AVARI: Animated Virtual Agent Retrieving Information

Lauren Cairco
Winthrop University
701 Oakland Avenue
Rock Hill, SC 29733
(803) 548-3079
caircol2@winthrop.edu

Dale-Marie Wilson, Vicky Fowler, Morris LeBlanc
The University of North Carolina at Charlotte
9201 University City Boulevard
Charlotte, NC 28223
(704) 687-7088
{DaleMarie.Wilson, vdfowler, msleblan}@uncc.edu

ABSTRACT
Avari is a virtual receptionist for the Computer Science department at The University of North Carolina at Charlotte. Her components include background subtraction to detect a person's presence, speech recognition, audio and visual devices to communicate with passersby. Deployed in a public setting, we investigate the reactions and interactions of passersby with Avari. We describe the design and architecture of the virtual human and discuss the effectiveness of a publicly deployed virtual human.

Categories and Subject Descriptors
H.5.1: Multimedia Information Systems: Artificial, augmented, and virtual realities; H.5.2 User Interfaces: Natural language, Voice I/O, User-centered design

General Terms
Design, Human Factors

Keywords
virtual humans, human-computer interaction, human-centered computing

1. INTRODUCTION
Previous research shows that virtual humans are effective in several applications that provide social conversation, training, and information. However, the current use of virtual humans is limited, with most applications used only within the labs where they are developed. These studies provide us with little information about how people generally interact with virtual humans and their effectiveness in the absence of previous exposure and expectations.

Avari (Animated Virtual Agent Retrieving Information) is a virtual receptionist that provides information, both formal and informal, about the computer science faculty members at the University of North Carolina at Charlotte (UNCC). She is capable of enticing passersby to approach and interact with her, recognizing a person's presence directly in front of her, and interacting with persons via speech and visual aids and directing conversations towards a specific goal. Avari communicates using natural speech and her personality is conveyed through her facial expressions, tone and verbal quips. The mixed-reality architecture consists of the following components: a wooden desk behind which Avari sits; two monitors (one for Avari's virtual head and the other for visual cues); a camera for motion detection, and a wooden frame for Avari's body. The main focus of this study is to investigate the approachability and effectiveness of a stand-alone virtual human.

Data collection focuses on the timing and content of user interactions. Avari was placed outside an undergraduate computer lab in the computer science building at UNCC, without perceivable supervision. Preliminary analysis of data reveals interesting trends in the behavior of persons approaching and interacting with a virtual human.

ACMSE '09 March 19-21, 2009, Clemson, SC, USA.

2. BACKGROUND
2.1 Social interaction with virtual humans
Researchers have shown that people respond to virtual humans in many of the same ways that they react to other people. Zanbaka et al. [12] found that the presence of a virtual human could affect task performance through social inhibition, and Slater et al. [9] found that in virtual meetings, avatars could elicit emotions in people such as embarrassment, self-awareness, and irritation. People interact and treat computers based on perceived human characteristics such as the computer's helpfulness, expertise, and friendliness [8]. Raij et. al. [7] examined perceived similarities and differences in experiencing an interpersonal scenario with a real and virtual patient, and found lower ratings on participants' rapport and conversational flow with the virtual patient. This was attributed to the limited expressiveness of the virtual patient.

Evidence suggests that human communication consists of a high bandwidth of modalities such as gestures, facial expressions, speech, and body language [6], and research shows that using both speech and gestures contributes to making virtual human interfaces more lifelike and believable [3]. Researchers have also shown that virtual human interfaces can provide feedback to human users using multiple channels such as speech, gestures, and facial expressions. For example, Rea, built by Cassell, et al., is a

virtual real estate agent capable of understanding speech and gaze, and of planning multimodal utterances from propositional abstract representations [3]. Rea also keeps a model of interpersonal distance with the user, and uses small talk to reduce interpersonal distance if she notices a lack of closeness with the user. Another example is Gandalf, a humanoid who guides a user through the solar system. Gandalf responds to user speech, gaze, and motion with appropriate gestures, speech, and head movement [10]. Gandalf's behavioral rules are derived from psychology literature on human to human interaction.

2.2 Virtual humans as receptionists

Babu, et al. [1,2] presented Marve, a virtual receptionist in a computer lab at UNCC. Marve uses computer vision and facial recognition to greet lab members and guests when they enter the lab, and is capable of social interaction with his users, including telling knock-knock jokes, commenting on the current weather or recent movies, and recording and relaying voice messages between lab members. Marve receives input through spoken keywords. Although most of his users were lab members, in this study, Bonnie, the building custodian also interacted with Marve. Bonnie provides some insight into how people with little to no computer experience might interact with virtual humans. Whether described by Bonnie or by members of the lab, the study showed that Marve was perceived and described by his users as a social entity instead of a computer. Other virtual agents used as receptionists include MIT's MACK [4], a mixed reality agent who gives directions to visitors of MIT's Media Lab and answers questions about the employees. MACK uses speech recognition to recognize user questions, and answers using speech and gestures. Another example is CMU's Valerie [5], a roboreceptionist who can give directions to places on campus and hold social conversations about campus gossip. Valerie accepts typed questions and responds via speech.

3. SYSTEM OVERVIEW

Avari is a virtual receptionist who is designed to answer user questions using both visual aids and speech. Her architecture is mixed-reality, consisting of real-word and virtual components. Avari's knowledge base consists of general, professional and personal facts about computer science professors at UNCC. She was developed via a user-centered design with a target population of inexperienced users with respect to virtual human functions and technologies.

To entice user interaction, Avari engages in several idle behaviors including joke telling, singing, sneezing, and complimenting passersby. During consecutive phases of inactivity, she performs a randomly selected behavior every 60 seconds. Avari's vision component can recognize a user standing directly in front of her. When a user is detected, Avari initiates the dialogue.

First, Avari introduces herself and explains her functionality. Next, she guides the user through a conversation, prompting the user for the faculty member that they would like to talk about, the category of interest, and finally, the question that they would like to have answered. This interaction represents a complete dialogue unit. Once completed, Avari presents the user with the following options (see figure 1):

- Initiate a new dialogue unit with same professor, same category
- Initiate a new dialogue unit with same professor, different category
- Initiate a new dialogue unit with different professor, different category
- Terminate conversation.

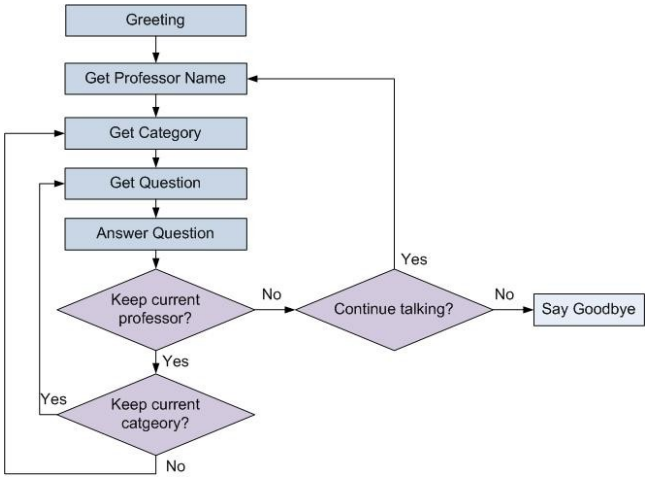

Figure 1: A state transition diagram for conversation flow

Additionally Avari presents verbal prompts that direct the dialogue flow and visual cues which are displayed on an accompanying monitor. The visual cues include pictoral representations of available topics (as shown in Figure 2), a "Listen" prompt indicating that the user should be listening, a "Talk" prompt indicating that Avari is listening for input, and sample questions. Avari's phrases are randomly selected in each conversational step, increasing the probability of unique exchanges over multiple interactions.

Figure 2: Visual cues for the "Hobbies and Interests" category

Following is a sample dialogue unit. This is an excerpt of actual conversation logged by Avari:

(user approaches)

Avari: Hey, I'm Avari! I know things about the professors here. What professor do you want information on?

(list of professors is displayed)

User: Min Shin

Avari: Oh, I know Min Shin. What would you like to know? Ask me about one of the categories on the other screen.

(list of categories is displayed)

User (as recognized by Avari): and background

(This is matched to the category "Background Information".)

Avari: All right, what do you want to know?

(prompt screen is displayed with images and keywords relating to facts in the background information category)

User (as recognized by Avari): well is the born

(This question is matched to "Where was he born?" in the database.)

Avari: Professor Min Shin said that he was born in Seoul, Korea. Do you want to keep talking about Min Shin?

User: No

Avari: Is there anything else I can tell you?

User: Yes

Avari: Okay, who else do you want to know about?

(list of professors is displayed)

User: Larry Hodges

(conversation continues)

3.1 Real-World Components

3.1.1 Physical appearance

Avari appears as a female receptionist sitting behind her desk. The desk is a hollow, wooden box that houses the hardware components. Avari's body consists of a wooden structure that extends from the top of the desk. A shirt stuffed to represent a female's upper torso is hung on the wooden frame. Avari is a young female in her early twenties and this is represented in her fashion. Her face, the virtual component, is depicted on a monitor mounted at eye level on top of her physical body. Included on Avari's desk are speakers for audio output, a microphone for speech input, and an additional monitor that displays visual cues and prompts. A camera is discretely positioned slightly below knee level of an average-sized user on the outside of her cabinet. The entire desk is mounted on lockable wheels that afford Avari's mobility. (See figure 3).

Figure 3: Avari's desk

3.1.2 Knowledge Repository

Seventeen computer science professors at UNCC answered twenty-five personal and professional questions through an online survey. These questions are divided into five categories. Examples include:

Category: General Information
Questions: Where is your office?
What is your e-mail address?

Category: Background
Questions: Where were you born?
Why did you choose this career?

Category: Hobbies & Interests
Questions: What is one interesting fact about you?
What is your favorite book?

Category: Opinions
Questions: What is your favorite thing about UNCC?
What would you do if every computer in the world crashed?

Category: Advice for Students
Questions: How did you survive your undergrad years?
What classes would you recommend?

Received responses are partitioned with respect to category and corresponding professor. The questions and responses are stored in the knowledge repository (KR) housed in a MySQL database. Studies show that people tend to use the same terms and phrases when requesting information [11]. Based on this premise

additional questions that would retrieve the same response were stored. Each question is indexed to a unique solution.

3.1.3 Vision

Avari uses computer vision to determine the presence or absence of a user. A camera is mounted on her desk, and her field of vision is marked by a blue mat on the floor directly in front of her. We used the Java Media Framework [13] and Java Advanced Imaging [14] tools to perform background subtraction. Background subtraction compares color values from two images to identify changed pixels. The presence of a user is indicated by a threshold of thirty percent pixels exceeding a discrepancy of fifteen color values. Avari uses the blue channel as represented by the blue mat placed in front of the camera. Access to the results of the vision program is performed every second. This affords Avari's acknowledgment of both the approach and departure of users. Detection of a user's presence prompts Avari to initiate a dialogue unit. Detection of a user's departure prompts Avari to end the current dialogue unit, terminate the conversation and return to an idle state.

3.2 Virtual Components

3.2.1 Physical Appearance and Personality

Avari's appearance, behavior, gestures, and speech are rendered through JavaScript calls to the Haptek plug-in for Internet Explorer. Her appearance is modeled via Haptek's People Putty and her gestures via Haptek's Figuremaker [15]. Avari utilizes text-to-speech (TTS) for communicating and her voice is synthesized using NextUp's NeoSpeech Kate16 voice [16]. Avari's idle audio behaviors including sneezing, coughing and singing required lipsyncing to the audio files. This synchronization is provided using HapPhonemeEdit [17]. Consistent with Avari's age, her virtual face wears sharply defined makeup. She cracks corny jokes, compliments passersby, and presents an amicable and outgoing personality. Avari's appearance and personality are specifically chosen to present a friendly and approachable virtual human.

3.2.2 Software Components

Avari's system runs in Internet Explorer 7. Static HTML pages produce the visual cues on one monitor, and Avari's speech and behavior are rendered through the Haptek plug-in for Internet Explorer. Hypertext Preprocessor (PHP) provides the communications between Avari's separate processes: database access, speech recognition, question interpretation, conversational flow control, timing, and vision. Additional technologies include: Speech Application Language Tags (SALT) for speech recognition [18], JavaScript and Active X for conversation flow, MySQL for the KR and Matlab for the vision.

3.2.3 Understanding Speech

Avari's interpretation of the user's question is based on an 'answers first' approach. Answers first focuses on the solutions/answers i.e., the stored knowledge. Once the knowledge base is identified, questions that would yield the stored solutions are retrieved and stored. Each question is indexed to a specific solution. Avari compares the recognized speech to the stored questions, upon matching the indexed solution is presented to the user. All questions and indexed solutions are stored in a MySQL database.

3.2.4 Error Recovery

Avari utilizes common speech interface error recovery practices. These practices include providing verbal and visual cues to the user upon misrecognition, non-recognition and the detection of silence errors. The cues presented are determined by the user's position in the dialogue unit and indicate possible words/phrases that would afford the user to continue along the dialogue unit. Following are examples of error recovery practiced in Avari:

- Misrecognition during the professor identification - for the first three misrecognitions, Avari displays a list of all available professors. If the user has not successfully recovered, Avari prompts the user to choose a different professor.
- Silence detection error - Avari provides up to three consecutive prompts for speech to the user. If unsuccessful, she assumes the user has walked away and terminates the dialogue unit and conversation.

4. USER STUDY

Avari was deployed in a public setting. She was placed in a hallway outside an undergraduate computer lab at UNCC for eleven weekdays to gather user interaction data. Avari was typically present in the hallway from 8:00 AM to 5:00 PM. All user interactions are logged into Avari's conversation log file. In order to meet the Internal Review Board (IRB) regulations, we were restricted in our data collection. We were unable to visually record users' interactions with Avari and also the reaction of users too intimidated to approach and/or interact with Avari. Only subjective analyses of the causes of the data can be performed.

4.1 Results

4.1.1 User conversations

Avari conducted 536 conversations. Approximately 57% of the conversations lasted less than 15 seconds as indicate in figure 4. Avari's initial greeting utilizes 17 seconds, thereby indicating that these users approached and left Avari without responding to her first question. This may suggest that approximately ½ of the user population were too intimidated to engage in conversations with Avari. Of the remaining 43%, the majority of these conversations consisted of a single dialogue unit.

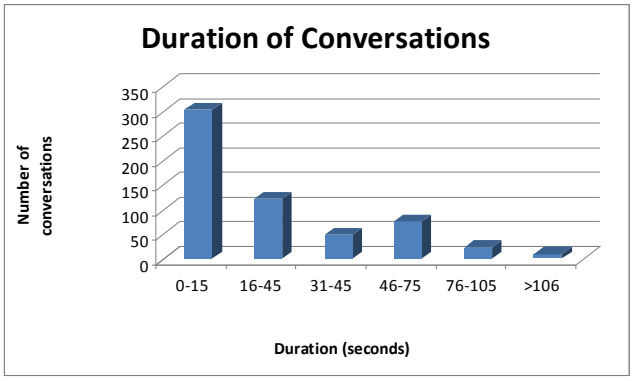

Figure 4: Duration of Conversations

Figure 5 shows the analysis of the category and content of each dialogue unit. There was no statistical significance in the popularity of the professors; however the 'Hobbies and Interests' category had the highest occurrence of 30% among its counterparts, with 'Advice for students' having the second highest occurrence of 25%. This suggests that the users engaged with Avari socially, acquiring subjective information about the professors over factual information.

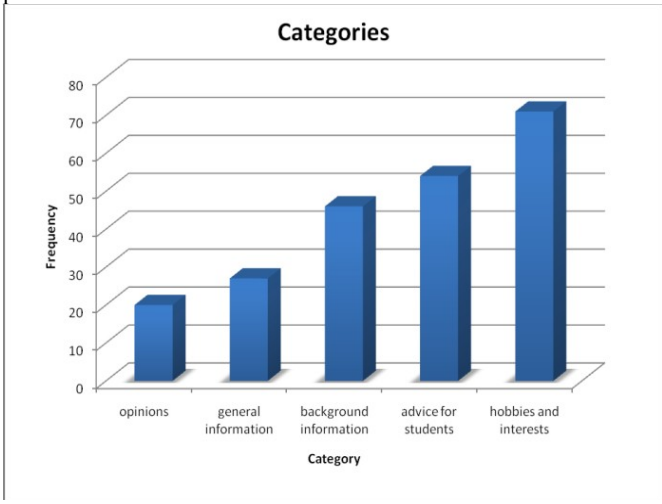

Figure 5: Category Popularity

4.1.2 User conversations

The duration between user conversations, idle time, had an average of 11.18 minutes. However, the data shows that on days of high user interaction, there was minimal idle time between user conversations. This suggests that the intimidation factor was reduced if the user witnessed another user interaction.

4.2 Difficulties in analysis

The main focus of this study is to investigate the approachability and effectiveness of a stand-alone virtual human. A previously undetected error surfaced in our software during the user study. This error caused Avari to cease communications during dialogue units. Avari would stop talking after the user asked a question, but before she responded to the user question. However, she would continue life-like expressions, winking, blinking, etc. When the user left her field of vision, she terminated the dialogue unit and concluded the conversation. This occurred in approximately 1/3 of user conversations. This software bug provided greater insight into users' perceptions of Avari. When users encountered this error, they continued to interact with Avari. They waited for a response for an average of 17.22 seconds. This suggests that users anticipated a response from Avari. They treated her as another human, exhibiting patience in waiting for her response. Anthropomorphism was applied to Avari allowing users to forgo typical expectations from a computer program, where users expect an immediate response. With respect to other people, users tend to be more patient in waiting for responses.

5. FUTURE WORK

We intend to deploy Avari in several settings to increase the diversity of users. Additionally, we would like to investigate the effects of Avari's appearance on user interaction. Changes to her appearance would include altering traits such as gender, race, age as well as changing the appearance of her rendered face and changing her physical clothing. Changes to Avari's disposition and personality will also be investigated.

6. ACKNOWLEDGEMENTS

The authors would like to thank Larry F. Hodges, Richard Souvenir, Toni Bloodworth, Amy Ulinski, and Louis Fletcher for their help with this project.

This work was supported in part by following NSF grants:
NSF-CNS 0540523 BPC-A: The STARS Alliance: A Southeastern Partnership for Diverse Participation in Computing (2006-2009)
NSF-CCF 0552631 REU Site: Computing Research for Undergraduates: Visualization, Virtual Environments, Gaming, and Networking (2006-2009)

7. REFERENCES

[1] Babu, S., Schmugge, S., Barnes, T., Hodges, L.F. 2006. "What Would You Like to Talk About?" An Evaluation of Social Conversations with a Virtual Receptionist. In Springer Lecture Notes on Artificial Intelligence LNAI, T. Gratch et al., Eds. Springer-Verlag.

[2] Babu, S., Schmugge, S., Inugala, R., Rao, S., Barnes, T., and Hodges, L.F. 2005. Marve: a prototype virtual human interface framework for studying human-virtual human interaction. In *Lecture Notes in Computer Science*, T. Panayiotopoulos, J.Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, Eds. Springer-Verlag, London, 120-133

[3] Cassell, J. 2000. Embodied conversational interface agents. *Commun. ACM 43*, 4 (Apr. 2000), 70-78.

[4] Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjalmsson, H. 2002. MACK: Media lab Autonomous Conversational Kiosk. In *Proceedings of IMAGINA '02*, (Monte Carlo, January 12-15, 2002), IMAGINA '02.

[5] Gockley, R., Bruce, A., Forlizzi, J., Moichalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A.C., Wang, J., 2005. Designing robots for long term social interaction. In *Proceedings of Intelligent Robots and Systems, 2006.IROS '05*. 1338-1343.

[6] Mehrabian, A., Friar, J., 1969. Encoding of attitude by a seated communicator via posture and position cues. Journal of Consulting and Clinical Psychology 33, 330-336.

[7] Raij, A., Johnsen, K., Dickerson, R., Lok, B., Cohen, M., Stevens, A., Bernard, T., Oxendine, C., Wagner, P., Lind, D. S. 2006. Interpersonal scenarios: virtual § real? In Proceedings of IEEE Virtual Reality 2006 (Alexandria, VA).

[8] Reeves, B., and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.

[9] Slater, M., Sadagic, M., Usoh, M., and Shroder, R. 2000. Small group behavior in a virtual and real environment: a comparative study. *Presence: Teleoperators and Virtual Envronments 9*, 37-51.

[10] Thorrison, K. R. 1997. Gandalf: an embodied humanoid capable of real-time multimodal dialogue with people. In *Proceedings of the First international Conference on Autonomous Agents* (Marina del Rey, California, United States, February 05 - 08, 1997). AGENTS '97. ACM Press, New York, NY, 536-537.

[11] Wilson, D. 2006. iTech: An Interactive Technical Assistant. PhD Dissertation, Auburn University.

[12] Zanbaka, C. A., Ulinksi, A. C., Goolkasian, P., and Hodges, L.F. 2007. Social responses to virtual humans: implications for future interface design. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM Press, New York, NY, 1561-1570.

[13] "Java Media Framework". Sun Developer Network. Sun Microsystems 12 Dec. 2008. <http://java.sun.com/javase/technologies/desktop/media/jmf/>.

[14] "Java Advanced Imaging". Java.net. Sun Microsystems. 12 Dec. 2008. <http://jai.dev.jav.net/>.

[15] "Haptek Figure Maker". Haptek Developers. Haptek Inc. 12 Dec. 2008. <http://haptek.biz/developers/figuremaker/>.

[16] "NextUp.com-NeoSpeech Kate16 Voice". Software Informer. NextUp.com. 12 Dec. 2008. <http://nextup-com-neospeech-kate16-voice.software.informer.com/>.

[17] "Haptek Phoneme Editor". Haptek Developers. Haptek, Inc. 12 Dec. 2008. <http://haptek.biz/developers/phonemeeditor/>.

[18] "Speech Application Language Tags (SALT)". Microsoft Developer Network. Microsoft Corporation. 12 Dec. 2008. <http://msdn.microsoft.com/en-us/library/ms994629.aspx>.