the system to understand a pose, there must already be information about the pose to be recognized. The team created a pose library that contains an array of pose objects.

To automate this, the research team created a separate module, the gesture capture tool. The capture tool enables the developer to select the relevant body parts for the pose capture, as shown in the left side of Figure 20. Then, it enables the capture of those joint angles. The process requires the developer to click a capture button while the pose actor is representing the pose in front of the sensor. In addition, the tool enables developers to capture the same pose several times from the same or different pose actors to improve accuracy. As different people do the same gestures differently, or the same person might slightly change posture between one attempt and another, the capture tool collects the data, analyzes it by calculating the maximum, minimum, and average angles, and estimates a range parting from the mean of the angles required to recognize this gestures in the majority of cases. This effectively calculates a margin of error, depending on the variety of poses captured.

The generated file is also in XML format and contains tags for every joint ID, its average angle, and its suggested margin of error.

After this process, the pose is then added to the pose recognition library and can now be named and detected. The resulting string of the detected pose can then be used to trigger events in other applications. Once the poses are stored in the library, we can build gestures from them. Because gestures
require movement, a gesture is defined as a sequence of poses. When the user follows the pose sequence, the tool detects the gesture. At its current state, the tool generates pose sequences that are insensitive to time. In other words, a gesture will be recognized when the human follows a pose sequence regardless of the speed with which it is executed. This is not optimal, as gestures can vary in meaning depending on speed of execution. In the future we will integrate timers that can be set between poses to add greater precision to the gesture recognition.

Another concern involves overlap of gestures. When gestures are not well defined, their margin of error might be higher than usual. If this happens across several gestures, there might be subsets of coordinates that fall between one or more gestures, making the recognizer unable to decide which gesture was actually executed.

Although the tool is limited in terms of dimensional space and post-processing data handling, it has proven to be useful and reliable for our current applications. Provided that there is post-processing of the pose library to minimize overlap, the tool performs well even though it is a lightweight application in comparison to commercial motion-capture systems or other recognizers that are unable to process information in real time. As it is, with its ECA front-end applications, the tool is applied to real-time interaction, real-time video annotation, and pose analysis.

4.1.7 Animations

In this section I explain the implementation of the tags for performing and recognizing actions. The do tag accesses the animation tree to produce a blend of animations, presented in the next section, to produce the intended motion for the agent. The system uses a layered approach to animation that enables developers to specify the agent’s actions with enough control to represent the desired situation and emotion but with enough abstraction as to avoid getting lost in detailed descriptions of animations. Movements of the body are specified in layers for locomotion, emotion, upper-body, and hands.
Figure 18: In this scene the agent uses a combination of animations to extend its hand to the user. The agent then prompts the user to grab her hand through a decide tag. With only one option, it is a required gesture to progress through the interaction. After the user performs a “Grab Hand” the control flow decides on the next episode to execute. At the end, the scene changes. On the virtual side, a new environment is loaded.

The *episode* in Figure 18 uses a locomotion layer, which enables the agent to walk around, move to the next waypoint, sit, crouch, crawl, or do any other movement that requires character displacement across the scene or lower body animations that change the current stance. In the example, the agent starts the scene by moving to a standing position, which is a valid transition because the default starting point for this particular scene was specified as crouching, even though agents generally start scenes in a standing position.

The *emotion* layer controls the configuration of the agent’s face. All transitions are weighted and transitioned gradually between emotions; the transitions are handled by the system’s additional animation-tree modules.

The *hands* parameter indicates the position or shape of the fingers. This is used for special circumstances such as pointing or making representational hand gestures.

Finally, the *upperBody* tag specifies the action of the torso and arms. In the design of the ECA system and the mark-up language, the team sought to avoid unnecessary or overspecialized animations. In the case of the scene in Figure 18, the agent is extending its hand towards the user to help him or her cross a gap in the virtual scenery. The upper-body parameter is pointing, because this enables reuse of another animation. A pointing animation includes an outstretched arm; without the hand performing a
pointing gesture, too, the pointing gesture is just an extended arm with an open hand. This is one of the ways in which work is reused to create new interactions without additional low-level development work.

After parsing these gestures, they become available as they were specified, in an ordered sequence when called by the episode in control of the dialog state. However, developers can incorporate (outside of the markup language declaration) conditions or variables that affect the agent’s gesture behavior depending on the runtime observations of the interaction. For example, a developer can specify within an episode a gesture containing indication of surprise for the upper body animation layer, and indication of happiness for the facial animation layer. If, however, this normally happy moment is obscured by a series of user transgressions that the developer kept track of throughout the interaction, the do tag queue can be accessed at runtime and animations can be overwritten.

In animating ECAs, run-time blending of animations can provide a large library of movements that increases the appearance of naturalness while decreasing the number of animations to be developed. This approach avoids the need to develop a costly full library of possible animations in advance of use. With our animation module, rather than creating over-detailed, fine-grained procedural animations or hundreds of motion-captured animation files, animators can include sets of their own animations for agents, blend them, and easily reuse animations, while constraining the ECA to use motions that would occur and transition naturally.

Perhaps some ECAs do not feature realistic animation because there is little need for naturalistic animation in most computer applications. Most commercial computer applications are metaphor-driven, where users interpret an icon’s image and associate it with its function without requiring a highly realistic animation. For example, users do not see or need a hyper-realistic 3D model of animated scissors every time they click on the “cut” function in common office applications. Likewise, users may not expect realistic animations for characters, when interpretation alone can suffice. Instead, users have come to expect non-realistic characters with limited functionality whose actions users have the responsibility to interpret.

Ironically, as agents become more realistic, the unnaturalness of their animations can become more evident. While some animation systems produce realistic and smooth movement for ECAs,
research and development of the agents typically focus on the appropriateness or categorization of gestures in taxonomies that define what should be displayed in response to different user behaviors. The responsibility for animation quality is often delegated to an artist, who is limited by the requirements of the particular agent and who has little creative freedom to improve the expressiveness of gestures. As a practical matter, this approach produces rough-and-ready results but does not provide specific guidelines or replicable techniques for producing ECA animations that are appropriately realistic. What is needed, then, is a systematic approach to realistically natural gesture animations for ECAs that can be applied by developers who are not experts in animation.

To respond to this need, the research team developed an automated animation system that can create a wide range of realistic animations based on a small set of states that can be blended, combined, layered, and transitioned. The team’s approach is based on an adaptation of principles of lifelike animation and is implemented via an animation graph in Unity’s Mecanim system. This approach contrasts with that of other ways of specifying ECAs and their interactions. Systems such as BEAT (Cassell, et al., 2004) and SPARK (Vilhjálmsson, 2005) provided a remarkable amount of detail for gesture based on discourse analysis, but unfortunately these approaches require large sets of micro-animations. And trying to save effort by generating these animations procedurally causes the animations to appear robotic and unnatural.

4.1.8 Animation Principles

Traditional animation theory has suggested twelve principles for creating “the illusion of life” (Thomas & Johnston, 1981). In order of importance, the twelve principles are squash and stretch, anticipation, staging, pose to pose, follow-through or overlapping action, slow in and slow out, arcs, secondary action, timing, exaggeration, solid action and appeal. Although these guidelines were meant for artists, applying the principles to ECAs highlights flaws in standard approaches to animating agents. Accordingly, I describe in detail each animation principle and how they were applied to our ECAs.

*Squash and stretch*. This principle provides the illusion of weight and volume. It is particularly useful in animating dialogue and facial expressions. In artistic terms, this is the most important factor, as it ensures that the character does not change shape and represents the character’s weight. For ECAs, it is
often preferable to simulate real-world physics so that users can better relate to the agent by acknowledging that the same real-world rules apply for both human and ECA. This principle might not apply to agents that are not human-like.

**Anticipation.** This principle involves the character’s movement in preparation for major actions that they are about to perform. Common examples include running, jumping, or even a change of expression. Almost all real action has major or minor anticipation, yet this is one of the most overlooked animation principles in ECA development. Agents often need to react in real time to users’ actions, often involving both speech and gesture. So by the time a system recognizes the user’s communicative behaviors and formulates an appropriate response, the system is likely to have to perform immediately the agent’s main gesture motion response, leaving no time to perform the anticipation animation. This is a key cause of agent actions appearing to be robotic, as it creates a move instantaneously with seemingly no previous thought or intent. To overcome this obstacle the system has to have a set of anticipation animations that can be used flexibly by predicting the animation that will be required; even as a broad guess can provide a more realistic animation through anticipation than omitting the anticipation animation.

**Staging.** This principle states that a pose or action should clearly communicate to the audience the mood, reaction, or idea of the character as it relates to the story and its continuity. Staging, in other words, directs the audience’s attention to the story or the idea being told. This represents a problem for ECAs because cinematography often uses camera angles and close-ups to make certain story elements salient for the audience. In addition, care must be taken when building scenery and backgrounds so that they do not compete with the animation for the audience’s attention. The main problem with staging, though, is that ECAs often do not have a proper stage on which to perform. This is a design issue that should be addressed early in development. In experiments (Gris et al., 2014), we have used staging techniques by providing our agents with a virtual environment. This approach has led to decreased memory loads and higher attention rates (Gris et al., 2014).

**Straight-ahead and pose-to-pose animation.** Due to the nature of 3-D animation software, we use pose-to-pose animation via key frames. But this does not mean that that animations should be planned
simply by creating an initial pose and a final pose, and then letting the 3D software interpolate the sequence automatically. Planning of animations should include transition between poses in the most natural way possible. To achieve this seamless effect, animations should be designed with the proper length, looping poses, and interruption segments so that animations can be combined or transitioned at any time.

*Follow-through and overlapping action.* In simple terms, this principle means that nothing stops all at once. When an agent is moving and the animation ends, the agent cannot simply go to a static pose. Rather, it should blend with a weight shift or another appropriate motion that provides the illusion of continuous realistic movement, even though the main underlying action has stopped. Following this principle can eliminate unnatural movements such stopping in the middle of a step while walking.

*Slow in and slow out.* This is one of the most important principles for agents who do conversational turn-taking. This principle enables animators to soften the animation and make it more lifelike. Similar to anticipation, attention is drawn at the beginning of the animation and at the end of the animation, as these are often indicators of turn-taking. At the beginning of the motion, people will usually try to guess what the reaction will be, and at the end people need an indication to know that the agent is about to finish and that the users will able to jump into the interaction again.

*Arcs.* In 3D modeling all organic characters, including humans and animals, are made of curves. In contrast, robots are made of sharp edges. Animating organic characters follow arcs or slightly circular paths because of the nature of our joints. This principle applies to almost all physical movement, including turns, eye motion, and walking paths. Accordingly, movements performed in a straight line will seem robotic or unnatural.

*Secondary action.* The principle of secondary action applies to actions that enrich the main action. Secondary action adds character. For example, envision a male character about to invite a female character on a date. If it approaches the female character slowly often changing direction, it gives the impression of a shy and unsure character. However envision that same character with many concurrent secondary actions, such as fidgeting and frequent gaze movement away from the target character. These
actions may not be necessary, but they enhance the animation by making it more obvious, clear, and natural (Lasseter, 1987).

**Timing.** This is a highly subjective principle for animations that establish mood, emotion, and character reaction to the situation. For example, a pointing animation displayed in a faster timing can indicate a sense of urgency, while a slow pointing animation can indicate laziness or lack of interest. Most animation systems for ECAs enable developers to specify the time the animation should be running, but it is harder to find systems that can accelerate the animation, which is equally important.

**Exaggeration.** In a play, actors do not behave like normal human beings. Instead, they exaggerate their gestures to make them observable by the audience. While developers of agents commonly try to make the animations as human-like as possible, there is a potential benefit in exaggerating animations to make them more noticeable. Of course, one must be careful not to over-exaggerate; the risk is that of becoming too theatrical or excessively animated. Although some research has examined how big animations should be, based on the introversion or extraversion of the subjects (Neff et al., 2010; Hostetter et al., 2012), the field still lacks detailed guidelines for gesture amplitude. Artists’ suggestions to use good taste and common sense are not very helpful in this regard. Moreover, perceived gesture amplitude depends on the distance from the display, the type of media, and the screen size (Detenber et al., 1996; Loomis et al., 2003).

**Solid drawing.** In the 1930s, animators used this approach to create a three-dimensional space from the two dimensions of drawings on paper. In the case of animated agents, though, this principle is redundant because agents’ 3-D models already include solid drawing.

**Appeal.** This principle usually refers to visual attractiveness, but it can also be used as part of the character design to provide an agent with clear intentions. This is another highly subjective trait that should not be underestimated. In one of the research team’s pilot studies, users perceived an agent who was supposed to help users in a game scenario as being actually a potential enemy. Our intention was, of course, to convey a helpful, friendly character. But due to our lack of attention to the agent’s visual appearance, our animated agent conveyed the opposite impression. In general, this is a trait that
combines visual representation, dialog, context, and gestures, and it is difficult to achieve when any of these elements is missing.

Although all twelve principles were developed originally for traditional animation and to design characters for passive audiences that do not interact with agents, they still help in developing agents that are more relatable, believable, and accurately representing what we want our agents to convey.

4.1.9 Animation Blend Tree

To enable developers of ECAs to use the twelve principles of animation, as adapted for agents, the research team developed an animation design approach implemented as a tree of blendable animations. The system, based on Unity’s Mecanim, creates an autonomous animation workflow.

In Mecanim, the animation tree is a state machine where each state contains an animation and a weight. The weights of the animations are used to transition between them or blend animations that affect the same body parts. When there is a single animation being executed, its weight is equal to one. Every animation that is not being executed has a weight of zero. Each state transition transfers the weight of the current state towards the next intended state, until the new state has a weight of one and the previous state has a weight of zero. This means that at the midpoint of the transition each animation is being displayed with half the strength, containing properties of both animation states simultaneously. These transitions are not themselves states but rather the equivalent of a cross-fade between animations.

Using Mecanim’s animation tree, we created, analyzed, and weighted a set of animations so that each state transition enforces adherence to the animation guidelines. That is, the structure and design of our implementation of the animation tree enables end users to link animations together while limiting the linking of unnatural animations (e.g., going from the running state to the crouching state), taking care to make transitions realistically gradual. Each state can include movements from any or all of the body-part layers. The blend tree can be expanded to include more movements.

In addition, the tree contains a set of layers that provide additional control for running animations procedurally in real time by enabling users to specify animations for fingers and hands, facial expressions, and upper and lower body separately. This effectively enables us to combine a happy facial expression while pointing with the left hand and walking, where otherwise it would require a specially
designed animation that would be useful only in this particular situation. Although it can be argued that crafting specific animations for each situation produces a more accurate representation of the intended movement, creating all of an agent’s possible animations in this way be impossible because the combinatorial explosion of the movements of the different parts of the body. In our approach, the goal of the layers is to enable a maximum number of combinations from a small subset of independent general animations, while maintaining the naturalness and movement flow described in the twelve guidelines.

Figure 19 presents a subset of our animation tree. (I note that animation tree is the official name for Unity Mecanim structures, but our modifications effectively turn this “tree” into a graph.)

Figure 19: A section of the animation tree detailing the permitted locomotion transitions for lower body movement. The labels of the animations differ in style (e.g., “Running” vs. “Jump”) because some animations are continuous loops (e.g., running, walking), and others are one-time movements (e.g., jump, stumble).
Given Mecanim’s basic architecture, the twelve animation principles, and our proposed solution to the combinatorial explosion, we developed our particular animation tree for ECAs based on a series of observations of natural conversations. Our corpus contained twelve dyadic conversations. Each conversant was recorded in an otherwise empty room using two Kinect sensors that recorded side and front views for each participant, who stood facing each other. Each recording contained video and depth information, and this information was interpreted by motion capture software and translated to one of our agents, a full-body virtual 3D character. Each animated file was then empirically analyzed and categorized based on movement similarity. Figure 21 shows the stages of motion capture.

Figure 20: From left to right, video and depth noise from the participant, motion capture data, and agent with animation data incorporated. The top right displays a thumbnail view of the raw front and side video as processed with iPi Soft.
Because the team sought to preserve naturalness of movement, we focused on the mechanics of the gesture rather than the gesture itself. That is, we studied the movements preceding the main gesture. For example, before performing an action such as pointing, leaning, or explaining, participants often prepared for several seconds before performing the intended gesture. Based on these observations, we classified the possible transition points between gestures, eliminating many possible but unnatural combinations that violated the animation principles. We also layered the captured animations by dividing the virtual character’s body into lower body (locomotion), upper body, head, face, and hands. Additional layers were created for lip-sync and blinking controls; these, however, did not come from the motion capture analysis. Figure 2.1 presents the animation layers corresponding to the agent’s body regions.

The special layer in Figure 2.1 is reserved for when there is the need to create an animation that is superimposed over all others, for example, a backflip, dancing, or other activity that requires simultaneous full-body coordination. Parameters detect the last pose that was animated and blends the new animation into it. If an animation has taken place but has not transitioned to a new state, the tree limits available movements. For example, if the agent sat down, it must remain sitting and cannot perform any other lower body animations until it stands up. In our approach, exceptions to this rule,
allowing an agent to sit while pointing or explaining, for example, must be explicitly identified via connections in the tree.

4.1.10 Virtual Environment Design

Development of ECAs has tended to focus on the character’s dialog capabilities, with less research on the design and effect of the agent’s voice and of the virtual environments in which the agent exists.

Changes in (1) an ECA’s virtual world and (2) an ECA’s voice can affect the way in which users interact with the ECA. We compared three versions of an immersive video game in which the user interacts with a life-size ECA, using the research team’s “Escape from the Castle of the Vampire King” game.

4.1.11 The Agent’s Voice

The quality of an agent’s speech is measured by the degree to which it replicates a human speaker. While there has been significant progress in this area, users are still unenthusiastic about most synthetic speech (Newell & Edwards, 2008). We hypothesized that increasing voice naturalness would make users less likely to exhibit frustration and interrupt the agent while it speaks. The first version of our game used Anna, the default voice provided by Microsoft operating systems. Twelve subjects interacted with the agent in two sessions each. After each session, subjects completed a survey that included an optional open question on what would they like to see improved. Subjects indicated that the voice was unclear, hard to understand, robotic, unemotional, unengaged, insensitive, monotone, and broken. For the second version of the game we used Salli, an American English voice from IVONA. For this version of the system, 22 subjects played the game, again across two sessions each. Most user comments noted a lack of emotion in the voice. In one participant’s words, “Even the smallest hint of empathy would greatly improve her demeanor.” Systems have been built for emotional synthesized speech (Schröder, 2001). Nevertheless, emotion in synthesized speech in real-time multimodal conversation is still in its infancy, and we were unable to find an emotional synthesized voice that we could satisfactorily adapt for our agent. For the third version we recorded over 200 different utterances, which were played in place of the synthesized voice. Only 4 out of the 58 participants complained about
a lack of emotion. We expect that the residual perception of lack of emotion arose from the use of emotion-neutral phrases that were used in multiple game situations.

As the agent’s voice improved, other issues of users’ engagement with the agent became more apparent, especially with respect to the relatively impoverished nature of the virtual environment in which the agent was presented. In our previous experiments across the three versions of our vampire game agent, we measured engagement by tracking gaze away from or toward the agent.

Table 3: Data from voice analysis.

<table>
<thead>
<tr>
<th></th>
<th>Version 1</th>
<th>Version 2</th>
<th>Version 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>User interruptions of agent</td>
<td>27</td>
<td>60</td>
<td>14</td>
</tr>
<tr>
<td>Agent utterances</td>
<td>871</td>
<td>878</td>
<td>948</td>
</tr>
<tr>
<td>Gaze at agent (seconds)</td>
<td>1453.6</td>
<td>3759.9</td>
<td>9603.5</td>
</tr>
<tr>
<td>Gaze away (seconds)</td>
<td>5793.8</td>
<td>3616.3</td>
<td>1444.1</td>
</tr>
</tbody>
</table>

### 4.2 Agent Features and Design

Given these results, a significant effort was placed in developing a high-quality interactive environment, and a sound story surrounding the agent. This helped improve the sense of awareness of the agent by creating a tight context that worked seamlessly with the story to create an engaging interaction that would allow us to represent, measure, and evaluate rapport in a setting that resembled a human-human interaction with the highest degree of automation possible.

#### 4.2.1 Interaction and Storytelling Design

The Survival on Jungle Island game comprises 23 scenes, which are interactions delimited by the environment and a concrete topic. All scene scripts can be found in Appendix A.

The first two scenes are meant to be instructive for the players, without being an obtrusive tutorial that breaks the impression of an immersive reality. These scenes are followed by a cinematic sequence, which is a non-interactive video, offers some contextual background for the situation. The cinematic shown in Figure 23 explains, through the voice of a narrator, how your ship is sinking in a
storm. The few sailors were washed off the deck by to a giant wave. The captain is unwilling to abandon his ship but sends you away with the hope that you might survive.

During the first scene, the player wakes up on a beach with someone --the agent-- looking at them. The agent introduces itself and ask a few questions, including if the participant is fine, if they can walk, and if they think there are other survivors. These questions help the user get immersed in the interaction.

There are two versions of the game, one which requires gesturing to perform some tasks, and another where the agent performs the tasks and requests feedback usually with a brief confirmatory dialog. If during the first scene the player is in the gesture-enabled condition, the agent will help them stand up by offering her hand. This is an important initial gesture, first, because although the participants are already standing, they understand that it is not necessary to walk around the room to be properly recognized, and that some gestures do not require crouching, laying down, or sitting. From the technical end, avoiding these situations increases our gesture-recognition accuracy, as all of the system’s gesture recognition assumes a standing participant. These types of requests from the agent are infrequent and were intentionally avoided when developing the storyline. These gestures are requested only when a virtual environment object makes it relevant (e.g., while climbing a virtual mountain, the agent has to make the user believe he or she is climbing). This also increases the user’s perception of the agent’s sense of awareness by realizing that gestures are recognized and reacted to.

One of the main system’s limitations is that there are no reactions for a lack of gesture over time. This is intentional: adding dialog that complains about lack of gesture would differentiate one version from the other significantly. For example, if the agent offers her hand during the first scene to help the player stand, and the player rejects the gesture, the ECA’s natural reaction should not be conductive to rapport. This would be more natural and believable, yet it would differentiate the interaction from the non-gesture enabled, where the agent would have no way of engaging in this type of behavior because the user was not allowed to refuse the action. A similar situation could possibly be implemented for the non-gesture version, where the lack of verbal response would lead to the same outcome. But this would introduce additional complexity to the interaction. Instead, we opted for a reminder, where if the user
does not respond after some time, whether a verbal is or a gesture response is being requested, the agent will remind the user after ten seconds of inactivity by repeating its last utterance. This is not an ideal solution, but it standardizes the interaction between subjects.

During the first scene, in addition to several of the open ended questions mentioned in the previous paragraph, the agent will introduce herself and ask for the player’s name. Most open-ended questions are not handled in the system by a NLP model. Instead, they are treated as wildcards, and the interaction is designed to traverse a dialog path through which any answer to those questions would seem recognized and valid (e.g., “Where are my manners? My name is Adriana. What is your name?”; after the user responds, the name is not purposely recognized, and the ECA checks merely for an audio input, to which her reply is always “Nice to meet you” and then continues the conversation). These personal questions seem to have seamless recognition for the users and are meant to be an indicator for the user that the agent is processing what it hears. This approach helps avoid command-like answers to questions, where users reply only with single words.

The purpose of the second scene is to progress the story by providing a plausible explanation why there is a person (the ECA) in the jungle and how she survived thus far. In previous experiments and pilot studies, the lack of context in these situations led people to believe that the agent was evil and had some ulterior motives. This is perhaps due to the nature of games, where there is normally a challenge, usually accompanied by some antagonist character. In the absence of any other ECAs, players may have assumed that the character they were interacting with would be part of the challenge.

This scene is also the first one to react appropriately to a user’s verbal response. At some point close towards the end of the scene, the ECA asks if they should stay on the same shelter or find a better place. Regardless of the user’s response, the agent will be inclined to find a better place, but the agent will acknowledge the player’s opinion by either agreeing or questioning if that is the best decision to make. This is meant to give the user a sense of presence for both the player itself and the ECA, as well as minimize the storytelling differences between playthroughs.
If the game relies heavily on branching, an experiment between subjects would be inadequate due to the different experience players might have, even in the same experimental condition. If the game offers no branching through, it creates the illusion of a linear story that disregards users’ input.

The scene path is the same for participants under both conditions, regardless of their choices. A virtual path in this context is the actual path where the agent walks and moves, along with the places and order in which human and ECA visit. Users’ responses to the agent’s inquiries might generate different feedback on behalf of the agent, but the scene will always continue through the same virtual path, leading to the same final outcome. To maintain the illusion that this is not the case, we incorporated a few exceptions, where players can take different virtual paths to accomplish the same goals. The earliest scene where this is possible is when the user and the agent are looking for water. While the user and the agent are inside a cave that they intend to use for shelter, they hear water droplets echoing. The agent is thirsty but afraid to explore the cave. The user can then choose either to gather water from the cave or to look for water somewhere else. If the user chooses the cave, they walk a few seconds and soon enough find water, whereas if they opt to look somewhere else they will find it outside the cave after walking a similar distance.

The scenes offer a controlled, scripted agent-disclosure timeline, meaning that each scene is written to incrementally disclose information about the agent and to request similar information from the user. This increases in frequency and intimacy over time, slowly building up the relationship between human and ECA. While looking for shelter, the agent will ask where the user is from (and reveal her own origins as well after the user’s response), and if there is anyone waiting for them back home. Much later, she will offer information about her own parents, her childhood memories, and her thoughts and hopes regarding the situation they are in. This establishes incremental intimacy and affinity at the verbal level. Since these key dialogs are static, meaning they do not change regardless of the path or previous decisions, it should not affect the extent to which players feel rapport with the agent.

Throughout the scenes, the players have the chance to learn about the agent’s current situation, her background, thoughts, fears, feelings, and likes. Some key scenes include their first encounter,
seeking and finding shelter, finding water, finding food, fishing, building an SOS signal, helping the agent after a death-threatening injury, and signaling a helicopter for help.

4.2.2 Agent Capabilities

As a summary of my contributions towards our ECA and virtual environments, I worked in the development of the core modules for speech and gesture recognition. I also designed the interaction, and contributed to the design of the scripting language for our agents. In addition, I built a network to connect all the different modules, although it has been significantly updated by the rest of the team. I also designed the animation tree, which controls animation selection and blending for the agent.

The ECA used in this experiment is not capable of performing the functions of an ideal agent as described at the beginning of this chapter. In particular, the lab setup makes it difficult to track and analyze facial expressions and gaze in real time. Natural-language communication is scripted and constrained to give the illusion of free dialog while actually being a guided conversation. Emotion is not perceived by the agent, and it is expressed by scripted voice acting to match each situation depending on the current scenario.

However, our agent does track and interact with people by reacting to most body postures and gestures. The agent is also capable of memorizing, (based on script decisions and carefully crafted storytelling) the events of every interaction. Our ECA is fully animated, so its movement can be easily manipulated to match a situation. Although not a perfect system, it proved to be advanced enough to interact successfully with humans for long sessions, to be engaging, and to generate feelings of rapport.

In the future, the research team is extending the AGENT Framework so that it can select an appropriate set of movements and coordinate its stance with the pose of the human participant, effectively mimicking and enhancing rapport to coordination.
Chapter 5: Data Collection

For this corpus the research group’s team and I recorded interactions between an ECA and a human. In this experiment, users’ actions were categorized, annotated, and analyzed.

5.1 Verbal Annotation

All personal information revealed by the user during the interaction was video recorded, some of which was gathered through a demographics survey prior to the experiment (see Appendix D). This information was also used by the agent to drive further inquiries about the user. The type of personal information gathered included where the user is from, their name, if they have family or friends, etc. Not all personal information was referenced, but information regarding the decisions they take during gameplay was annotated and caused path branching. Because the agent uses a very basic form of language processing, only the scene sequence, which is a direct consequence of their choices, was annotated.

5.2 Nonverbal Annotation

The only way to interact with the agent was by verbal or non-verbal actions. As there were no direct commands or explicit and visible interface elements such as text fields or buttons, the participants in the experimental condition had to make specific movements to progress through the scenarios. These tasks were simple and used for the purpose of maintaining the interaction immersive and to allow virtual manipulation of objects that reside in the agent’s virtual environment. From the player’s perspective, these tasks were accomplished by performing movements to follow the agent, look around the virtual environment, and to get across simple obstacles. These mandatory behaviors under the gesture condition are not annotated.

Rapport-building gestures are not necessary to advance through the chapters but through their frequency rate and amplitude across sessions can demonstrate and enforce the sense of physical connection. These behaviors can carry additional meaning. In other words, these behaviors are used to react to the agent’s paralinguistic behavior, or to invite the agent to interact. These behaviors are recorded, but due to the volume of data, only changes in stance were annotated for type and frequency.
To aid with the annotation tasks, the research team and I developed a flexible system using a Kinect™ that detects a set of human poses and gestures in real time and annotates them chronologically (Gris, et al., 2015). In addition, the system communicates with the ECA, allowing the agent to “see” the human and react in real time to full-body pose and gesture. Although the agent’s behavior is controlled and limited by its grammar and a predefined pool containing the set of movements from which it chooses its reaction, the user’s behavior is recorded and recognized positions are annotated in a log file (see Figure 13) while the agent reacts in real time to full body posture.

Figure 13: Annotation of nonverbal behavior during the interactions. The red line represents the agent and the blue one, the user.

The graph in Figure 13 is a representative file generated by the annotation system for nonverbal behavior. The Y-axis represents the gesture ID, and the X-axis is the timeline in seconds.

The Y-axis is not indicative of any gesture correlation. The purpose of these graphs is to visualize posture shifts across time and to compare them with the posture shifts of the agent, and then use these to detect users’ shifts that may have been caused by the ECA’s behavior.

Each file contains a graph and a respective table with fine-grained, more detailed timestamps, as the longer the interaction is the harder it is to visualize the interaction timeline. The corpus contains 23
generated graphs and tables per participant, one for every scene, for a total of 1472 annotated files, for an approximate total of 2368 minutes (39.5 hours) of interaction.

The timeframe portrayed in Figure 13 begins at the 26 minute mark, towards the end of the game. The graphs were segmented so that each file, table, and graph represent one scene. A scene is a unit of interaction that takes place on the same virtual environment (e.g., beach, cave, jungle). Each scene is divided into dialog units called episodes. The full structure of the game is discussed in chapter 5.

Table 4: Gesture IDs used to graph the interaction timelines. IDs are unique but do not follow a particular sequence. Missing ID numbers are due to animations or gestures that were not used for this experiment. The agent gestures table (left) contains gestures exclusive to the agent.

<table>
<thead>
<tr>
<th>ID</th>
<th>Gesture</th>
<th>ID</th>
<th>Agent Gestures</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Arms Up</td>
<td>23</td>
<td>Pointing</td>
</tr>
<tr>
<td>1</td>
<td>Hi Five</td>
<td>24</td>
<td>Thinking</td>
</tr>
<tr>
<td>2</td>
<td>Normal Stance</td>
<td>25</td>
<td>Stretch</td>
</tr>
<tr>
<td>3</td>
<td>Hands on Hip</td>
<td>26</td>
<td>Weight Shift</td>
</tr>
<tr>
<td>4</td>
<td>Crossed Arms</td>
<td>28</td>
<td>Throwing</td>
</tr>
<tr>
<td>5</td>
<td>Right Hand Left Shoulder</td>
<td>29</td>
<td>Explaining</td>
</tr>
<tr>
<td>6</td>
<td>Left Hand Right Shoulder</td>
<td>30</td>
<td>Surprised</td>
</tr>
<tr>
<td>7</td>
<td>Hands Front</td>
<td>31</td>
<td>Excited</td>
</tr>
<tr>
<td>8</td>
<td>Hands on face</td>
<td>32</td>
<td>Listening</td>
</tr>
<tr>
<td>9</td>
<td>Right Hand Head</td>
<td>33</td>
<td>Scared</td>
</tr>
<tr>
<td>10</td>
<td>Left Hand Head</td>
<td>34</td>
<td>Cold</td>
</tr>
<tr>
<td>11</td>
<td>Balancing</td>
<td>35</td>
<td>None</td>
</tr>
<tr>
<td>14</td>
<td>Hold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Ventilate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Spear Throw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Striking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Lift Hand</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Lift Hands</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For this experiment, task gestures and non-task gestures were configured to be exclusive; that is, the recognition will not pick up any non-task gestures while the user is in a task interaction state and vice versa. To count only the non-task gestures the automatic annotation module, which is a log file of every gesture, was post-processed to disregard flickering gestures, noise, and non-task gestures.
5.3 Demographics

I enlisted 80 UTEP students to participate in this experiment, out of which 64 provided usable data: 11 females and 53 males. Ages ranged between 18 and 36 years, with the mean age being 22.5, with a standard deviation of 3.77.

UTEP’s unique location provided a unique corpus as well, with 50 participants identifying themselves as Hispanic. To participate in this experiment participants had to be able to stand for an hour and perform full-body motions; in addition, participants were required to be fluent in English and not have any hearing impairments. Thirty three participants reported English as their first language, and 27 Spanish. There were only two exceptions, one subject reported Bengali as his first language; another reported sign language but had no hearing impairment at the time of the experiment.

5.4 Personality

In addition, I asked users to complete a personality test based on the Myers-Briggs Type Indicator ® (MBTI ®) personality inventory. Table 2 shows on the first row the number of participants whose test results categorized them in each category. The second row presents the personality average, that is, the aggregated average of all individual scores per participant.

Table 5: Summary of personality types based on the MBTI.

<table>
<thead>
<tr>
<th></th>
<th>Extrovert</th>
<th>Introvert</th>
<th>Sensing</th>
<th>Intuitive</th>
<th>Thinking</th>
<th>Feeling</th>
<th>Perceiving</th>
<th>Judging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>51.64%</td>
<td>48.36%</td>
<td>59.16%</td>
<td>40.83%</td>
<td>55.83%</td>
<td>44.17%</td>
<td>51.26%</td>
<td>48.82%</td>
</tr>
</tbody>
</table>

Although this does not play any particular role on this experiment, ongoing research at our lab suggests that personality type significantly affects users’ perception of agents. In addition, it shows that no particular personality type was favored, reducing the risk of biased results based on personality perceptions of the agent (e.g., a group full of extroverts might be more likely to gesture). The full list of personality results can be found in Appendix B.
Chapter 6: Results

So far I have developed a paralinguistic model of rapport. To improve the naturalness, effectiveness and believability of human-ECA interactions I have relied on user’s engagement in a state of rapport. To achieve this, I have designed an interaction that encourages users to express rapport-building nonverbal behaviors towards the agent. All models of rapport, including the one developed by our AGENT team, lack a definitive guide that defines precise, replicable gestures that are indicative of rapport. Furthermore, the gestures that have been identified as rapport-building behaviors were observed from human-human interactions. It is a novel approach to attempt to replicate these gestures, and questions quickly arise as to whether these gestures are appropriately represented in terms of acceleration or frequency, gender, culture, personality and situation among other parameters.

6.1 Rapport Scores Evaluation

A two-tailed t-test indicated that there was no significant main effect for physical connection (p=0.56) or understanding (p=0.23). There was, however a significant main effect for coordination (p=0.03). The survey for rapport perception included twelve questions, four for each dimension of rapport. Understandably, the physical connection questions evaluated the users’ perception of the agent’s movements. Since the animations were very similar under both conditions, no main effect was expected or desired for physical rapport.

Similarly, the sense of understanding was not significantly affected. These items evaluate the agent’s attention towards the user. However, gaze, an indicator of attention, did not vary across conditions. With the exception of the dialogs where the agent would specifically ask the user to perform a gesture by helping with some survival tasks (e.g., “Hurry, yell to the helicopter!” vs. “Hurry, signal the helicopter!”), the utterances were otherwise identical.

In contrast, the significant results for coordination suggest that feelings of trust, connection, friendliness and collaboration were positively affected by the agent that elicited rapport building behaviors.
This results consistent with the hypothesis that suggests that under the non-gesture condition, the subjects would score lower emotional connection and mutual understanding, while the sense of physical connection should not change.

Table 6: Non-gesture and gesture condition average scores across all participants. The header row initial stands for the dimension of rapport the item evaluates (Understanding, Physical, and Coordination). A positive number on the final row indicates an overall preference for the gesture agent while a negative number indicates preference for the non-gesture one.

<table>
<thead>
<tr>
<th>Item</th>
<th>Rapport Dimension</th>
<th>N</th>
<th>G</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>The agent understood me</td>
<td>Understanding</td>
<td>3.34</td>
<td>3.45</td>
<td>0.11</td>
</tr>
<tr>
<td>The agent seemed engaged</td>
<td>Understanding</td>
<td>3.43</td>
<td>3.67</td>
<td>0.24</td>
</tr>
<tr>
<td>The agent was excited</td>
<td>Physical</td>
<td>4.43</td>
<td>4.00</td>
<td>-0.43*</td>
</tr>
<tr>
<td>The agent's movements were natural</td>
<td>Physical</td>
<td>2.53</td>
<td>2.51</td>
<td>-0.02</td>
</tr>
<tr>
<td>The agent was friendly</td>
<td>Emotional</td>
<td>4.68</td>
<td>4.61</td>
<td>-0.07</td>
</tr>
<tr>
<td>The agent was paying attention to me</td>
<td>Understanding</td>
<td>3.28</td>
<td>3.29</td>
<td>0.01</td>
</tr>
<tr>
<td>The agent and I worked towards a common goal</td>
<td>Emotional</td>
<td>4.21</td>
<td>4.71</td>
<td>0.50*</td>
</tr>
<tr>
<td>The agent and I seemed to connect</td>
<td>Emotional</td>
<td>3.06</td>
<td>3.61</td>
<td>0.55*</td>
</tr>
<tr>
<td>I sensed a physical connection with the agent</td>
<td>Physical</td>
<td>2.21</td>
<td>2.38</td>
<td>0.17</td>
</tr>
<tr>
<td>The agent’s gestures were lively</td>
<td>Physical</td>
<td>3.37</td>
<td>3.29</td>
<td>-0.08</td>
</tr>
<tr>
<td>I feel the agent trusts me</td>
<td>Emotional</td>
<td>4.21</td>
<td>4.32</td>
<td>0.11</td>
</tr>
<tr>
<td>I understood the agent</td>
<td>Understanding</td>
<td>4.31</td>
<td>4.61</td>
<td>0.30*</td>
</tr>
</tbody>
</table>

According to table 6, while there was not much difference in terms of the perceived friendliness of the agent across conditions. However, users agreed that there was a positive difference in terms of sense of connection and collaboration, and to a minor extent, trust between the user and the ECA on the gesture condition.

In the next section, we analyze the gesture frequency differences across both conditions.
6.2 Gesture Evaluation

In this section I compare the nonverbal behavior of participants between the gesture and non-gesture eliciting agents.

For this analysis, I labeled the 64 participants from G01 to G32 and from N01 to N32 based on the condition and participant number. I removed the data gathered from subjects G04, G09, G20, G24, G27 and G32 from the gesture eliciting condition, and N16 and N32 from the non-gesture eliciting condition due to technical difficulties or noise in the recordings or gesture log files.

Table 7 presents the total number of gestures per participant and the cumulative total under each condition. The gesture data come from the participants’ gesture log files, which automatically recorded the most common poses or gestures. Although the videos were not completely analyzed, rated, or annotated, corroborated the soundness of the log files for this analysis - that is, the gesture recognizer was performing adequate recognition. All gestures that users performed in the gesture condition as a task requirement interaction were removed.

Overall, participants in the gesture condition seemed more likely to perform nonverbal behaviors, and were more active during the interaction. This confirms the other hypothesis, which proposed that subjects interacting physically with the agent would display more rapport-building behaviors than those who interact with the non-physical agent, and report a higher sense of rapport compared to interactions with agents displaying only verbal behaviors.

The following tables (Table 7, 8, 9, 10, 11, 12 and 13) analyze the number of gestures based on the control condition and personality type.

Table 7: Total number of gestures per participant. The full table containing gestures per scene per participant can be found in Appendix B.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G01</td>
<td>160</td>
<td>N01</td>
<td>219</td>
</tr>
<tr>
<td>G02</td>
<td>96</td>
<td>N02</td>
<td>176</td>
</tr>
<tr>
<td>G03</td>
<td>169</td>
<td>N03</td>
<td>71</td>
</tr>
<tr>
<td>G05</td>
<td>157</td>
<td>N04</td>
<td>467</td>
</tr>
<tr>
<td>G06</td>
<td>410</td>
<td>N05</td>
<td>62</td>
</tr>
<tr>
<td>G07</td>
<td>299</td>
<td>N06</td>
<td>106</td>
</tr>
<tr>
<td>G08</td>
<td>96</td>
<td>N07</td>
<td>281</td>
</tr>
<tr>
<td>G10</td>
<td>99</td>
<td>N08</td>
<td>300</td>
</tr>
<tr>
<td>G11</td>
<td>435</td>
<td>N09</td>
<td>143</td>
</tr>
</tbody>
</table>
In previous studies, such as the ones described in chapter two and three, participant’s personality played a role in the perception of the agent. Throughout those studies, agent’s gestures were varied in amplitude to suggest a familiarity feeling. A two-tailed t-test between introverts and extroverts in the gesture condition \((p=0.68)\) and another between introverts and extroverts in the non-gesture condition \((p=0.55)\) suggests that the difference between the gesture and non-gesture condition for non-task gesture production is significant and unlikely to be a byproduct or an artifact as a result of differences in perception between extraverts and introverts with respect to rates of non-task gesture production.

Table 8: T-test results for different condition comparisons.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE-GI</td>
<td>0.680001</td>
</tr>
<tr>
<td>GE-NE</td>
<td>0.148358</td>
</tr>
<tr>
<td>NE-NI</td>
<td>0.558677</td>
</tr>
<tr>
<td>GI-NI</td>
<td>0.200516</td>
</tr>
</tbody>
</table>
Table 9: Standard deviation and average gesture rate among different personalities and conditions, where G is the gesture condition, N is non-gesture condition, I stands for Introvert and E indicates an extrovert personality trait.

<table>
<thead>
<tr>
<th></th>
<th>GE</th>
<th>GI</th>
<th>NE</th>
<th>NI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVG</td>
<td>234.7</td>
<td>215.7</td>
<td>178.37</td>
<td>154.25</td>
</tr>
<tr>
<td>Std</td>
<td>116.9</td>
<td>57.9</td>
<td>92.68</td>
<td>114.82</td>
</tr>
</tbody>
</table>

Table 10: Comparison between introverts and extroverts under the gesture condition.

<table>
<thead>
<tr>
<th>Gesture E</th>
<th>Gesture I</th>
</tr>
</thead>
<tbody>
<tr>
<td>G01</td>
<td>160</td>
</tr>
<tr>
<td>G02</td>
<td>96</td>
</tr>
<tr>
<td>G06</td>
<td>410</td>
</tr>
<tr>
<td>G07</td>
<td>299</td>
</tr>
<tr>
<td>G08</td>
<td>96</td>
</tr>
<tr>
<td>G10</td>
<td>99</td>
</tr>
<tr>
<td>G11</td>
<td>435</td>
</tr>
<tr>
<td>G12</td>
<td>296</td>
</tr>
<tr>
<td>G16</td>
<td>219</td>
</tr>
<tr>
<td>G18</td>
<td>158</td>
</tr>
<tr>
<td>G19</td>
<td>369</td>
</tr>
<tr>
<td>G21</td>
<td>365</td>
</tr>
<tr>
<td>G22</td>
<td>247</td>
</tr>
<tr>
<td>G23</td>
<td>59</td>
</tr>
<tr>
<td>G28</td>
<td>334</td>
</tr>
<tr>
<td>G30</td>
<td>205</td>
</tr>
<tr>
<td>G31</td>
<td>144</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>1726</strong></td>
</tr>
</tbody>
</table>

Table 11: Comparison between introverts and extroverts in the non-gesture condition

<table>
<thead>
<tr>
<th>Non-Gesture E</th>
<th>Non-Gesture I</th>
</tr>
</thead>
<tbody>
<tr>
<td>N01</td>
<td>219</td>
</tr>
<tr>
<td>N02</td>
<td>176</td>
</tr>
<tr>
<td>N03</td>
<td>71</td>
</tr>
<tr>
<td>N05</td>
<td>62</td>
</tr>
<tr>
<td>N06</td>
<td>106</td>
</tr>
<tr>
<td>N08</td>
<td>300</td>
</tr>
<tr>
<td>N13</td>
<td>108</td>
</tr>
<tr>
<td>N15</td>
<td>387</td>
</tr>
<tr>
<td>N17</td>
<td>127</td>
</tr>
<tr>
<td>N18</td>
<td>223</td>
</tr>
<tr>
<td>N04</td>
<td>467</td>
</tr>
<tr>
<td>N07</td>
<td>281</td>
</tr>
<tr>
<td>N09</td>
<td>143</td>
</tr>
<tr>
<td>N10</td>
<td>41</td>
</tr>
<tr>
<td>N11</td>
<td>148</td>
</tr>
<tr>
<td>N12</td>
<td>121</td>
</tr>
<tr>
<td>N14</td>
<td>190</td>
</tr>
<tr>
<td>N19</td>
<td>149</td>
</tr>
<tr>
<td>N24</td>
<td>42</td>
</tr>
<tr>
<td>N25</td>
<td>44</td>
</tr>
</tbody>
</table>

80
Table 12: Number of participants under each personality type, under both conditions and their cumulative gestures

<table>
<thead>
<tr>
<th>NONGESTURE</th>
<th>GESTURE</th>
<th>GESTURE</th>
<th>NONGESTURE</th>
<th>GESTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>16</td>
<td>2831</td>
<td>E</td>
<td>17</td>
</tr>
<tr>
<td>I</td>
<td>12</td>
<td>1824</td>
<td>I</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>4655</td>
<td>Total</td>
<td>25</td>
</tr>
<tr>
<td>N</td>
<td>7</td>
<td>1320</td>
<td>N</td>
<td>3</td>
</tr>
<tr>
<td>S</td>
<td>19</td>
<td>2742</td>
<td>S</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>4062</td>
<td>Total</td>
<td>25</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>1579</td>
<td>F</td>
<td>7</td>
</tr>
<tr>
<td>T</td>
<td>16</td>
<td>2451</td>
<td>T</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>4030</td>
<td>Total</td>
<td>22</td>
</tr>
<tr>
<td>J</td>
<td>13</td>
<td>1607</td>
<td>J</td>
<td>11</td>
</tr>
<tr>
<td>P</td>
<td>13</td>
<td>2653</td>
<td>P</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>4260</td>
<td>Total</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 13: Normalized average gestures in gestures per across all personality types for the gesture and non-gesture conditions.

<table>
<thead>
<tr>
<th>NORMALIZED</th>
<th>N</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>177</td>
<td>235</td>
</tr>
<tr>
<td>I</td>
<td>152</td>
<td>216</td>
</tr>
<tr>
<td>N</td>
<td>189</td>
<td>199</td>
</tr>
<tr>
<td>S</td>
<td>144</td>
<td>233</td>
</tr>
<tr>
<td>F</td>
<td>158</td>
<td>242</td>
</tr>
<tr>
<td>T</td>
<td>153</td>
<td>205</td>
</tr>
<tr>
<td>J</td>
<td>124</td>
<td>235</td>
</tr>
<tr>
<td>P</td>
<td>204</td>
<td>222</td>
</tr>
</tbody>
</table>
Furthermore, a two-way ANOVA indicates that there is not a significant difference among the conditions when paired with personality (gesture-extrovert, non-gesture-extrovert, gesture-introvert, non-gesture-introvert), with the possible exception of gesture-introvert (Figure 14). This analysis was performed to further examine the effects of personality in our rapport-building agent, however, further analysis, such as that for the other personality variants falls out of the scope of this research.

One of the reasons the gesture-introvert condition might have come off as significant could be related to introverts finding gesturing uncomfortable in general, although it is more likely that these results are an artifact of our limited data for this analysis.

![Figure 14: Gesture rates across conditions, where A represents gesture-extrovert, B non-gesture-extrovert, C non-gesture-introvert, D gesture-introvert.](image)

Some of these non-task gestures made the human-ECA connection stronger. An interesting example were the participants who danced. At the end of the interaction, while the credits rolled, the agent would start dancing to the credits music. Many users did attempt to dance with the agent. In some
cases, they tried to mimic or imitate the dance steps, while in others they offered dance lessons and performed their own dance moves. Dancing with or imitating the agent are clear examples of non-verbal rapport-building behaviors.

In another example, a user, which was not uniquely in this situation, offered to help the agent in various non-task scenes. He accompanied his gestures with verbal help offerings and a concerned tone of voice for the agent’s wellbeing. This particular user I am referring to, stretched his hand towards the agent in soothing motions to clearly calm her pain, as this occurs in a scene where the ECA is wounded and in distress. Other participants in the same situation would be more aggressive in both, tone of voice and gestures, while still caring for the agent. For example, another participant in the same scene pointed out commandingly at the ECA and said loudly “do not go to sleep”, in a scene where the agent is about to faint. This rough act of caring shows that rapport is being maintained and increased, as this is something that might be unacceptable during the initial phases of a relationship, but becomes understandable and tolerable as the relationship progresses and positivity becomes less important.
Chapter 7: Conclusion

7.1 Contributions

One of the questions to address in ECA research is how agents should appropriately represent human behaviors to create and maintain states of rapport with humans. In human-human conversations, this is can be a seamless unconscious process that occurs naturally by interacting with another human being. In most cases however, it is often not so easy for people to build rapport; there are also many cases where people fail to connect, or to fail to connect as well as they might have. To build rapport, humans display behavioral cues which include gesture mimicry, gaze, head nodding, body orientation, and body posture that indicate their synchronicity with another person. These cues are regularly misinterpreted, unobserved, or inappropriate in time or context, which is why modeling and replicating rapport in human-agent interactions requires careful planning and rarely leads to optimal results.

A line of research tries to build agents that express rapport through their nonverbal behaviors. Based on molar (high-level, abstract) or molecular (fine-grained, overdetailed) evaluation methods that are not inclusive of all gestures and tend to generalize or over specify gesture qualities, these models succeed in replicating some behaviors imitated from social cues in human-human interaction. In this dissertation I have expanded this line of research by creating a rapport-eliciting agent - that is, an agent that requests and expects rapport-building gestures based on user engagement and non-task user gestures. This work assumed that gesture frequency is as important and indicative as gesture quality. The frequency was measured by comparing two physically identical agents interacting with subjects under a gesture-eliciting condition and a non-gesture condition. I analyzed users’ gestures log files produced by our full-body tracking system to gain insight into the effectiveness and importance of rapport eliciting ECAs.

7.1.1 Physical Engagement as a Way to Increase Rapport

The main contributions to the field are twofold. First, eliciting gestures improves the perception of emotional rapport in human-ECA interactions and increases the sense of connection, collaboration and trust. Second, gesture eliciting increases engagement (gesture frequency) of non-task gestures in users. These suggest that agents and interactions should be developed with physical engagement in mind.
to more fully exploit the benefits of rapport. These results, combined with the tools we have developed can lead to more engaging, socially present agents and new ways to interact with virtual characters.

Interaction analysis showed the importance of gesture elicitation. Technical limitations constrained this awareness to a subset of nonverbal behaviors related to pose changes, but they nonetheless proved to be an effective starting point for analyzing users’ rapport gestures. I also provided an overview of the different components required to build an ideal agent and explained the modules that enable speech recognition, gesture recognition, and the game logic to function as a single entity.

7.1.2 Multimodal Information in Human-ECA Interactions

To react to as many gestures as possible, I presented several modules from the AGENT Framework. For gesture, I discussed modules for creating pose libraries, analyzing gesture in real-time throughout the interaction, and logging the results for all agent and human gestures throughout the interaction. Furthermore, video recordings were captured for all participants, and a personality analysis provide additional perspectives that can be used in conjunction with the gesture log to explain human-ECA behaviors.

7.1.3 Behavioral Analysis of Virtual Rapport

Finally, this research presented an extensive analysis of current knowledge on rapport. I combined temporal models, with relational models, and connected them to the traditional definition of rapport. This enabled me to identify and target what is arguably the most difficult composite dimension of rapport: coordination. By establishing a model that requires a synergistic human-ECA relationship, I ensure that rapport is effectively a state and not a feeling or a one-sided representation of an ideal behavior. In doing so, I enabled the agents to engage in rapport by using a constrained interaction situated in the context of a jungle survival scenario as a base for developing intimacy and affinity.

I also measured virtual rapport across the eliciting and non-eliciting agents. I showed that agents that create or engage in situations that require human nonverbal behavior to complete are more relatable, and likeable, and thus users interact to a greater extent than with agents that are not aware and do not elicit these behaviors.
7.2 Limitations and Future work

In my study, all human-ECA conversations took place with a physically identical agent. Due to the immersive nature of the interaction, some of the video recordings were obscured and unusable. Our agents and gesture annotation modules were fully automated. Under exceptional circumstances (i.e., computer system automatic update, low network speeds or computer processor saturation), the game or the agent malfunctioned; most often this was visible in the form of a half-second lag between recognition and response, or agent spoken dialog and nonverbal expression. In some cases this was perceived by the user as an intended interaction or behavior; I removed all known faulty interactions. The rapport-eliciting model developed in this dissertation analyses the gesture frequency only through the interaction and the users’ perception of rapport; it does not presume to express more than the appropriate categorical classification of rapport-building gestures.

Other aspects of the model can be improved upon. The agent walks, effectively translating in virtual space. These translations often result in the ECA looking away from the user. Gaze was not a factor in rapport-building or rapport-eliciting behavior in the application used in this study, although it is known to be an important indicator of attention, which is a basic dimension of rapport.

Aside from the relatively modest technical annoyances, this work makes way for interesting research challenges. It has always been relatively difficult for humans to elicit rapport in everyday interactions; sometimes it is not required by the situation or context, or there is not enough time for people to converge in a state of rapport. As with any social trait that is not clearly defined, generalized, and modeled, it is quite difficult to translate and design them within an ECA system. Future work should address current behavioral limitations for everyday rapport, such as how can we build intimacy by referring to prior interactions in long-term settings. In particular, it is challenging to decide what information to store, and more importantly, when to retrieve it for the desired intimacy effect. It is also difficult to integrate emotions into rapport. Most of the emotions generated in our system were situational, designed to be triggered when specific events occurred. As a follow up, creating emotional agents capable of recognizing, analyzing and reacting to emotion would be the first step towards emotional computing in automated human-ECA interactions. This area has been partially explored by
using emotional coloring (Acosta and Ward, 2011) or skin color features (Ramirez, et al., 2014), but it has yet to be applied to emotional models and virtual agents.

It is also difficult to design immersive experiences for rapport. Current approaches, like the one presented in this research, immerse the user in the world of the agent, that is, the user plays by the agent’s rules and manipulated the course of action in a virtual environment rather than in our own physical space. Creating immersive experiences where users interact with agents in the physical space, or where the users exist virtually in the same plane (perhaps using emerging virtual reality technologies) could allow a more direct paralinguistic approach at generating rapport and may lead to adjustments in the currently limited models of virtual rapport.

Another research challenge that can now be addressed thanks to the development of our fully autonomous agents is rapport over time. It is well known that rapport can be build, maintained, and destroyed, however, it has not been possible to explore this in human-agent settings yet because it requires extensive interaction sessions, which are unfeasible for current WoZ systems. With our system we can now explore adaptive rapport, or how rapport changes over time in order to increase or maintain it in human-ECA interactions.

Ultimately, rapport is a combination of behaviors and perception, and as such, it is difficult to isolate. The model proposed in this work should be integrated with respective verbal models of rapport and nonverbal gestures that address other conversational characteristics out of the scope of rapport-elicitation. In particular, the quality of rapport-eliciting behaviors should be analyzed and factored in measuring the success of rapport-building agents at a gesture level in addition to a perceptual one.

7.2.1 Attitude Gestures in Virtual Rapport

Not all gestures are equal. Nonverbal expression is a complicated behavior that depends on an extensive list of factors that modify the intention or the expression of the gesture. One notable gesture modifier is attitude. Gestures can be weighted based on intimacy, so that a gesture that may appear rude in private might seem acceptable in public. This perception is not based on the physical quality of the gesture itself but rather its context and the setting. Attitude is similar: identical gestures can be interpreted according to the perceived mood or real attitude of the gesturer. Although attitude
recognition in human-computer interaction is loosely based on facial recognition, inspiring attitude analysis for gesture interpretation may eventually lead to increased accuracy and improved gesture adaptation in human-ECA interactions.

7.2.2 Shared Interpersonal Knowledge in Virtual Rapport

Adriana, the ECA extensively discussed in this work, can perform a wide variety of automated functions that help simulate natural and believable behavior. Many of these functions are based on a predefined context. The agent is only aware of this context because it is carefully crafted and programmed into her. Shared interpersonal knowledge in this experiment was static—that is, the agent always revealed the same information, at the same point in time for every participant. Players would also be requested to reveal information during specific stages of the interaction, and although they were not limited or constrained from revealing information at other points during the interaction, the information would only be appropriately processed if provided or requested during the predefined allotted slots.

Creating a more generic method for gaining, remembering, using, and re-using interpersonal knowledge would produce a major improvement on rapport-building ECA design and development.

7.2.3 Personality as an Indicator of Rapport

Previous studies with earlier versions of our research team’s ECA in a more constrained interaction created awareness of gesture perception based on personality types. If personality types can be connected with gesture perceptions, then develop agents may be able to adapt their gesture amplitude, speed, or frequency to match those that a user finds most natural and comfortable. This study collected such information but did not analyzed these data, as this analysis was beyond the scope of the research presented here. In an ideal setting, adaptive personality ECAs could perhaps lead researchers toward creating agents with better social behavioral skills than humans, as they would adapt to a person’s preference, something we humans are not particularly skilled at.
Bibliography


Gris, Iván ; Novick, David ; Rivera, Diego A. ; Gutiérrez, Mario (2014) UTEP's AGENT Architecture. Workshop on Architectures and Standards for IVAs, held at the "14th International Conference on Intelligent Virtual Agents (IVA 2014)", Boston, USA, August 26

Gris, Ivan, Adriana Camacho, David Novick “Full-Body Gesture Recognition for Embodied Conversational Agents” In Press. GESPIN 2015


Hudson Casey “Mass Effect 3”, 2012.


Molyneux Peter “Fable”, 2004.


Willis, Janine, and Alexander Todorov. "First impressions making up your mind after a 100-ms exposure to a face." Psychological science 17, no. 7 (2006): 592-598.


Appendix A: Scene Scripts

This appendix contains all scripts for the gesture condition. The dialog is highly redundant in the non-gesture version, with a consistent difference. All decide gesture tags are changed to decide speech in the non-verbal condition. By the time of this publication, minor edits and corrections might have occurred on this script. Also, note that the dialog is not linear and it is not necessarily read from top to bottom. For additional information on how to interpret the scripts read chapter 5.

<scene "Scene_1_Beach">
<episode "beach1">
    <do emotion "Surprised" hands "Idle" Locomotion "Crouch" upperBody "Scared" isSpecial "false" special "none">
        <speak "beach1.wav" "Hey wake up, can you hear me? Hello? Hellooo?">
            <pause "1">
                <decide speech>
                    <createRule "wildcard">
                        <go "beach1b"> </go>
                        <tag "wc">
                            <items "wildcard">
                            </tag>
                    </rule>
                </decide>
            </tag>
        </speak>
    </episode>
<episode "beach1b">
    <speak "beach2.wav" "Are you okay? Man glad to see you’re alive!">
        <pause "1">
            <speak "beach3.wav" "How are you feeling? Any broken bones or anything?">
                <pause "1">
                    <decide speech>
                        <createRule "wildcard">
                            <go "beach1c"> </go>
                            <tag "wc">
                                <items "wildcard">
                            </tag>
                        </rule>
                    </decide>
                </tag>
            </speak>
        </speak>
    </episode>
<episode "beach1c">
    <speak "beach4.wav" "You look all right to me. Are you alone?">
        <pause "1">
            <decide speech>
                <createRule "wildcard">
                    <go "beach1e"> </go>
                    <tag "we">
                        <items "wildcard">
                    </tag>
                </rule>
            </decide>
        </tag>
    </episode>
<episode "beach1e">
    <speak "beach5.wav" "I’m sorry I’ll let you catch your breath. Do you think there are any other survivors?">
        <pause "1">
            <decide speech>
                <createRule "wildcard">
                    <go "beach2"> </go>
                    <tag "we">
                        <items "wildcard">
                    </tag>
                </rule>
            </decide>
        </tag>
    </episode>
</scene>
<episode "beach2">
    <do emotion "Neutral" hands "Idle" Locomotion "Crouch" upperBody "Scared" isSpecial "false" special "none">
        <speak "beach6.wav" "I don't see anyone else around here.">
        <pause "1">
            <speak "beach7.wav" "Can you walk?">
                <pause "1">
                    <decide speech>
                        <createRule "wildcard">
                            <go "beach2b"> </go>
                        </tag "wc">
                    </decide>
                </rule>
            </speak>
        </speak>
    </do>
</episode>

<episode "beach2b">
    <speak "beach8.wav" "Well, you'd better gather up your strength anyway because I won't be able to carry you around">
        <pause "1">
            <go "beach2c"></go>
        </speak>
</episode>

<episode "beach2c">
    <do emotion "Caring" hands "Idle" Locomotion "StandUp" upperBody "Pointing" isSpecial "false" special "none">
        <speak "beach9.wav" "Here, let me help you up. Grab my hand.">
            <pause "1">
                <decide gesture>
                    <go "beach3"> "Lift Hand" </go>
                </decide>
            </speak>
        </speak>
    </do>
</episode>

<episode "beach3">
    <do emotion "Sad" hands "Idle" Locomotion "Walking" upperBody "Thinking" isSpecial "false" special "none">
        <speak "beach10.wav" "It looks like we are trapped here.">
            <pause "1">
                <speak "beach11.wav" "It might take us a while to get off of this place, in the meantime, we might as well get used to it.">
                    <pause "1">
                        <speak "beach12.wav" "I think we have a greater chance of survival if we stay together.">
                            <pause "1">
                                <speak "beach13.wav" "Oh, I'm sorry where are my manners, my name is Adriana, it's nice to meet you, although I wish we could have met under better circumstances.">
                                    <pause "1">
                                        <speak "beach14.wav" "How about you, what's your name?">
                                            <pause "1">
                                                <decide speech>
                                                    <createRule "wildcard">
                                                        <go "beach3a"> </go>
                                                    </tag "wc">
                                                </decide>
                                            </speak>
                                        </speak>
                                    </speak>
                                </speak>
                            </speak>
                        </speak>
                    </speak>
                </speak>
            </speak>
        </speak>
    </do>
</episode>

<episode "beach3a">
    <speak "beach15.wav" "Nice to meet you">
        <pause "1">
            <go "beach4"></go>
        </speak>
</episode>

<episode "beach4">
    <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Explaining" isSpecial "false" special "none">
        <speak "beach16.wav" "Well I've been on this island for a few days now.">
            <pause "1">
                <speak "beach17.wav" "It's a nice place, but tougher than it looks. I've barely managed to stay alive.">
                    <pause "1">
                        <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Explaining" isSpecial "false" special "none">
Oh man, it looks like the storm that brought you here is still coming after you.

C'mon let's go to my shelter, we can talk there.

The storms out here can get pretty rough, let's hope this one doesn't hit too hard.

This jungle is so calm and peaceful, it's hard to believe it can have such bad weather.

We are almost there, we just need to cross the pond now.

Careful here, it's not a big fall from the tree but try to keep your balance anyway. Look, use your arms, like this.

Well, this is it, home sweet home. Not what you expected I suppose.

It's not an ideal place to live, but it's all I could make on my own.

Do you think it will stand the storm if it hits, or should we try to find another place?

Well, actually, maybe staying here isn't such a good idea after all. The nights here are cold and it looks like the wind will blow this in no time. Let's try to find higher ground and solid walls.

Good idea. Let's try to find higher ground and solid walls. We may be able to find a cave up in that mountain over there. Let's move.

Do you think it will stand the storm if it hits, or should we try to find another place?
"I didn’t think it was such a good idea to move away from the beach in case some boat or plane happened to pass by."

"but I don’t think staying here is an option any more."

"Oh, I almost forgot my stuff. It might look like trash but who knows when it can turn out to be useful."

"It’s not that heavy but we might have to cover a big distance with these things."

"All right, all set, let’s get moving."

"You don’t have to be so quiet. I’m sure we will be just fine. We are alive at least."

"It gets pretty quiet when you have spent a few days alone."

"So tell me, where are you from?"

"Ah, so Texas huh?"

"Im from Texas too! It’s very hot there most of the year. At least it’s not as humid as this place. I hate mosquitoes!"

"Im from a small town" "Im from out there" "las cruces" "new mexico" "canada" "albuquerque" "england" "america" "india" "spain" "barcelona" "bangladesh"
"Ah, Mexico huh?"

"Oh, you must be a Spanish speaker then. I've heard great things about Mexico, never been there though and my Spanish sucks, so keep it in English please."

"Cool, I've never been there. At least it's not as humid as this place I suppose. I hate mosquitoes!"

"I'm just a town girl from Texas."

"How long have you lived there?"

"That must be nice. I've lived alone for a while."

"I don’t mean here. But hey, let's cheer up. I'm sure you'll see them again."
<episode "splitpaths4">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Explaining" isSpecial "false" special "none">
    <pause "2">
      <speak "splitpaths15.wav" "Have you ever been in the wild? I mean, camping or that type of thing?">
        <decide speech>
          <createRule "camping">
            <go "splitpaths4a" />
            <tag "campyes">
              <items "yes" "yeah" "when i was a kid" "long ago" "every year" "i go often" "a few times" "once" "a couple of times" "not in a tent but in a cabin" "i would go hunting" "yes i love the outdoors" "i always go camping" "i am an outdoors person" "i have been camping but not in a place like this" "i have been to the woods" "once or twice" "my family and I would go on camping trips" "i have gone fishing" />
            </tag>
            <go "splitpaths4b" />
            <tag "campno">
              <items "no" "nope" "never been" "no but I always wanted to" "i hate camping" "i am scared of camping" "i dont like the outdoors" "im from the city" "i avoid nature" "i am not really into this" />
            </tag>
          </rule>
        </decide>
      </speak>
    </do>
  </episode>
</episode>

<episode "splitpaths4a">
  <do emotion "Caring" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <speak "splitpaths16.wav" "Well, that's good. Your skills will be helpful later on.">
      <pause "1">
        <go "splitpaths5" />
      </speak>
    </do>
  </episode>
</episode>

<episode "splitpaths4b">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <speak "splitpaths17.wav" "Well I have. I'm no expert, but together we might have a shot.">
      <pause "1">
        <go "splitpaths5" />
      </speak>
    </do>
  </episode>
</episode>

<episode "splitpaths5">
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Thinking" isSpecial "false" special "none">
    <speak "splitpaths18.wav" "So, where to now? Should we take the grassy path which leads to the jungle or take the rocky path which leads to the mountains?"
      <pause "1">
        <decide speech>
          <createRule "paths">
            <go "grassy" />
            <tag "grass">
              <items "grassy path" "grassy" "lets take the grassy path" "lets take the path no one has been to" "the grassy path" "the jungle" "the grassy path, i am afraid of the mountains" "the grassy path is safer" "the safest path" "the left one" "to the left" "the one on the left" />
            </tag>
            <go "rocky" />
            <tag "rock">
              <items "lets take the rocky path" "rocky path" "rocky" "mountains" "the mountain path" "the rocky road" "the right one" "the one on the right" "lets seek altitude" "lets go to the mountains" "we should take the rocky path" "the rocky path to the mountain" />
            </tag>
          </rule>
        </decide>
      </speak>
    </do>
  </episode>
</episode>

<episode "grassy">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Scared" isSpecial "false" special "none">
    <speak "grassy1.wav" "Are you sure about that? Who knows what we will find if we go deeper into the jungle.">
      <pause "1">
        <speak "grassy2.wav" "Not that I'm afraid or anything but I'm not to sure about finding shelter in a humid, insect-infested low ground place.">
          <pause "1">
            <speak "grassy3.wav" "We might be able to find a cave if we go on the rocky path. Lets explore the jungle once we have shelter."
              <pause "1">
                <speak "grassy4.wav" "I'll lead the way."
                  <pause "1">
                    <do emotion "Caring" hands "Idle" Locomotion "Walking" upperBody "Excited" isSpecial "false" special "none">
                      <pause "2">
                      </do>
                    </speak>
                  </pause>
                </speak>
              </pause>
            </speak>
          </pause>
        </speak>
      </pause>
    </speak>
  </do>
</episode>
<episode "rocky">
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Excited" isSpecial "false" special "none">
    <pause "2"/>
    <speak "rocky1.wav" "Hmm, sounds good to me! I don't want a bunch of insects crawling all over while I sleep."
      <pause "1"/>
    <speak "rocky2.wav" "Hey, maybe we can even find a cave or something! As long as it's dry and on higher ground we should be fine. I haven't seen any large animals around here, so I don't think we'll have to worry about that either."
      <pause "1"/>
    <speak "rocky3.wav" "I'll lead the way."/>
    <pause "2"/>
  </do>
</episode>

<nextscene "Scene_3_Mountain">
</scene>

<scene "Scene_3_Mountain">
<episode "mountain1">
  <do emotion "Excited" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none">
    <speak "mountain1.wav" "I haven't explored this far up the mountain before."
      <pause "1"/>
    <speak "mountain2.wav" "I'm starting to get kind of tired, I hope we can find some shelter soon."
      <pause "5"/>
  </do>
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Pointing" isSpecial "false" special "none">
    <speak "mountain3.wav" "Well speak of the devil, look at that!"
      <pause "3"/>
    <speak "mountain4.wav" "I think that cave is as good as it gets. Let's go check it out."
      <pause "6"/>
  </do>
</episode>

<nextscene "Scene_3A_EnterCave">
</scene>

<scene "Scene_3A_EnterCave">
<episode "entercave1">
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none">
    <speak "entercave1.wav" "This looks like a good place to stay; Now we just need some water and food, and I wouldn't mind if we had fire too."
      <pause "1"/>
    <speak "entercave2.wav" "Let's not get too comfortable though, I still have plans to get off this island."
      <pause "1"/>
  </do>
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Listening" isSpecial "false" special "none">
    <speak "entercave3.wav" "Wait, I hear something!"
      <pause "1"/>
    <speak "entercave4.wav" "I think that's water, but it's pretty dark, I can't see a thing. What do you think?"
      <pause "1"/>
    <speak "entercave5.wav" "Should we, go in or search for water somewhere else? I'm not going alone in there!!"
      <pause "1"/>
  </do>
  <decide speech="true">
    <createRule "searchwater1">
      <go "entercave1a">
        <tag "go">
          <items "lets go" "lets look for water" "ill go" "ill get the water" "i guess ill go" "lets go look for water" "im too thirsty ill do anything" "im thirsty i want water" "yeah im thirsty" "lets go in the cave im thirsty" "Yeah lets go" "Let's go together" "Yes" "Yeah!" "Sounds like fun"/>
        </go>
      </tag>
      <go "entercave1b">
        <tag "none">
          <items "i dont think so" "look for water somewhere else" "lets look somewhere else" "im not going unless you go" "im not going alone either" "im scared" "i think we passed some fresh water somewhere else" "lets just go somewhere else" "its too dark in there" "lets keep following the path and see what else we can find" "im scared of the dark" "i would go in there either its dangerous" "search for water somewhere else" "no i dont think its a good idea to go inside" "lets go find water somewhere else" "if you dont want to go in the cave we can search somewhere else" "lets go somewhere else" "No way" "Go somewhere else" "Let's get out of here" "Look elsewhere" "No" "None" "Nope">
        </go>
      </tag>
    </rule>
  </decide>
</episode>
<episode "entercave1a">
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none">
    <pause "1"/>
    <speak "entercave6.wav" "Really? Well, after you Indiana Jones, hope you are not afraid of snakes and spiders."/>
  </do>
  <pause "1"/>
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none">
    <pause "1"/>
    <speak "entercave7.wav" "Now that I think about it, even if you are afraid of them I don’t think we will be able to see them in this darkness."/>
  </do>
  <pause "2"/>
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none">
    <speak "entercave8.wav" "Look water! Do you think we should drink that?"/>
    <pause "1"/>
  </do>
  <decide speech>
    <createRule "searchwater2">
      <go "entercave2" />
      <tag "gowater">
        <items "i think so" "lets drink" "you drink first" "yes definitely" "definitely" "probably" "maybe if we boil it" "yes it looks clean" "bottoms up" "yes" "yeah" "Let’s drink it" "Sure" "Take it">
          <tag/>
        </tag>
      </go>
      <tag "nopewater">
        <items "i dont think so" "the water will make us sick" "the water looks funny" "probably not" "it depends on the color" "it might be poisonous" "it might be dirty" "no lets boil it first" "not before filtering it" "no i dont think we should drink water in a cave" "lets check how clear it is first" "lets boil it to make sure" "no" "nope" "Don’t drink it" "No way">
          <tag/>
        </tag>
      </go>
    </rule>
  </decide>
  </episode>

<episode "entercave1b">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <pause "2"/>
    <speak "entercave9.wav" "Yeah, let’s look for water elsewhere, I heard some water running along the way, might be a waterfall or a river perhaps."/>
    <pause "1"/>
    <speak "entercave10.wav" "That water should be safer to drink, it might even taste better let’s go look for it."/>
    <pause "1"/>
    <go "Scene_3A2_LookforWater" />
  </do>
</episode>

<episode "entercave2">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <pause "2"/>
    <speak "entercave11.wav" "Wait! I don’t think we should. This is pretty embarrassing but I once got diarrhea from drinking water from my dog’s bowl when I was a kid and I learned my lesson."/>
    <pause "1"/>
    <speak "entercave12.wav" "I think we should at least try to boil the water first, if it’s not running water it may have parasites or something."/>
    <pause "1"/>
    <go "entercave4" />
  </do>
</episode>

<episode "entercave3">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <pause "2"/>
    <speak "entercave13.wav" "Yeah, you are probably right. This is pretty embarrassing but I remember, when I was a kid, I got diarrhea from drinking water from my dog’s bowl, I think I’ve learned my lesson">
      <pause "1"/>
      <speak "entercave14.wav" "We can boil it later or just look for another water source, what do you think?"/>
      <pause "1"/>
      <decide speech>
        <createRule "boil">
          <go "entercave4" />
          <tag "goboil">
            <items "lets boil it" "lets just get this over with and boil it" "theres no time for looking around" "we should boil it" "we should try boiling it" "you boil it" "ill boil it" "we should boil it" "i would suggest boiling it" "boil" "boil the water" "id boil it if its safer" "lets boil it because we may not find water somewhere else" "yes" "yeah" "boil it">
              <tag/>
            </tag>
          </go>
          <tag "nopeboil">
            <items "boiling is too much work" "lets go find water thats safer" "lets go somewhere else" "I would prefer from a fresher water source" "no" "nope" "look for another source" "look somewhere else">
                <tag/>
              </tag>
            </tag>
          </go>
        </rule>
      </decide>
    </speech>
  </do>
</episode>
<episode "entercave3b">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <speak "entercave15.wav" "Yeah, let’s look for water elsewhere, I heard some water running along the way, might be a waterfall or a river perhaps."></speak>
    <pause "1"></pause>
    <speak "entercave16.wav" "That water should be safer to drink, it might even taste better let’s go look for it."></speak>
    <pause "1"></pause>
    <go "Scene_3A2_LookforWater"> </go>
  </do>
</episode>

<episode "entercave4">
  <do emotion "Caring" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <speak "entercave17.wav" "Great! Now all we need is to build a fire."></speak>
    <pause "1"></pause>
</episode>

<nextscene "Scene_4_BuildFire">
</scene>

<scene "Scene_3B_RiverWater">
  <episode "riverwater1">
    <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Excitement" isSpecial "false" special "none">
      <speak "riverwater1.wav" "I think we are getting closer">
      <pause "2"></pause>
      <speak "riverwater2.wav" "Wow this place is full of fresh water! Let's freshen up!">
      <pause "1"></pause>
      <speak "riverwater3.wav" "I saw that guy in Man versus Wild gather up rain water from big leaves… now that I think about it, I saw him drinking urine too.">
      <pause "2"></pause>
      <do emotion "Angry" hands "Idle" Locomotion "Walking" upperBody "HeadShake" isSpecial "false" special "none">
        <speak "riverwater4.wav" "We are lucky we have all this water at our disposal!">
        <pause "1"></pause>
      </do>
    </do>
  </episode>
  <nextscene "Scene_4_BuildFire">
</scene>

<scene "Scene_3C_JungleMistake">
  <episode "junglemistake1">
    <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Thinking" isSpecial "false" special "none">
      <speak "junglemistake1.wav" "Wood. A lot of wood! Hey don’t look at me like that; do you know how hard it is to find DRY wood in this place? We can use it to build a fire."/>
      <pause "1"></pause>
      <speak "junglemistake2.wav" "First things first though, we need water and we seem to be getting further away from it, I can’t hear that sound of running water anymore."></speak>
      <pause "1"></pause>
      <speak "junglemistake3.wav" "Let’s go back and look for that river, if there is one I mean. "></speak>
      <pause "2"></pause>
</episode>

<nextscene "Scene_3B_RiverWater">
</scene>

<scene "Scene_4_BuildFire">
</scene>
<episode "buildfire1">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Thinking" isSpecial "false" special "none">
    <speak "buildfire1.wav" "Now that we've had enough to drink, we might want to find a way to stay warm during the night">
    <pause "1">
    <speak "buildfire2.wav" "So, I guess that if we want to build a fire, we need to do it like they do it in the movies. Have you seen how they do it?">
    <pause "1">
    <decide speech>
      <createRule "firebuild">
        <go "buildfire1a"/>
        <tag "yepfire">
          <items "you mean with a rock and a stick" "i know how to, i was a boy scout" "yes ive even tried before" "with a rock" "of course i have" "duh i dont live under a rock"
            "like the cowboys do it" "i was a boy scout" "i was a girl scout" "im a scout leader"
            "why yes i have" "sure rub two sticks together" "get gasoline and pour it on top of stuff and light it with a match"
            "yeah i saw castaway" "yeah i think we need to find some rocks and make friction" "yeah we need to find flint and tinder"
            "yes" "yeah" "Of course" "Sure" "I have">
        </tag>
      </go "buildfire1b"/>
      <tag "nahfire">
        <items "no but we wont know unless we try" "no why dont we give it a shot" "i have no idea" "i dont think that will ever work" "i dont watch tv" "ive never seen a movie so i dont know how to" "not really no" "i dont know how to make a fire" "no" "nope" "I dont think so" "No way" "never have">
      </tag>
    </rule>
  </decide>
</episode>

<episode "buildfire1a">
  <pause "1">
  <speak "buildfire3.wav" "Quite a cliche but I think it should work.">
  <pause "1">
  <speak "buildfire4.wav" "Lets gather as many dry materials as we can to use as tinder. Of course you already knew that, you've seen it.">
  <pause "1">
  <speak "buildfire5.wav" "Lets head back out to look for the stuff we need.">
  <pause "1">
</episode>

<episode "buildfire1b">
  <pause "2">
  <speak "buildfire6.wav" "Quite a cliché but I think it should work.">
  <pause "1">
  <speak "buildfire7.wav" "We need to find something we can use as tinder, something that will ignite very easily.">
  <pause "1">
  <speak "buildfire8.wav" "Theres a lot of stuff around here, but most of it is either wet, or we may need to grind it a little. Dry grass or tree bark should be fine.">
  <pause "1">
  <speak "buildfire9.wav" "Oh, we also need kindling. Then we can build a tepee and it should burn inside out.">
  <pause "1">
  <speak "buildfire10.wav" "Lets head back out to look for the stuff we need.">
  <pause "2">
</episode>

<nextscene "Scene_4A_GatherWood">
</nextscene>

<scene "Scene_4A_GatherWood">
</scene>

<episode "gatherwood1">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Excited" isSpecial "false" special "none">
    <speak "gatherwood1.wav" "Oh yeah! It looks like we hit the jackpot! We should be able to find everything we need here.">
    <pause "3">
    <speak "gatherwood2.wav" "The sun seems to be going down pretty quick though so we better hurry!">
    <pause "1">
    <speak "gatherwood3.wav" "Lets gather as much as we can.">
    <do emotion "Happy" hands "Idle" Locomotion "Default" upperBody "Throwing" isSpecial "false" special "none">
      <speak "gatherwood4.wav" "Here! Catch these!">
      <pause "2">
      <speak "gatherwood5.wav" "Its dry animal poo. We can use that tinder! Just kidding, havent seen any animals here now that I think about it">
      <pause "2">
      <speak "gatherwood6.wav" "Its dry animal poo. We can use that tinder! Just kidding, havent seen any animals here now that I think about it">
      <speak "gatherwood7.wav" "No way! Never have we seen anything like this before."
        "Surely they are waiting for someone like us to come along and find their treasure."
        "So, we can use that to our advantage."
        "Great!"
        "Lets go and gather as much as we can."
        "Yay!"
        "Yes!">
      <pause "2">
    </do>
</episode>
"Ah, that's just some dry palm bark, nothing to fear. Hah, but you should've seen the look on your face!"

"Here, catch this too!"

"OK, looks like this is as much as I can carry. Let's head back."

"Here, catch this too!"

"OK, looks like this is as much as I can carry. Let's head back."

"Finally we made it!"

"Good job, high five!"

"Ha, I guess I am a little cheery, given that we are stranded in the middle of God knows where..."

"Do you know how to do it? I mean, other than using a magnifying glass and Doritos or Cheetos, have you seen how those things burn?"

"Anyway, if we happened to have some, I'd prefer eating them rather than burning them, besides, it's dark, so there's no use for a magnifying glass or mirror even if we had one."

"Wanna take a shot at building that fire, or should I do it?"

"All right hot shot! Show me how it's done."

"Let me try" "Yeah I can"

"I've got this" "Let me try" "I can do it" "I've got this" "Let me try" "I can do it"

"I'll give it a shot" "let me do it" "are you sure you can" "ill give it a try" "ill try first" "ill try it first" "absolutely ill do it" "i guess" "i guess since I have a little bit more knowledge than you" "yah ill do it" "yes ill do it" "ill do it" "ill build the fire" "I can do it" "I've got this" "Let me try" "Yeah I can"
I placed the tinder in the center and built the tepee.

All you have to do is strike the magnesium bar until there are some sparks that will get this thing to burn, and the rest should work out by itself.

Striking

Oh come on! That’s fine, I’ll do it.

You just have to generate more friction until it gets hot enough to burn, ugh.

It’s no use. I’m too tired and I’m not generating enough heat.

Come on, now you know how it’s done, just strike the magnesium bar as fast as you can

Faster! You almost have it! I can see a spark.

That’s impressive. To be honest I didn’t think we would be able to pull this off.

I’m so hungry, and so thirsty! But I guess it’s too dark to go out looking for food now.

Let’s not risk it, we can wait until tomorrow. Let’s enjoy what’s left of the night.

Look at those stars. Have you ever seen anything like it?

You must be from a rural area

You must be from around the city then. They are beautiful.

You must be from a rural area

You must be from around the city then. They are beautiful.

"in a dream" "i never have" "not as clear as here" "so many stars" "its a beautiful night" "no the stars are unique and beautiful" "only in my dreams" "no i have never seen anything like it" "No" "Not really" "I've never noticed" "Nope" "Never have" "no theres too much light back home"

"I placed the tinder in the center and built the tepee."

"All you have to do is strike the magnesium bar until there are some sparks that will get this thing to burn, and the rest should work out by itself."

"Striking"

"Oh come on! That’s fine, I’ll do it."

"You just have to generate more friction until it gets hot enough to burn, ugh."

"It’s no use. I’m too tired and I’m not generating enough heat."

"Come on, now you know how it’s done, just strike the magnesium bar as fast as you can"

"Faster! You almost have it! I can see a spark."

"That’s impressive. To be honest I didn’t think we would be able to pull this off."

"I’m so hungry, and so thirsty! But I guess it’s too dark to go out looking for food now"

"Let’s not risk it, we can wait until tomorrow. Let’s enjoy what’s left of the night."

"Look at those stars. Have you ever seen anything like it?"

"You must be from a rural area"

"You must be from around the city then. They are beautiful."
The night is beautiful. It reminds me when I would stay home in bed until late, reading, and drinking coffee. Ahh, I miss coffee.

Its nice to have some company for a change. I think I’ve started to value all the simple things a lot more while being here. You appreciate food, any kind of food, water, even whatever hole in the ground you can call home.

All the simple pleasures you know, we have it all granted outside of this place, but here, alone… Ah, don’t let me ruin the mood.

It was a lot worse before you got here; I think I even started talking to myself. Would’ve talked to a volleyball if one had washed up.

Well Im gonna hit the hay. Im so tired, goodnight.
<episode "discussfood3a">
  <do emotion "Happy" hands "Idle" Locomotion "Default" upperBody "Excitement" isSpecial "false" special "none">
    <speak "discussfood9.wav" "Awesome! Let's go find some food.">
    <pause "1"></pause>
    <speak "discussfood10.wav" "Maybe we can find some fish there.">
    <pause "1"></pause>
  </do>
  <go "discussfood4"></go>
</episode>

<episode "discussfood3b">
  <do emotion "Angry" hands "Idle" Locomotion "Stomp" upperBody "Default" isSpecial "false" special "none">
    <speak "discussfood11.wav" "Sure, like you have more important things to do. Think I am going to leave you here and do all the work? Come on!">
    <pause "1"></pause>
  </do>
  <go "discussfood4"></go>
</episode>

<episode "discussfood4">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Pointing" isSpecial "false" special "none">
    <pause "2"></pause>
    <speak "discussfood12.wav" "Here, I saved these from the kindle we brought yesterday. I had planned to use them as walking sticks, but I have a better idea.">
      <pause "1"></pause>
      <speak "discussfood13.wav" "So far we haven’t seen any traces of animals other than birds and fish.">
        <pause "1"></pause>
        <speak "discussfood14.wav" "I would think catching a live bird for eggs or to make stranded-island-roasted chicken is out of the question. ">
          <pause "1"></pause>
          <speak "discussfood15.wav" "However, I do think, or at least I hope, they’re good enough for fishing">
            <pause "1"></pause>
          </speak>
        </speak>
      </speak>
    </speak>
  </do>
  <nextscene "Scene_5A_Fishing"></nextscene>
</episode>

<scene "Scene_5A_Fishing">
  <episode "fishing5">
    <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none">
      <speak "fishing1.wav" "If we do catch some fish I hope they taste like salmon, those are my favourite. Especially with butter and garlic...">
        <pause "1"></pause>
        <speak "fishing2.wav" "I think it's obvious I've been trapped here for a while haha">
          <pause "3"></pause>
          <speak "fishing3.wav" "Well here we are, let's try this spot.">">
            <pause "3"></pause>
            <speak "fishing4.wav" "Have you ever fished? I mean, without a fishing rod?">
              <pause "1"></pause>
            </speak>
          </speak>
        </speak>
      </speak>
      <decide speech="">
        <createRule "fished">
          <go "fishing5a"></go>
          <tag "havefished">
            <items "yes" "yeah" "yup" "yeah i have" "yes i have" "yup i have" "i always go fishing without a fishing rod" "i go spear fishing" "i used to spear fish" "i used to fish all the time" "i never use a fishing rod" "i use my bare hands" "only once" "always" "sometimes" "all the time" "i use a net to fish" "i always use a net to fish" "i always use a spear to fish">
              </tag>
          </go>
          <tag "neverfished">
            <items "no" "nope" "nah" "never" "not once" "no i never have" "nope i never have" "nah I never have" "ive never fished without a fishing rod" "i dont like fishing without a rod" "i have never fished without a fishing rod" "i always use a fishing rod" "i cant fish without a rod" "ive never fished" "i dont go fishing" "i hate fishing" "im terrible at fishing" "im horrible at fishing" "never in my life" "i dislike fishing" "im too lazy to fish" "im not the best at fishing">
              </tag>
          </go>
        </rule>
        <decide>
        </decide>
      </decide>
    </do>
  </episode>
</scene>
I used to go fishing with my dad when I was younger but we always had our fishing rods and bait, and even then patience isn’t my greatest virtue."

"I guess trying to stab them would be the best way to catch them".

"Well don’t worry, I used to go fishing with my dad all of the time."

"It’ll be a piece of cake even without a fishing rod, I’ll show you how to catch them.

"Aiming and stabbing at them would be the best way to catch them, just be careful.

"Oh you were so close! Try again!"

"These fish are pretty slippery... but, we only need a couple more!"

"Not quite like that. Try straightening your arm, and stabbing with a quick motion!"

"Yeah get that last one, I know you can do it!"

"Yeah nice one, you got that one dead center! Hurry try for the next one!"
<speak "fishing17.wav" "Who would have known you had such sick fishing skills? Great Job!"> <pause "1"> </go "fishing9"> </go>
</episode>

<episode "fishing8b">
<do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
<speech "fishing18.wav" "Argh, the one that got away!"> <pause "1"> </go "fishing8"> </go>
</episode>

<episode "fishing9">
<do emotion "Surprised" hands "Idle" Locomotion "Default" upperBody "Surprised" isSpecial "false" special "none"> <speech "fishing19.wav" "Three fish?! Okay my turn to try!"> <pause "2"> </go "fishing9"> </go>
<do emotion "Happy" hands "Idle" Locomotion "Default" upperBody "Throwing" isSpecial "false" special "none"> <pause "2"> </go "fishing20.wav" "Yeah, I got one!"> <pause "1"> </go "fishing21.wav" "This is enough. We should only get what we are going to eat, otherwise, if there are any predators in this place we might attract them with the left overs."> <pause "1"> </go "fishing22.wav" "That was pretty fun though, let's go back and cook"> <pause "1"> </go>
</episode>

<nextscene "Scene_6_Cooking">
</scene>

<scene "Scene_6_Cooking">
<episode "cooking1">
<do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none"> <pause "1"> </go "cooking1.wav" "Let’s heat this up."> <pause "2"> </go "cooking2.wav" "Woah, weve got to be careful, the fire is about to die, and I really dont want to light it up again."> <pause "1"> </go "cooking3.wav" "I can see some embers still burning. Here Ill clean up the fish while you revive our fire."> <pause "1"> </go "cooking4.wav" "Just use your hands to vent it so that it catches up."> <pause "1"> </go>
<decide gesture="Ventilate"> <go "cooking2"> "Ventilate" </go>
</decide>
</episode>

<episode "cooking2">
<do emotion "Happy" hands "Idle" Locomotion "Default" upperBody "Excitement" isSpecial "false" special "none"> <pause "1"> </go "cooking5.wav" "Great job, now lets get to work on these fish."> <pause "1"> </go "cooking6.wav" "Im not particularly fond of fish, much less of raw fish."> <pause "1"> </go "cooking7.wav" "Dont get me wrong, fishing is fun, its just not my favorite dish."> <pause "2"> </go "cooking8.wav" "I guess I liked fishing because it was my chance to spend some time with dad."> <pause "1"> </go "cooking9.wav" "$Do you miss your family?" > <go "cooking3"> </go>
</episode>

<episode "cooking3">
<do emotion "Neutral" hands "Idle" Locomotion "SittingDown" upperBody "Default" isSpecial "false" special "none"> <pause "2"> </go "cooking9.wav" "$Do you miss your family?" > <createRule "havefamily"> <tag "familyyes"> <go "cooking3a"> </go>
</tag>
</createRule>
</episode>
"yes" "yeah" "of course i do" "yes, just like anybody else" "i think i do" "i hope theyre still living" "i guess i do" "just my pets" "my grandma" "my spouse" "my boyfriend" "my partner" "my wife" "my husband" "my fiance" "my children" "i do parents and a brother" "yeah my wife" "yeah my husband" "yeah my girlfriend" "yeah my boyfriend" "yes my beautiful mother" "yes my father" "yeah i do my parents and siblings" "yeah like i said friends family and a girlfriend"
Be careful not to swallow any bones!

Let's dig in!

I was also thinking that we should set up a smoke signal just in case any ships or helicopters pass by near the beach.

We have a water and food supply, and a somewhat safe shelter, we've done some exploring and we have fire.

I guess it's time to start thinking about how we can get off this island. We should try to signal for help.

I've been thinking... we are making ourselves comfortable here.

The fastest and easiest thing to do probably, is to create an SOS signal out of wood or rocks close to the beach.

I guess it's time to start thinking about how we can get off this island. We should try to signal for help.

I was also thinking that we should set up a smoke signal just in case any ships or helicopters pass by near the beach.

Worst thing that could happen is that no one sees it and we spend a day or two working at it. It's not like there is that much to do meanwhile either.
<pause "1"> <speak "cooking40.wav" "These fish are sooooooo delicious! I guess when you are this hungry everything tastes like heaven.""> <pause "4"> <do emotion "Neutral" hands "Idle" Locomotion "StandUp" upperBody "Default" isSpecial "false" special "none"> <pause "1"> <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none"> <speak "cooking41.wav" "As much as I want to take a nap, we should probably head back to the beach and start our rescue operation!"> <pause "4"> </episode>

<nextscene "Scene_6A_Help">
</scene>

<scene "Scene_7_AgentCry">
<episode "agentcry1">
<do emotion "Scared" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none"> <pause "2"> <speak "agentcry1.wav" "Ahhhhh! Help! By the river!"> <pause "1"> </episode>

<nextscene "Scene_7A_Injury">
</scene>

<scene "Scene_7A_Injury">
<episode "injury2">
<do emotion "Fear" hands "Idle" Locomotion "Crouching" upperBody "Scared" isSpecial "false" special "none"> <pause "1"> <speak "injury1.wav" "Oh I'm so glad you're here! I'm stuck, and it hurts... badly ahhhh!"> <pause "1"> <speak "injury2.wav" "Please, please, please! Don't let it be broken!."> <go "injury3" />
</episode>

<episode "injury3">
<do emotion "Fear" hands "Idle" Locomotion "Crouching" upperBody "Scared" isSpecial "false" special "none"> <pause "1"> <speak "injury3.wav" "Thank you. Ahh I don't think I'll be able to walk... I can try though."> <pause "1"> <do emotion "Fear" hands "Idle" Locomotion "StandUp" upperBody "Default" isSpecial "false" special "none"> <pause "2"> <do emotion "Fear" hands "Idle" Locomotion "Crouching" upperBody "Point" isSpecial "false" special "none"> <pause "1"> <speak "injury4.wav" "AGH! AGH! Dammit! I CAN'T walk on it. Here, help me up."> <pause "1"> <decide gesture>
  <go "injury4"> "Lift Hand" </go>
</decide>
</episode>

<episode "injury4">
<do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Exhaustion" isSpecial "false" special "none"> <pause "1"> <do emotion "Sad" hands "Idle" Locomotion "Default" upperBody "Exhaustion" isSpecial "false" special "none"> <pause "1"> <speak "injury5.wav" "Thank you, trying to stand on my own was a bad idea!"> <pause "1"> <speak "injury6.wav" "Next time I AM using the improvised fishing rod as a walking stick"> <pause "1"> <speak "injury7.wav" "I hope there is a natural painkiller around. I've seen some mushrooms but I think it would be dangerous to try them"> <pause "1"> <speak "injury8.wav" "How are your first aid skills? Do you have any medical training?"> <pause "1"> <decide speech>
  <createRule "firstAidSkills">
    <go "injury4a"> </go>
  <tag "firstaidskillsyes">
    <items "yes" "yeah" "yup" "i used to be a nurse" "i used to be a doctor" "i might be of assistance" "i have a little bit of first aid training" "a little bit" "i took a course" "i can try" "im a certified nurse" "im a certified doctor" "only from what i remember from
I took health class in high school. I took health class in high school. Lots of experiences. You'll be fine. I know some basic stuff. Very little.

Tag "firstaidskillsno"

No. No. I've never done this before. None whatsoever. No but I can try. No but you don't have any options. I don't know anything. I'm no nurse. I'm no doctor. I'm sorry I don't have medical training.

Rule

Episode "injury4a"

Do emotion "Happy" hands "Idle" Locomotion "StandUp" upperBody "Nodding" isSpecial "false" special "none"

Speak "injury9.wav" "Ohhhh, that's a relief well when we are in the cave you're gonna have to do your best to patch me up then."<pause "2">

Episode "injury4b"

Do emotion "Sad" hands "Idle" Locomotion "StandUp" upperBody "Default" isSpecial "false" special "none"

Speak "injury10.wav" "Darn, well maybe we can figure something once we get to the cave."

Episode "injury5"

Do emotion "Sad" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none"

Speak "injury11.wav" "I actually just wanted some time alone."

Do emotion "Sad" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none"

Speak "injury12.wav" "I was sure I could handle it by myself."

Speak "injury13.wav" "I was just walking along the edge of the river, but the current was stronger than I thought."

Episode "injury6"

Speak "injury14.wav" "After surviving what we have, I feel like the guy who used to jump from the top of Niagara falls, only to die by slipping over an orange peel."

Episode "injury6"

Speak "injury15.wav" "Don't worry, I'm very much alive but it hurts, I need to rest, lets head to the cave."

Episode "Scene_8_CureFootAtTheCave"

Scene "Scene_8_CureFootAtTheCave"

Episode "curefoot1"

Do emotion "Sad" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none"

Speak "curefoot1.wav" "I'm just going to lay down over there, I'm feeling dizzy."

Do emotion "Sad" hands "Idle" Locomotion "Stumble" upperBody "Default" isSpecial "false" special "none"

Episode "curefoot2"

Speak "curefoot2.wav" "Can you tell me how bad it is?"
"It hurts! Ah"
"What?! Oh, no no no…something bit me……I’m not feeling so well."
"Everything is getting darker"
"Tell me the truth! Please tell me I’m going to be okay."
"Okay okay okay….I’m going to be okay… thanks, I’m starting to calm down."
"How could you say that?! This isn’t a joke, please help me."
"I know I can trust you, I wouldn’t have been able to survive as long as I did without you."
"We just need to take this one step at a time and figure out what to do."
"I dont think we are equipped to deal with this, our best bet is to find help soon."
<episode "curefoot6">
    <speak "curefoot12.wav" "Mushrooms are out of the question, they might make me worse.">
    <pause "1"/>
    <speak "curefoot13.wav" "There is no one else on this island either, or we would have found them by now.">
    <pause "1"/>
    <go "curefoot7" /></episode>

<episode "curefoot7">
    <pause "3">
        <do emotion "Sad" hands "Idle" Locomotion "LayingDown" upperBody "Default" isSpecial "false" special "none">
            <pause "1"/>
            <speak "curefoot14.wav" "I've been here for days now, surely someone will come look for us soon">
            <pause "2">
            <speak "curefoot15.wav" "Ugh, I feel like I'm burning up. This was not what I had in mind to fight the cold during the night.">
            <pause "1"/>
            <speak "curefoot16.wav" "It might be a good idea to head to the beach, if no one comes and I don't get any better, this could be my only way out of this">
        </do>
    </episode>

<episode "curefoot8">
    <do emotion "Neutral" hands "Idle" Locomotion "LayingDown" upperBody "Default" isSpecial "false" special "none">
        <pause "1"/>
        <speak "curefoot17.wav" "I'm just going to try to rest for a little bit. I feel exhausted. Maybe the fever will go away when I wake up">
        <pause "3">
    </episode>

<nextscene "Scene_8A_LightSignal">
</scene>

<scene "Scene_8A_LightSignal">
    <episode "lightsignal1">
        <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
            <pause "1"/>
            <speak "lightsignal1.wav" "Remember, just strike the bar to build a fire, just like we did with the campfire. I'm trusting you!">
            <pause "1"/>
            <decide gesture>
                <go "lightsignal2" "Striking" /></decide>
        </do>
    </episode>

    <episode "lightsignal2"/>
    <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
        <speak "lightsignal2.wav" "Great job, I knew you could do it!">
        <pause "3">
    </episode>

<nextscene "Scene_8B_ReturntoAgent">
</scene>

<scene "Scene_8B_ReturntoAgent">
    <episode "returnagent1">
        <do emotion "Neutral" hands "Idle" Locomotion "SittingDown" upperBody "Default" isSpecial "false" special "none">
            <speak "returnagent1.wav" "Oh...nm you're... bacce.k ">
            <pause "1"/>
            <do emotion "Sad" hands "Idle" Locomotion "LayingDown" upperBody "HeadShake" isSpecial "false" special "none">
                <go "returnagent2" /></do>
        </do>
    </episode>

    <episode "returnagent2"/>
    <do emotion "Surprised" hands "Idle" Locomotion "SittingDown" upperBody "Default" isSpecial "false" special "none">
        <speak "returnagent2.wav" " Argh Did I doze off? Please don’t let me fall asleep again. This fever is giving me nightmares. I... I might black out">
        <pause "1"/>
    </episode>

<nextscene "Scene_8C_FinalHour">
</scene>
<scene "Scene_8C_FinalHour">

<episode "finalhour1">
  <do emotion "Neutral" hands "Idle" Locomotion "SittingDown" upperBody "Default" isSpecial "false" special "none">
    <pause "1"/>
    <speak "finalhour1.wav" "It's been a while, and I am not feeling any better."/>
    <pause "3"/>
    <speak "finalhour2.wav" "Everything is spinning. I will just... close my eyes now..."/>
    <pause "2"/>
    <do emotion "Scared" hands "Idle" Locomotion "LayingDown" upperBody "Cold" isSpecial "false" special "none">
      <speak "finalhour3.wav" "That... sound... helicopter. Go... signal it... you're... our only... chance."/>
      <pause "1"/>
      <go "finalhour2"></go>
  </episode>

<episode "finalhour2">
  <do emotion "Scared" hands "Idle" Locomotion "Walking" upperBody "CrossedArms" isSpecial "false" special "none">
    <pause "2"/>
    <do emotion "Scared" hands "Idle" Locomotion "Walking" upperBody "CrossedArms" isSpecial "false" special "none">
      <pause "2"/>
  </episode>

<nextscene "Scene_8F_RunBeach">
</scene>

<scene "Scene_8F_RunBeach">

<episode "runbeach1">
  <do emotion "Neutral" hands "Idle" Locomotion "Walking" upperBody "Default" isSpecial "false" special "none">
    <speak "runbeach1u1.wav" "Hurry signal the helicopter before it flies away!"/>
    <pause "1"/>
    <decide gesture>
      <go "runbeach2" "Arms Up"></go>
    </decide>
  </episode>

<episode "runbeach2">
  <do emotion "Neutral" hands "Idle" Locomotion "Default" upperBody "Default" isSpecial "false" special "none">
    <pause "1"/>
  </episode>

<nextscene "Scene_End">
</scene>

<scene "Scene_End">

<episode "end1">
  <do emotion "Happy" hands "Idle" Locomotion "Default" upperBody "Excitement" isSpecial "false" special "none">
    <pause "1"/>
    <speak "end1.wav" "I survived thanks to you. I'll never forget it."/>
    <pause "60"/>
  </episode>
</scene>
Appendix B: Interaction Results

Table 14: Survey items as seen by participants and their respective ID.

<table>
<thead>
<tr>
<th>Item</th>
<th>Question ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>The agent understood me</td>
<td>U1</td>
</tr>
<tr>
<td>The agent seemed unengaged</td>
<td>U2</td>
</tr>
<tr>
<td>The agent was not paying attention to me</td>
<td>U3</td>
</tr>
<tr>
<td>I didn’t understand the agent</td>
<td>U4</td>
</tr>
<tr>
<td>The agent was excited</td>
<td>P1</td>
</tr>
<tr>
<td>The agent’s movements were not natural</td>
<td>P2</td>
</tr>
<tr>
<td>I sensed a physical connection with the agent</td>
<td>P3</td>
</tr>
<tr>
<td>The agent’s gestures were not lively</td>
<td>P4</td>
</tr>
<tr>
<td>The agent was friendly</td>
<td>C1</td>
</tr>
<tr>
<td>The agent and I worked towards a common goal</td>
<td>C2</td>
</tr>
<tr>
<td>The agent and I did not seem to connect</td>
<td>C3</td>
</tr>
<tr>
<td>I feel the agent trusts me</td>
<td>C4</td>
</tr>
</tbody>
</table>

Table 15: Non-gesture condition questionnaire results per participant without reverse coding. All values represent positive aspects of the agent.

<table>
<thead>
<tr>
<th></th>
<th>U1</th>
<th>U2</th>
<th>P1</th>
<th>P2</th>
<th>C1</th>
<th>U3</th>
<th>C2</th>
<th>C3</th>
<th>P3</th>
<th>P4</th>
<th>C4</th>
<th>U4</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>N01</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4.17</td>
</tr>
<tr>
<td>N02</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3.75</td>
</tr>
<tr>
<td>N03</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>3.58</td>
</tr>
<tr>
<td>N04</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2.58</td>
</tr>
<tr>
<td>N05</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4.58</td>
<td></td>
</tr>
<tr>
<td>N06</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4.00</td>
</tr>
<tr>
<td>N07</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3.75</td>
</tr>
<tr>
<td>N08</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3.17</td>
</tr>
<tr>
<td>N09</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>N10</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3.67</td>
</tr>
<tr>
<td>N11</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3.83</td>
</tr>
<tr>
<td>N12</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>3.00</td>
</tr>
<tr>
<td>N13</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3.00</td>
</tr>
<tr>
<td>N14</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>3.42</td>
</tr>
<tr>
<td>N15</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>2.50</td>
</tr>
<tr>
<td>N16</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3.67</td>
</tr>
<tr>
<td>N17</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3.58</td>
</tr>
<tr>
<td>N18</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3.17</td>
</tr>
<tr>
<td>N19</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4.25</td>
</tr>
<tr>
<td>N20</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3.17</td>
</tr>
<tr>
<td>N21</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2.83</td>
</tr>
<tr>
<td>N22</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4.50</td>
<td></td>
</tr>
<tr>
<td>N23</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>2.83</td>
</tr>
<tr>
<td>N24</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3.92</td>
</tr>
<tr>
<td>N25</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>3.67</td>
</tr>
</tbody>
</table>
Table 16: Gesture condition questionnaire results per participant without reverse coding. All values represent positive aspects of the agent.

<table>
<thead>
<tr>
<th></th>
<th>U1</th>
<th>U2</th>
<th>P1</th>
<th>P2</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>P3</th>
<th>P4</th>
<th>C4</th>
<th>U4</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>G01</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G02</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G03</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G04</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G05</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>G06</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G07</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G08</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G09</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G10</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G11</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>G12</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G13</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G14</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G15</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G16</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>G17</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G18</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G19</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>G20</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>G21</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G22</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G23</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G24</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G25</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G26</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G27</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G28</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>G29</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G30</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>G31</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>AVG</td>
<td>3.45</td>
<td>3.68</td>
<td>4.00</td>
<td>2.52</td>
<td>4.61</td>
<td>3.29</td>
<td>4.71</td>
<td>3.61</td>
<td>2.39</td>
<td>3.29</td>
<td>4.32</td>
<td>4.61</td>
</tr>
</tbody>
</table>
### Appendix C: Demographics

Table 17: Demographics including gender, age, college level, and first, second and third languages.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Race/Ethnicity</th>
<th>Year in College</th>
<th>[English]</th>
<th>[Spanish]</th>
<th>[Other]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>30</td>
<td>African American</td>
<td>fifth year or above</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Asian</td>
<td>first year</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>28</td>
<td>Asian, White</td>
<td>second year</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>18</td>
<td>Hispanic</td>
<td>first year</td>
<td>first</td>
<td>second</td>
<td>third</td>
</tr>
<tr>
<td>Male</td>
<td>18</td>
<td>Hispanic</td>
<td>second year</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Hispanic</td>
<td>second year</td>
<td>first</td>
<td>second</td>
<td>third</td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>20</td>
<td>Hispanic</td>
<td>third year</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>20</td>
<td>Hispanic</td>
<td>second year</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>Hispanic</td>
<td>third year</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>Hispanic</td>
<td>third year</td>
<td>first</td>
<td>second</td>
<td>third</td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>Hispanic</td>
<td>third year</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>22</td>
<td>Hispanic</td>
<td>third year</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>22</td>
<td>Hispanic</td>
<td>fifth year or above</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>23</td>
<td>Hispanic</td>
<td>fifth year or above</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>26</td>
<td>Hispanic</td>
<td>third year</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>26</td>
<td>Hispanic</td>
<td>fifth year or above</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>26</td>
<td>Hispanic</td>
<td>fifth year or above</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>28</td>
<td>Hispanic</td>
<td>second year</td>
<td>first</td>
<td>third</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>24</td>
<td>Hispanic, White</td>
<td>fifth year or above</td>
<td>first</td>
<td>second</td>
<td>third</td>
</tr>
<tr>
<td>Male</td>
<td>24</td>
<td>Hispanic, White</td>
<td>fifth year or above</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>White</td>
<td>fourth year</td>
<td>first</td>
<td>third</td>
<td>second</td>
</tr>
<tr>
<td>Female</td>
<td>22</td>
<td>White</td>
<td>fourth year</td>
<td>first</td>
<td>third</td>
<td>second</td>
</tr>
<tr>
<td>Male</td>
<td>24</td>
<td>White</td>
<td>fifth year or above</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25</td>
<td>White</td>
<td>fifth year or above</td>
<td>first</td>
<td>second</td>
<td>third</td>
</tr>
<tr>
<td>Male</td>
<td>27</td>
<td>White</td>
<td>fifth year or above</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>28</td>
<td>White</td>
<td>fourth year</td>
<td>first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>29</td>
<td>White</td>
<td>fifth year or above</td>
<td>first</td>
<td>second</td>
<td>third</td>
</tr>
<tr>
<td>Male</td>
<td>32</td>
<td>White</td>
<td>fifth year or above</td>
<td>first</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>27</td>
<td>Asian</td>
<td>fifth year or above</td>
<td>second</td>
<td></td>
<td>first</td>
</tr>
<tr>
<td>Female</td>
<td>18</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>second</td>
<td></td>
<td>first</td>
</tr>
<tr>
<td>Male</td>
<td>18</td>
<td>Hispanic</td>
<td>first year</td>
<td>second</td>
<td></td>
<td>first</td>
</tr>
<tr>
<td>Male</td>
<td>18</td>
<td>Hispanic</td>
<td>first year</td>
<td>second</td>
<td></td>
<td>first</td>
</tr>
<tr>
<td>Male</td>
<td>18</td>
<td>Hispanic</td>
<td>first year</td>
<td>second</td>
<td></td>
<td>first</td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Hispanic</td>
<td>second year</td>
<td>second</td>
<td></td>
<td>first</td>
</tr>
<tr>
<td>Gender</td>
<td>Age</td>
<td>Ethnicity</td>
<td>Year</td>
<td>Grade</td>
<td>College</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-----------------</td>
<td>-------------</td>
<td>--------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Hispanic</td>
<td>first year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Hispanic</td>
<td>first year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Hispanic</td>
<td>first year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19</td>
<td>Hispanic</td>
<td>first year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>20</td>
<td>Hispanic</td>
<td>second year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>20</td>
<td>Hispanic</td>
<td>third year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>20</td>
<td>Hispanic</td>
<td>third year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>Hispanic</td>
<td>third year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>21</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>22</td>
<td>Hispanic</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>22</td>
<td>Hispanic</td>
<td>fourth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>23</td>
<td>Hispanic</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25</td>
<td>Hispanic</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25</td>
<td>Hispanic</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>26</td>
<td>Hispanic</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>28</td>
<td>Hispanic</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>36</td>
<td>Hispanic</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>22</td>
<td>White</td>
<td>fifth year</td>
<td>second</td>
<td>first</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>23</td>
<td>Hispanic, White</td>
<td>fifth year</td>
<td>third</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>first</td>
<td></td>
</tr>
</tbody>
</table>
Figure 23: Demographics survey
Appendix D: Agent Scripting Language Schema

This appendix contains the latest version of the agent scripting language schema, which includes several updates that occurred after the Jungle Game tests. For additional information on how to write or interpret this scripts go to chapter 5.

```xml
<?xml version = "1.0" encoding = "UTF-8" ?>
<xs:schema xmlns:xs = "">
    <xs:element name = "scene" type="xs:string">
        <xs:complexType>
            <xs:element name = "episode" type = "xs:string">
                <xs:complexType>
                    <xs:element name = "do" type="xs:string">
                        <xs:complexType>
                            <xs:element name = "animation">
                                <xs:complexType>
                                    <xs:element name = "face" type="xs:string">
                                        <xs:complexType>
                                            <xs:attribute name = "id" type = "xs:string" use = "required"/>
                                            <xs:attribute name = "speed" type = "xs:positiveInteger" use = "required"/>
                                            <xs:attribute name = "amp" type = "xs:positiveInteger" use = "required"/>
                                        </xs:complexType>
                                    </xs:element>
                                    <xs:element name = "hands" type="xs:string">
                                        <xs:complexType>
                                            <xs:attribute name = "id" type = "xs:string" use = "required"/>
                                            <xs:attribute name = "speed" type = "xs:positiveInteger" use = "required"/>
                                            <xs:attribute name = "amp" type = "xs:positiveInteger" use = "required"/>
                                        </xs:complexType>
                                    </xs:element>
                                    <xs:element name = "lowerBody" type="xs:string">
                                        <xs:complexType>
                                            <xs:attribute name = "id" type = "xs:string" use = "required"/>
                                            <xs:attribute name = "speed" type = "xs:positiveInteger" use = "required"/>
                                            <xs:attribute name = "amp" type = "xs:positiveInteger" use = "required"/>
                                        </xs:complexType>
                                    </xs:element>
                                    <xs:element name = "upperBody" type="xs:string">
                                        <xs:complexType>
                                            <xs:attribute name = "id" type = "xs:string" use = "required"/>
                                            <xs:attribute name = "speed" type = "xs:positiveInteger" use = "required"/>
                                            <xs:attribute name = "amp" type = "xs:positiveInteger" use = "required"/>
                                        </xs:complexType>
                                    </xs:element>
                                    <xs:element name = "fullBody" type="xs:string">
                                        <xs:complexType>
                                            <xs:attribute name = "id" type = "xs:string" use = "required"/>
                                            <xs:attribute name = "speed" type = "xs:positiveInteger" use = "required"/>
                                            <xs:attribute name = "amp" type = "xs:positiveInteger" use = "required"/>
                                        </xs:complexType>
                                    </xs:element>
                                </xs:complexType>
                            </xs:element>
                        </xs:complexType>
                    </xs:element>
                </xs:complexType>
            </xs:element>
        </xs:complexType>
    </xs:element>
</xs:schema>
```
<xs:complexType>
    <xs:attribute name="ep" type="xs:string" use="required"/>
</xs:complexType>
</xs:element>

<xs:element name="tag" type="xs:string">
    <xs:complexType>
        <xs:element name="items">
            <xs:complexType>
                <xs:element name="i" type="xs:string"/>
            </xs:complexType>
        </xs:element>
    </xs:complexType>
    <xs:attribute name="id" type="xs:string" use="required"/>
</xs:element>
</xs:complexType>
</xs:element>

<xs:element name="nextScene">
    <xs:complexType>
        <attribute name="id" type="xs:string" use="required"/>
    </xs:complexType>
</xs:element>
</xs:complexType>
Vita

Ivan Gris recently earned his Ph.D. from the University of Texas at El Paso, where he worked with David Novick. He was named outstanding doctoral student in Computer Science at UTEP in the spring of 2015. Ivan has authored or co-authored 14 refereed publications. He develops full-body embodied conversational agents and immersive environments at UTEP's Interactive Systems Group, as part of the Advanced Agent Engagement Team. Ivan is also a co-founder of Inmerssion, where he works on creating automated virtual characters and virtual reality applications. He likes to explore novel human-computer interaction techniques to deliver digital content through believable characters and engaging storytelling scenarios.

Permanent address: 306 E Paisano Dr. PMB 117
El Paso, TX, 79901

This dissertation was typed by Ivan Gris Sepulveda.