The "Vampire King" (Version 2) Corpus

Ivan Gris, David Novick, Mario Gutierrez, Diego A. Rivera
The University of Texas at El Paso
500 W. University Ave., El Paso, TX 79912 USA
E-mail: ivangris4@gmail.com, novick@utep.edu, mgutierrez19@miners.utep.edu, darivera2@miners.utep.edu

Abstract

As part of a study examining nonverbal and paralinguistic behaviors in conversations between humans and embodied conversational agents (ECAs), we collected a corpus of human subjects interacting with an ECA in an adventure game. In the interaction, the ECA served as a narrator for a game entitled "Escape from the Castle of the Vampire King," which was inspired by text-based computer games such as Zork. The corpus described here is based on Version 2 of the game, in which a map of the castle was displayed on the wall behind the ECA. The system was not a Wizard-of-Oz simulation; the system responded using speech recognition and utterance generation. The corpus includes 20 subjects, each of whom interacted with the game for 30-minute sessions on two consecutive days, for a total of approximately 1200 minutes of interaction. All 40 sessions were both audiovisually recorded and automatically annotated for speech and basic posture using a Kinect sensor. The corpus includes (a) the automated annotations for speech and posture and (b) manual annotations for gaze, nods and interrupts.

Keywords: multimodal, real conversation, rapport

1. Introduction

This paper reports the collection of a corpus of interactions between humans and an embodied conversational agent (ECA). We developed the corpus to support a study of human-ECA rapport.

One of the main goals of researchers on real-time interaction with ECAs is to strive for increased realism in agents' behavior. One issues is maintaining and adapting to long-term interaction, particularly with respect to rapport. In our view (Novick & Gris, in press), paralinguistic rapport comprises three dimensions: a sense of emotional connection, a sense of mutual understanding, and a sense of physical connection. Because our research focuses on the physical dimension, the corpus was aimed at understanding the results of using an agent with different nonverbal behaviors (familiar and non-familiar). Studies of human-human dialog have suggested that people signal increased familiarity by, among other things, increasing the amplitude of nonverbal communicative behaviors such as hand gestures and head nods (Neff et al., 2010; Cafaro et al., 2012; Clausen-Bruun, Ek, & Haake, 2013). Thus in our system the agent communicated increased familiarity by increasing the amplitude of its gestures.

Because our underlying research on the development of human-ECA rapport depended on having subjects engage in multiple sessions over time, we needed to provide an interaction experience that was highly engaging; participants should want to return for later sessions. Toward this end, we developed an adventure game based on text games such as Zork (Anderson & Galley, 1985) or Colossal Cave (Crowther, Woods & Black, 1976) that follows the same gameplay format. In our game, the user tries to escape from the castle of an evil vampire king. In the place of the traditional console text interface, though, our agent narrated the game situation to the user. Players verbally indicated how they want to react to the presented situation, either by moving to different rooms, picking up items, using items, examining and uncovering secret passages or fighting the vampires.

In this paper we describe our automated and semi-automated annotation systems and our corpus-collection techniques. We provide examples of both verbal and nonverbal annotations generated by our system as well as how we use this functionality to recreate past interactions. We briefly describe the differences between familiar and non-familiar agent behaviors. Finally, we conclude with a discussion of limitations of our study and avenues for future research and applications of our corpus and corpus annotation systems.

2. The Computer Game

The human-ECA interaction took place in the form of a spoken-language version of a text-based adventure game entitled "Escape from the Castle of the Vampire King." The player controlled the game through speech commands rather than through text input, and the game had a graphical interface with a full-sized embodied conversational agent that served as the game's narrator. As the player explored the world of the game, the interface projected on the wall incrementally displayed a map showing where the player had been.

The game was written to appear as natural as possible. The game dialogs spoken by the agent were scripted in such a way so that even with limited commands, users could answer with natural language restricted to a specific context. Table 1 presents an excerpt of a player's interaction, showing the simple form of the player's verbal production.

Agent:	You enter the courtyard. A fountain rests in the room's center. A chapel lies in the distant west and the lobby to the south.
Player:	Examine the fountain.
Agent:	The water is green with algae and unsafe to drink. You notice a sign that reads: No magic allowed.
Player:	Use the magic potion.
Agent:	You dump the magic potion into the overgrown fountain. The vines within the fountain begin to smoke and hiss, finally burning away until you can see a dirty red gem lying over the drain in the fountain's base.
Player:	Pick up the red gem.
Agent:	You pick up a dirty red gem that looks like an eye.
Player:	Go to the chapel.

Table 1. An interaction transcript from the first session.

The vampire game comprises 26 different rooms, each with its own items, secret passages, points of interest, descriptions, and vampires. The agent is voiced by a text-to-speech engine that responds to several versions of four available commands (e.g., take the potion, pick up the potion, grab the potion). The commands are *move*, *take*, *use*, and *examine*; these commands can be applied to locations or items. For its part, the agent can respond to misunderstandings or unknown commands in five different ways.

3. Corpus Collection

We developed a first version of the system before the version from which the corpus reported here was collected. With Version 1, players were given two sheets, 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. We found that in Version 1, players would concentrate their gaze on the sheets rather than on the agent. For the rapport study to be effective, we needed the players to be looking at the agent so that the players would perceive differences in the agent's behaviors, our independent variable. So to immerse the players and fix their gaze towards the agent, we developed Version 2 of the game, which featured a small help box in the upper-left corner of the projection and a map displayed behind the agent that was automatically updated as the user progressed through the game. This also reduced the cognitive load required to play the game, as memorizing every place that players visited and every item they carried at any point in time would make the game impractical and effectively unplayable.

The game play took place in the Immersion Lab of UTEP's Interactive Systems Group. A full-body life-sized ECA was projected on a wall, roughly 18 feet diagonal, with a displayed background that resembles other walls of the Immersion Lab, which we intended to suggest that both the player and the agent were co-located in the same physical space. Figure 1 shows one of the authors conversing with the ECA during a game.

In each session the agent displayed nonverbal behaviors that reflected the study's independent variable of familiarity vs. non-familiarity. Although it is possible to slowly transition from the non-familiar to the familiar animations in a single session, we opted to include only

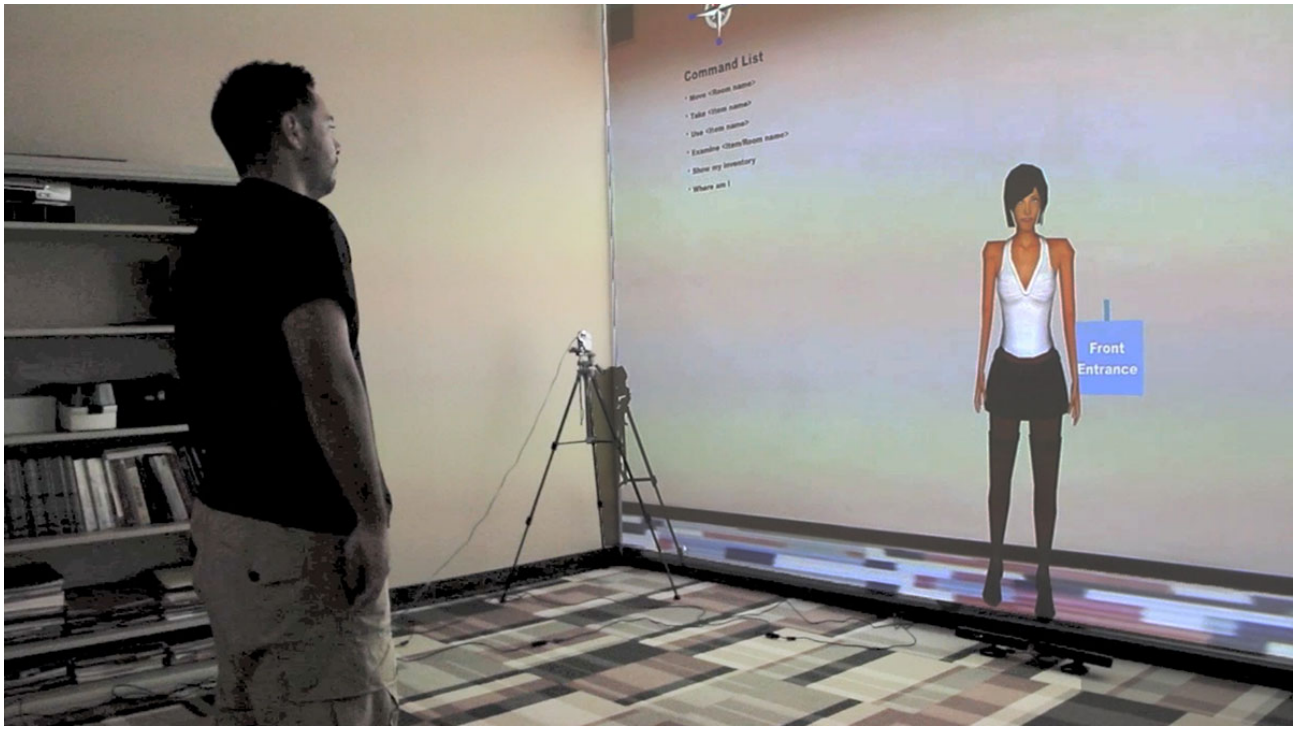

Figure 1. A conversation in the Immersion Lab between a player and the ECA

one type of behavior (i.e., low or high amplitude) per session to make a clear distinction between them and to ensure that subjects find differences in each behavior. The initial conversations exhibited non-familiar behaviors (low amplitude). The second sessions alternated between the behaviors (half with non-familiar and half familiar).

We recruited 20 undergraduate students to play with the agent over two days, in thirty-minute sessions; the subjects were assigned randomly to the familiar or non-familiar condition in the second session. We recorded both video and audio in each session from two angles, one from a Microsoft Kinect and one from a regular digital camcorder. The Kinect recorded the locations and angles of twenty user joints (see Figure 2). A normal stance and crossed arms were automatically detected and annotated on the log file; however the agent did not react to any position. We tested and recorded a total of 40 conversations, two for each of the 20 participants.

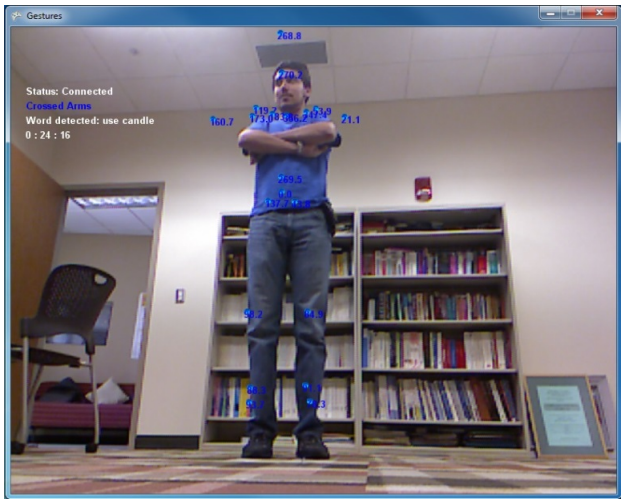

Figure 2. Pose recognizer detecting crossed arms

4. Annotations

Each episode was automatically annotated using two different methods, a game-save file and a log file.

With the game-save-file method, subjects were asked to save their game after their first session so that the agent would remember their previous interactions when they returned for the second session. These save files contain a list of all the valid interactions that led to a state change in the game; these valid interactions are immediately and silently recreated when the game is loaded.

The log-file method created a log file with a time stamp. These log files contain the current pose (normal stance or arms crossed) and what was understood by the agent. The log file was updated after every utterance that was heard by the ECA. Figure 3 presents examples of both a log file and a save file.

5. Limitations

Pose annotations were limited, and while the ECA logged them it did not react to particular positions. Because the Kinects were visible to the players, most players were aware that their pose might be recognized, and some even consciously attempted to make the agent react to their movements. In addition, the game task required the players to remember a considerable amount of game information, even when we displayed the map. Consequently, there were extended periods of silence or inactivity while players attempted to recall something.

Finally, the physical position of the Kinect sensor was not optimal. Because the wall served as a projection screen, the Kinect had to be placed on the floor close to the wall. The Kinect does not have high-resolution cameras, so images at this distance were difficult to analyze. In particular, even though we dedicated one of the Kinects specifically to the subjects' facial expressions, we failed at effectively recording and annotating facial gestures. Figure 4 shows the Kinect tracking the face of a person playing the game in the Immersion Lab.

6. Future Work

We expect to improve and expand the system by using annotations from unrecognized arbitrary poses to create new detectors. In particular, we want to collect additional data from the pose detector. As it is, we can recognize and annotate particular poses with their timestamp; however, creating the detection for these poses is a lengthy process. Figure 5 shows a manually coded detector of frustration gestures based on our corpus. It includes the angles between joints per participants and

```
1  Log File                                                                                                    | Save File
2  ------------------------------------------------------------------------------------------------------------------------
3  Time: 0 : 23 : 850 -- Word recognized: load game player one   -- Pose: No pose detected    | load game
4  Time: 0 : 33 : 816 -- Word recognized: take rusty sword       -- Pose: Normal Stance       | take rusty sword
5  Time: 0 : 43 : 416 -- Word recognized: go to the dinning hall -- Pose: Normal Stance       | move dinning hall
6  Time: 0 : 48 : 916 -- Word recognized: go to the clock tower  -- Pose: Normal Stance       | move clock tower
7  Time: 0 : 53 : 933 -- Word recognized: go to the courtyard    -- Pose: No pose detected    | move courtyard
8  Time: 1 : 7  : 200 -- Word recognized: use rusty sword        -- Pose: No pose detected    | use rusty sword
9  Time: 1 : 31 : 466 -- Word recognized: go to the library      -- Pose: Normal Stance       | move library
10 Time: 1 : 44 : 200 -- Word recognized: use holy water         -- Pose: Normal Stance       | use holy water
11 Time: 1 : 58 : 783 -- Word recognized: pick up ancient book   -- Pose: Normal Stance       | take ancient book
```

Figure 3. Example of a log file (left) and save file (right)

several statistical measures to calculate efficient margins of error. The next step is to collect the information related to the angles between joints and create new poses from them. We also hope to improve the illumination, camera, microphone and sensor location, and file compression to attain portable, high quality media that automatically provides additional information to improve the behavior of our agents in real time.

A corpus for Version 3 of the Escape from the Castle of the Vampire King game will be forthcoming. The new corpus will differ primarily with respect to improved game-play, including using recorded speech for the ECA and having backgrounds that represent the world of the game rather than the virtual reality of the Immersion Lab. For the longer run, we are building a new game, based on a jungle survival scenario, that is designed to support a more conversational style of dialog, advanced gesture recognition, longer-term interaction, and, at least to a limited extent, the mutual-understanding dimension of rapport.

7. Acknowledgments

The authors thank Guillaume Adoneth and David Manuel for their contributions to the design of this study, Jonathan Daggerhart for permission to adapt his original text-based adventure game into "Escape from the Castle of the Vampire King," and Alex Rayon, Adriana Camacho, Baltazar Santaella, Juan Vicario, Joel Quintana and Anuar Jauregui for their help in developing the game.

8. References

Anderson, T., Galley, S.: The history of Zork. *The New Zork Times*, 4(1-3) (1985).

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., Valgarðsson, G. S.: First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Intelligent Virtual Agents* (pp. 67-80). Springer Berlin Heidelberg (2012).

Clausen-Bruun, M., Ek, T., Haake, M.: Size certainly matters–at least if you are a gesticulating digital character: The impact of gesture amplitude on addressees' information uptake. In *Intelligent Virtual Agents* (pp. 446-447), Springer Berlin Heidelberg (2013).

Crowther, W., Woods, D., Black, K.: Colossal cave adventure. *Computer Game* (1976).

Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In *Intelligent Virtual Agents* (pp. 222-235). Springer Berlin Heidelberg (2010).

Novick, D., Gris, I.: Building rapport between human and ECA: A pilot study. In *Proceedings of HCI International 2014* (in press).

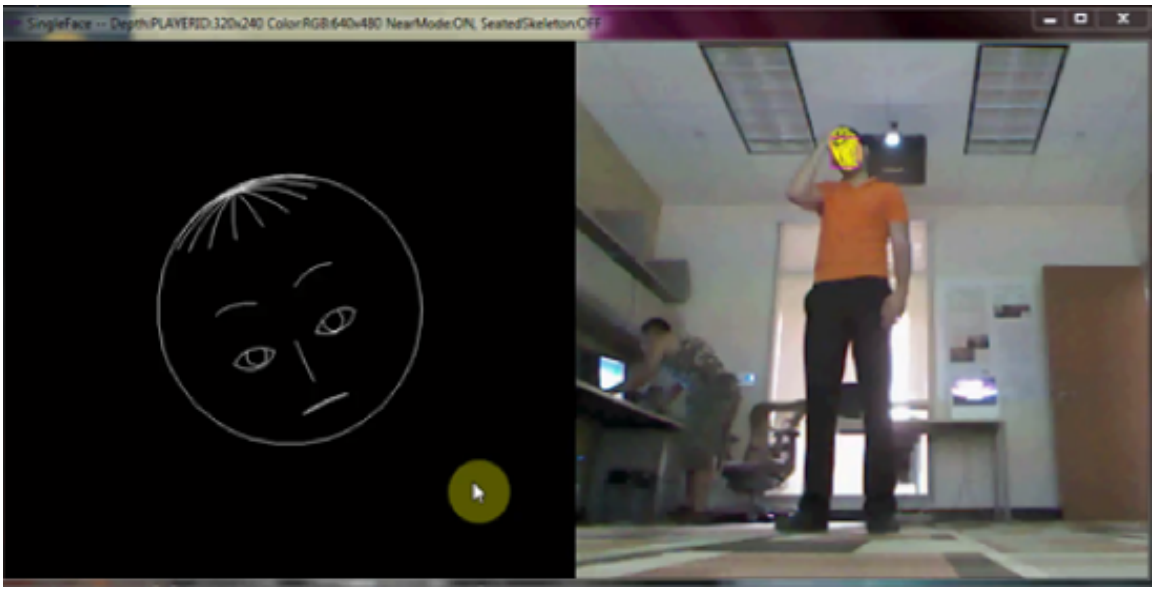

Figure 4. Facial gesture recognition

Frustration 2	P1	P2	P3	P4	P5	P6	MAX	MIN	AVG	DIF	DIF2	DIF3	RANGE
Shoulder Center - Shoulder Left	52	35	36	33	43	6	65	33	44.0	21.0	11.0	32	10.0
Shoulder Left - Elbow Left	338		135	316		32	338	135	277.8	60.3	142.8	203	-82.5
Elbow Left - Wrist Left	213	213	211	219	250	23	250	211	224.0	26.0	13.0	39	13.0
Wrist Left - Hand Left	260	251	241	225	249	26	268	225	249.0	19.0	24.0	43	-5.0
Shoulder Center - Shoulder Right	146	152	141	149	157	15	157	141	149.7	7.3	8.7	16	-1.3
Shoulder Right - Elbow Right	112	95	98	112	103	11	118	95	106.3	11.7	11.3	23	0.3
Elbow Right - Wrist Right	88	92	92	102	89		102	88	92.6	9.4	4.6	14	4.8
Wrist Right - Hand Right	143	93	96	102	108		143	93	108.4	34.6	15.4	50	19.2

Figure 5. Coding of frustration gestures