Recorded Speech, Virtual Environments, and the Effectiveness of Embodied Conversational Agents

Ivan Gris, David Novick,
Adriana Camacho, Diego A. Rivera, Mario Gutierrez, Alex Rayon

Department of Computer Science, The University of Texas at El Paso
500 W. University Ave., El Paso, TX 79912 USA
ivangris4@gmail.com, novick@utep.edu,
{accamacho2, darivera2, mgutierrez19, amrayon2}@ miners.utep.edu

Abstract. Development of embodied conversational agents (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. For a study of human-ECA rapport, we iteratively developed three versions of a game featuring an ECA, where each version of the game had a different combination of speech generation and virtual environment. Evaluations of the users' interactions with the different versions of the game enabled us to assess the effects of changes in the agent's voice and of changes in the agent's virtual world.

Keywords. Embodied conversational agents, virtual environments, human-agent dialog

1 Introduction

In this paper, we examine the effect on users' interaction with an embodied conversational agent (ECA) resulting from changes in (1) the ECA's virtual world and (2) the ECA's voice. Specifically, our study compares three versions of an immersive video game in which the user interacts with a life-size ECA. The game we developed is an adventure game, inspired by text-based games such as Zork (Anderson & Galley, 1985) and Colossal Cave (Crowther, Woods & Black, 1976), where the user tries to escape from the castle of an evil vampire king.

The human-ECA interaction reported in this paper took place in a spoken-language adventure game entitled "Escape from the Castle of the Vampire King." The game had a graphical interface with a full-sized ECA that served as the game's narrator, and the player controlled the game through speech commands. The game comprised 26 different rooms, each with its own secret passages, exits, items and clues. Players' interactions occurred in 20-minute sessions on two consecutive days, for a total of approximately 40 minutes per participant. The purpose of the game was to support a study of extroversion-based rapport-building behaviors over time (see Novick & Gris, in press).

2 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.

3 Increasing Engagement

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 system, users interacted with first-person perspective rather than the third-person perspective associated with avatars (see Serrels, 2011).

In our experiments across the three versions of the game, we measured gaze away from or toward the agent. We hypothesized that (a) reducing cognitive load would increase the proportion of time that users directed their gaze toward the agent and (b) placing the agent in a virtual world related to the game's story would also increase the proportion of time that users direct their gaze towards the agent.

4 Versions of the Game

Before presenting our results, we review the three versions of the "Escape from the Castle of the Vampire King" game with their respective voices and environments.

- **Version 1**. In the first version of the game, players were given two sheets of paper, one with a printed set of commands and their respective examples and a second with a template for drawing a map to mark the player's progress. This version used the default speech synthesizer provided by the Windows operating system.
- **Version 2**. The second version of the game featured a quick-reference command list in the upper-left corner of the projection and a dynamic map behind the agent. This version of the game used the IVONA Salli speech synthesizer.
- **Version 3.** The third version of the game included 3D scenery, recorded speech, agent movement based on motion-capture, and the quick-reference commands and incremental map display.

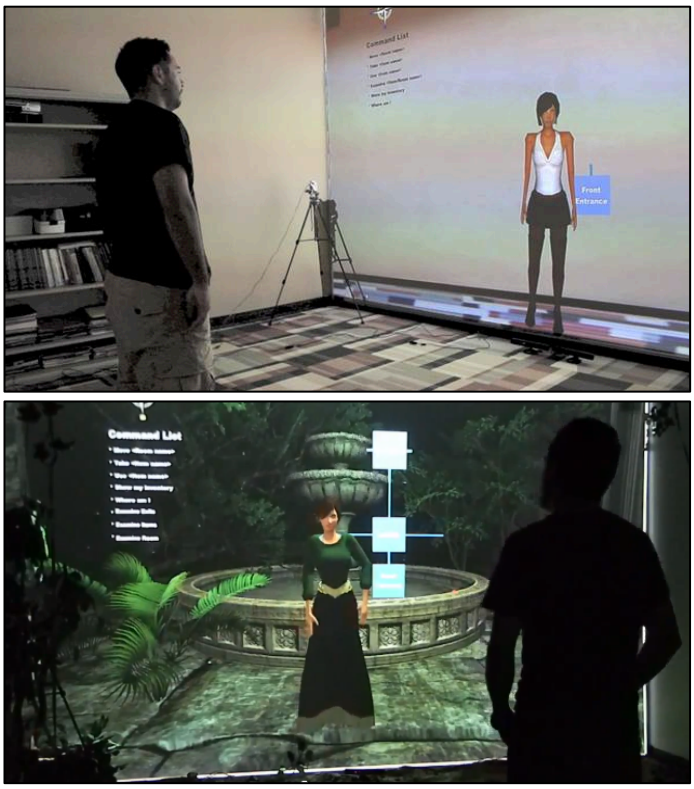

Fig. 1. Top: Version 2 of the game, with quick-reference commands, visible map, and Salli voice. Bottom: Version 3 of the game, with game scene, quick-reference commands, visible map, and recorded voice.

5 Results

In Section 2, we hypothesized that increasing the voice naturalness would make users less likely to interrupt the agent. The data, shown in Table 1, showed a significant ($\chi^2 < 10^{-6}$) reduction in the relative number of times that users interrupted the agent. In

Section 3, we hypothesized that reducing cognitive load would increase the proportion of time that users directed their gaze toward the agent. The data, also shown in Table 1, showed a significant ($\chi^2 < 10^{-6}$) increase in the proportion of time that the users gazed at the agents. We also hypothesized that placing the agent in a virtual world related to the game's story would increase the proportion of time that users directed their gaze towards the agent. The data showed a significant ($\chi^2 < 10^{-6}$) increase in the proportion of time that the users gazed at the agent.

Table 1. Data for analysis of hypotheses.

	Version 1	Version 2	Version 3
User interruptions of agent	27	60	14
Agent utterances	871	878	948
Gaze at agent (seconds)	1453.6	3759.9	9603.5
Gaze away (seconds)	5793.8	3616.3	1444.1

Table 2. Gaze shifts away from the agent.

	Gaze Shifts Away	Total Time (Seconds)	Average (Seconds/Shift)
V2	742	7376.2	9.94
V3	272	11047.6	40.62

This study was subject to three key limitations. First, the number of subjects varied across the three versions of the game. Second, as similar codings have consistently had high Kappas, we did not calculate interrater reliability. And, third, because the project involved iterative improvements to a system intended for a study of human-agent rapport, the study did not isolate the changes in voice and environment.

References

1. Anderson, T., Galley, S.: The history of Zork. *The New Zork Times*, 4(1-3) (1985).
2. Crowther, W., Woods, D., Black, K.: Colossal cave adventure. *Computer Game* (1976)
3. Newell, C., & Edwards, A. (2008). Place, authenticity time: a framework for synthetic voice acting. *International Journal of Performance Arts and Digital Media*, *4*(2-3), 155-180.
4. Novick, D., & Gris, I. (in press). Building rapport between human and ECA: A pilot study. *HCI International 2014*.
5. Schröder, M. (2001, September). Emotional speech synthesis: a review. In *Inter-Speech 2001* (pp. 561-564).
6. Serrels, M. (2011, April 18). *In Real Life: First Or Third Person - What's Your Perspective?* Retrieved from Kotaku: http://www.kotaku.com.au/2011/04/first-or-third-person-whats-your-perspective/.