Extending Empirical Analysis of Usability and Playability to Multimodal Computer Games

David Novick(✉) and Laura M. Rodriguez

Department of Computer Science, The University of Texas at El Paso,
500 West University Avenue, El Paso, TX 79968-0518, USA
novick@utep.edu, lmrodriguez3@miners.utep.edu

Abstract. The published research examining usability and playability of games is largely theoretical. A prior empirical study of a game with an embodied conversational agent found that most frustration episodes could be understood in terms of both usability and playability, but this study was based on a game in which the interaction by both player and agent were limited to verbal communication. To explore whether these results would hold for a game in which the player and agent communicated with both speech and gesture, we conducted an empirical formative user-experience evaluation of a multimodal game. Our findings strongly confirmed that frustration episodes can be understood as issues of both usability and playability. However, the relative frequencies of the categories of usability and playability issues differed between the speech-only and the speech-and-gesture games. Much of this difference likely arose because higher levels of engagement and rapport between player and agent in the speech-and-gesture game led to the players having greater, and in many cases unfulfilled, expectations for the capabilities of the agent.

Keywords: Embodied conversational agent · User experience · Playability · Usability · Gesture · Engagement · Rapport

1 Introduction

User-experience evaluation typically assesses the effectiveness, efficiency, and user-satisfaction of a system, given its goals [1]. For most user interfaces, the purpose of usability testing is to make the application as easy to use as possible. But for games, the very point of the application is that it not be easy: the application should present the user with interesting challenges. For office applications, user-experience goals center on factors such as task completion, error elimination, and workload reduction; for games, user-experience goals center on factors such as entertainment, the fun of overcoming obstacles, and workload increase [2]. For this reason, the relationship between playability and usability has remained problematic. And as embodied conversational agents (ECAs) [3] become increasingly ubiquitous, developers of ECA-based games, adventures, and other experiences could benefit from an understanding of evaluation methodologies that more clearly explains the relationship between usability and playability.

A. Marcus (Ed.): DUXU 2016, Part I, LNCS 9746, pp. 469–478, 2016.
DOI: 10.1007/978-3-319-40409-7_44

The constructs of usability and playability have been explored with respect to games, but primarily on a theoretical basis. The only known published application of formative empirical evaluation of a computer game evaluated the usability and playability of an ECA-based adventure game [4]. In that study, however, users navigated the application only through relatively simple speech commands. The application's user interface did not enable users to use natural speech or physical gestures as modalities of interaction. In the present study, we address this limitation by reporting on the formative user-experience evaluation of a multimodal ECA-based adventure game in which the player could communicate with the on-screen agent through more natural, unstructured utterances and through upper-body gestures.

In this paper, then, we briefly review the state of the art of user-experience evaluation of video game, explain the challenge for analysis of the relationship between usability and playability, describe the study's methodology, report the results of the formative evaluation, and discuss the implications of these results. We conclude with a contrast of these results with those in [4] and discuss the limitations of our study.

2 Background

Studies of evaluation of the playability and usability video games have, with very limited exceptions, not extended to empirical evaluation of games through user studies. Rather, the research literature of playability and usability has tended to focus on heuristic evaluation of games (e.g., [5–7]), and the research literature on user-centered evaluation of games has tended to remain at the theoretical level (e.g., [1, 8, 9]). In contrast, commercial practice appears to rely largely on empirical usability tests, with little use of heuristic evaluation [10]; unfortunately, the results of the commercial studies tend not to be published. At the same time, the relationship between playability and usability remains unclear. There are evident differences between playability and usability [1, 2]. For example, an office application seeks to make the user's task as easy as possible, while a game seeks to make the user's task interestingly difficult. But from the standpoint of empirical testing of games, the theoretical divisions between playability and usability appear to diminish in practice.

An initial study that assessed playability and usability of a computer game through empirical user testing suggested that the evaluation technique for playability can be the same as for usability: there is really a single technique of empirical testing of the user's experience in computer games, regardless of whether this is called usability testing or playability testing [4]. This paper remains the only published empirical study of user-experience testing of computer games of which we are aware. However, the game that was tested in that study, although it included animated visuals, was entirely verbal in its interaction. The game, called "Escape from the Castle of the Vampire King," although implemented as an immersive game with an ECA, was essentially an immersive version of a text-based adventure game such as Zork [11]. Do these results hold for an immersive game that goes beyond the limitations of verbal interaction to combine verbal interaction with gestural interaction and movement in the virtual world by the human and the agent?

3 Methodology

To address the question of whether the results for relatively simple speech-based interaction hold for more complex, multimodal interaction, we studied the development of an immersive computer game, entitled "Survival on Jungle Island" [12], in which the human user interacted with the ECA through both speech and gesture as the human and the agent moved through the game's virtual world. A brief video of excerpts of the game is available at http://www.cs.utep.edu/novick/jungle.mp4. In the game, each participant partners with the agent to survive on and escape from a deserted jungle island.

The study took place in the Immersion Lab at the University of Texas at El Paso (UTEP). This lab consists of a rectangular room about 14 feet by 17 feet, with the image projected on one of the shorter walls, which is covered with a reflective paint to serve as a screen; the wall is fully covered by the projection. For the jungle game, the lab set-up included artificial plants and other scenery elements that helped make the setting more immersive. The artificial plants were spread out between the player and the projected scene. A camera, placed behind a tree at the lower left corner of the projection wall, recorded the user's experience. The system recognized the player's speech and body gestures via a Microsoft Kinect, which was placed at the bottom center of the projection wall. The person running the experiments was behind the projector at the control table. Figure 1 shows the setup of the room in which experiments took place. Figure 2 shows the interior of lab, including scenery for the Jungle game, with the placement of the projector, Kinect, and camera.

The jungle game, created to study rapport between embodied conversational agents (ECAs) and humans, comprised a series of scenes in which a single user interacts with the ECA through speech and gesture, such as a high-five gesture. The research studied whether users would feel increased rapport with an ECA that elicits and perceives gestures than with one that elicits only speech; inclusion of gestures in the game was essential for achieving the aims of the research.

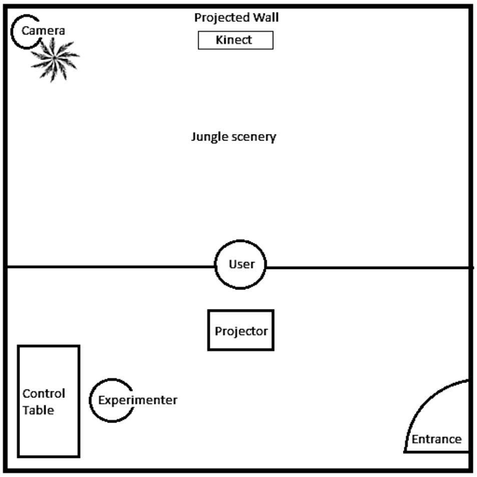

Fig. 1. Experimental setup in UTEP's immersion lab

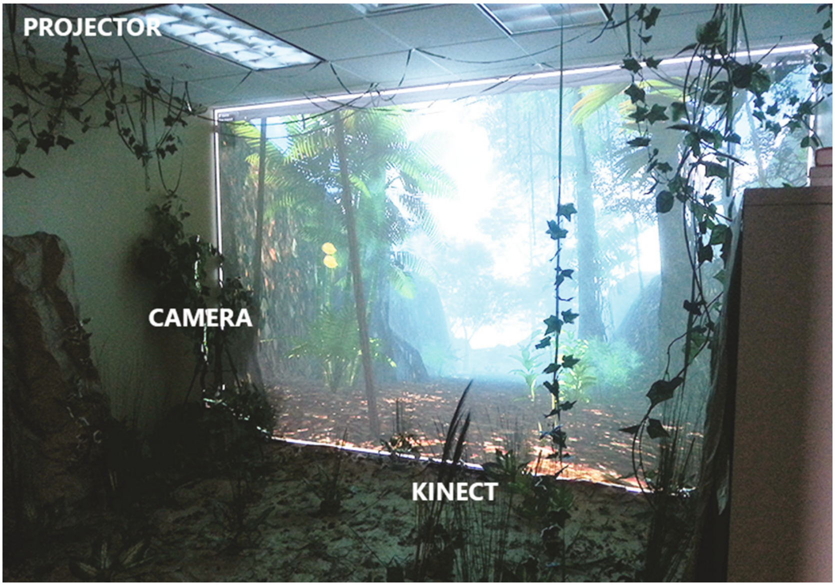

Fig. 2. Interior of the immersion lab, with scenery for the "Survival on Jungle Island" game

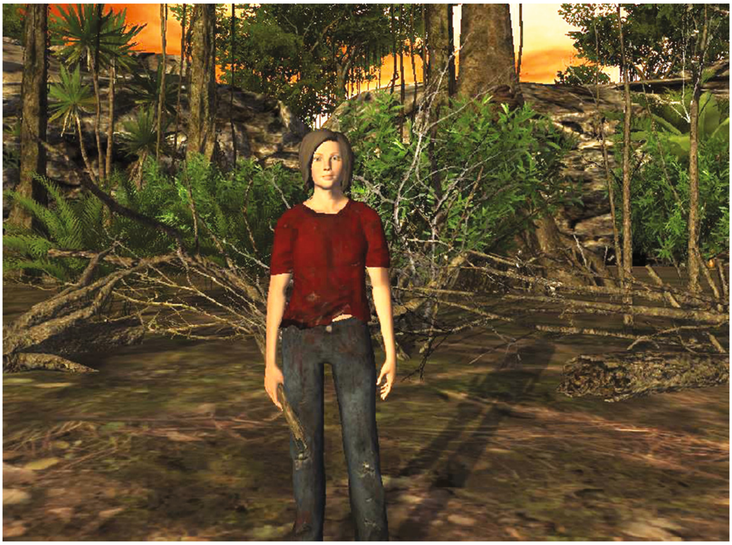

Fig. 3. Adriana, the embodied conversational agent in "Survival on Jungle Island"

A session with the Jungle game typically lasted 40–60 min, depending on the player. The players interacted with a life-sized ECA named Adriana who had been stranded in the island shortly before the player arrived. Figure 3. Depicts Adriana in one of the game's scenes. Adriana would partner with the player to survive on and escape from the deserted island. The participants were video-recorded throughout the game to make note of their reactions. At the game's conclusion, the participants completed a survey in which they communicated their experience with and perceptions of the agent during the game.

For the user-experience evaluation, an experimenter was present in the lab to note the details of the projection at the time of a frustration episode. The experimenter also could step in to troubleshoot, as the system was being evaluated formatively; because it was still in development, the system, on some infrequent occasions, froze.

The formative user evaluation of the Jungle game reported here involved four participants, three male, and one female. As the system uses speech and gesture recognition, it was not practical to use common usability approaches such as think-aloud techniques. We introduced each participant to the game and explained that the session would be recorded for research purposes. Then the experimenter started the game and observed the projection as the game progressed. We considered a frustration episode when the user tried to do something and the system did not respond as the user expected producing observable or expressed reactions of irritability or confusion. After the game, the experimenter asked the participant about specific frustration episodes so that the participant could explain the reasons for frustrations encountered.

4 Results

In the user tests, we observed a total of unique 44 frustration episodes; for the purposes of this analysis, subsequent identical episodes in the same session were not counted. The number of unique frustration episodes per user ranged from 6 to 14, with a mean frequency of 11.0 unique episodes per user. We coded each episode for playability with the six facets of playability developed in [1] (intrinsic, mechanical, interactive, artistic, personal, and social), and coded each episode for usability with the categories developed in [13] for heuristic evaluation (dialog, speaking the user's language, user's memory load, consistency, feedback, clearly marked exits, accelerators, error messages, error prevention, and other). Table 1 summarizes these results.

We found that all of the frustration episodes could be categorized as both a playability issue and a usability issue. For example, an episode where the player became unnerved by the facial expression of the agent as she said words containing the vowel "o" and was smiling because the agent's animation seemed creepy we coded for usability as speaking the user's language and for playability as artistic. An episode where the agent asked the player to not let her fall asleep, but there was no way for the player to keep the agent awake, we coded for usability as consistency and for playability as intrinsic. Table 2 presents four examples of frustration episodes, all in the "dialog" usability category but in three different playability categories.

Overall, of the 34 unique problems identified in the system through the user-experience testing, 18 involved speech production and recognition, 9 involved gesture production and recognition, and 7 involved other factors.

Table 1. Frustration episodes in the Survival in Jungle Island categorized using both usability and playability.

	Playability						
Usability	Intrinsic	Mechanical	Interactive	Artistic	Personal	Social	Total
Dialog	3		3		4		10
Speaking users' lang			14	1			14
User's memory load							0
Consistency	2	1	2	1			5
Feedback	1	2	10				13
Clearly marked exits							0
Accelerators							0
Error messages							0
Error Prevention							0
Other							0
Total	6	3	29	2	4	0	44

Table 2. Example issues and coding into usability and playability categories

Issue	Usability	Playability
Agent's utterance was too fast to understand, so participant did not catch the joke and was confused about why the agent was smiling.	Dialog	Interaction
Agent encouraged participant to not be so quiet in one dialog when in the rest of the game agent talks more than she listens.	Dialog	Intrinsic
Participant complained that agent "doesn't let me talk!"	Dialog	Personal
Participant asked agent about what happened and why she was on the island.	Dialog	Intrinsic

4.1 Frustration Episodes Involving Speech

We encountered several issues regarding the ECA's verbal communication. For example, the agent asked rhetorical questions and did not give the user enough information to know that it was rhetorical or provide time to respond. This upset the players, who felt ignored; some even verbally expressed their complaint to the agent, only to be ignored again, thus compounding the problem and making them even more upset. We coded these frustration episodes for usability as *dialog* and for playability as *personal*.

Players also encountered a problem in an early scene where the agent encouraged the user to talk more. Because the game's designers sought to encourage interaction with the user, they thought that this prompt from the agent would be appropriate. However, the agent's encouragement made the players overconfident about what the agent could understand. As a result, three of the four testers started asking the agent questions in the next scene, but the agent was not designed to answer spontaneous questions from the player; as a result, the agent ignored the players' questions. The testers explained in their post-session interviews that this upset them not only because answering questions was not within the agent's capabilities but because the agent had initially encouraged them to interact more. Because of these problems, most of the testers viewed the agent as a bossy extravert. We coded these frustration episodes for usability as *dialog* and for playability as *intrinsic*.

4.2 Frustration Episodes Involving Gesture

We now turn to playability and usability issues related to gesture. The hardest gestures to perform in the game were spearfishing and starting a fire, and the way these activities are performed depends on their context. Unfortunately, the early version of game evaluated in this study presented little physical context for these activities. Without information on what tool they were using for spearfishing, the testers produced varied gestural responses to the agent's prompts. One of the testers tried to catch fish with his hands.

Similar problems occurred when the agent asked the players to start a fire using two sticks. People who already know how to do this—and possibly people who have merely seen it done—are likely to come up with the way to move their hands to simulate the gesture. The system expected that the players would have both hands in front of their body, close to elbow level, with one hand at an angle striking the other in such a way that one hand lands on top of the other. But in practice, the testers' intuitive gestures for striking two sticks to start a fire in this study were hugely different from this model. One participant tried to strike the two sticks in front of her with her arms perpendicular to the floor. She then tried striking her hands at an angle in which her right hand was on top of the left while both of her hands would come together directly in front of her. Even though the motion was partially right, it was not being performed at the correct angle with respect to the elbows. The participant tried slight variations of this gesture until she got feedback. We coded this frustration episode for usability as *speaking the user's language* and for playability as *interactive*.

Another participant was striking his hands in front of him with a wide and swift arm motion. Each time that he did the gesture and did not get feedback, he would do the motion slower, until eventually he came to a complete stop. We coded this frustration episode for usability as *feedback* and for playability as *interactive*.

Another problem with gestures arose because objects or events to which the agent referred were not accurately depicted in the virtual world in the early version of the game being tested. This confused the testers and distracted them from their task at hand. We coded these frustration episodes for usability as *consistency* and for playability as *interactive*.

5 Conclusion

Our analysis confirmed the findings in [4] that frustration episodes can be viewed as both playability and usability problems. In the current study, all 44 of the episodes could be coded into one of the six facets of playability and could be coded into one of the categories of heuristic usability other than "other."

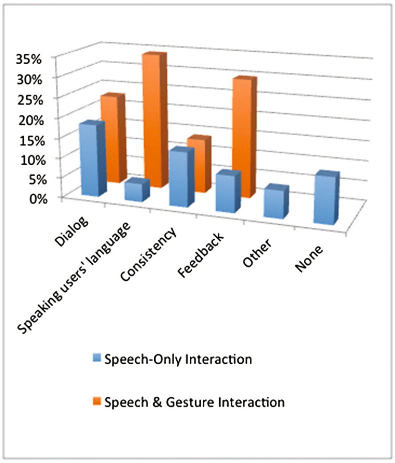

Fig. 4. Comparison with respect to relative frequency of usability issues of speech-only and speech-and-gesture games. (Color figure online)

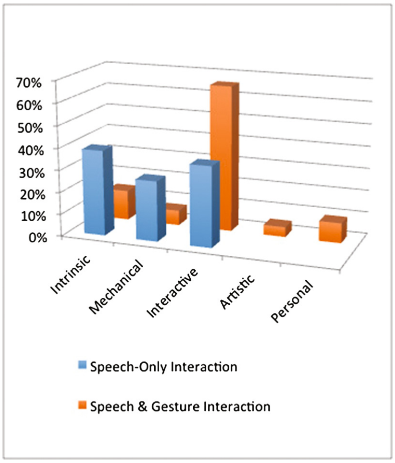

Fig. 5. Comparison with respect to relative frequency of playability issues of speech-only and speech-and-gesture games. (Color figure online)

In terms of playability, we classified 66 % of the frustration episodes as interactive, which is associated with player interaction and videogame user interface development. This contrasts with the results in [4], where only 36 % of the episodes were interactive. In terms of usability, the frustration episodes clustered in the categories of dialog, speaking the user's language, and feedback. This again contrasts with the results in [4], where the frustration episodes clustered in the categories of error prevention and dialog. Figures 4 and 5 contrast the respective playability and usability distributions between [4] and the present study.

This evidence suggests that the modality of games affects the categories of usability and playability issues that users experience. While it may seem paradoxical that the speech-only game had fewer problems with *dialog* and *speaking the user's language* than did the speech-and-gesture game, this effect likely resulted from differences in the way language was used in the two games. The speech-only game, "Escape from the Castle of the Vampire King," was effectively a text-based adventure game with animated pictures. Its speech recognition relied on the players producing highly structured utterances such as "go to lobby" and "pick up castle key," and its agent produced relatively simple utterances. In contrast, the speech-and-gesture game, "Survival on Jungle Island," enabled the players to communicate with unstructured utterances, and the agent's utterances were longer and more conversational. Consequently, the Jungle game had greater opportunity for players to explore the space of speech they produced.

A similar dynamic likely led to the differences in frequency of playability issues. In particular, the speech-only game, with its Zork-like restriction to structured utterances and connected rooms, presented players with greater challenges in their accomplishing goals, thus leading to a greater relative frequency of intrinsic playability issues. In contrast, the speech-and-gesture game, with its promise of apparently free-form spoken interaction, provided greater opportunity for players to reach the limits of the game's natural-language capabilities, thus leading to a relatively higher frequency of interactive usability issues. The Jungle game was not designed so that the agent could answer questions from players. Thus although the game did well making players feel comfortable with the agent, in some of the players this enhanced level of engagement and rapport tended to provoke a sense of curiosity. These users began asking questions of the agent, only to get frustrated at her perceived rudeness. This phenomenon did not appear to occur in the speech-only game, probably because the interaction led to a much lower level of engagement and rapport between player and agent.

6 Postscript

The formative user-experience evaluation reported in this paper enabled the development team to address both the usability and playability issues identified in the user testing. The "Survival on Jungle Island" game went on to receive the award for outstanding demonstration at the 2015 International Conference on Multimodal Interaction [14].

Acknowledgments. The authors acknowledge with gratitude the team, in addition to the authors, that developed Survival on Jungle Island: Ivan Gris, Adriana Camacho, Diego Rivera, Mario Gutierrez, Alex Rayon, Joel Quintana, Anuar Jauregui, Timothy Gonzales, Alfonso Peralta, Victoria Bravo, Jacqueline Brixey, Yahaira Reyes, Paola Gallardo, Chelsey Jurado, Guillaume Adoneth, David Manuel, Brynne Blaugrund, and Nick Farber.

References

1. Sánchez, J.G., Simarro, F.M., Zea, N.P., Vela, F.G.: Playability as extension of quality in use in video games. In: 2nd International Workshop on the Interplay between Usability Evaluation and Software Development (I-USED) (2009)

2. González Sánchez, J.L., Padilla Zea, N., Gutiérrez, F.L.: From usability to playability: introduction to player-centred video game development process. In: Kurosu, M. (ed.) HCD 2009. LNCS, vol. 5619, pp. 65–74. Springer, Heidelberg (2009)
3. Cassell, J.: Embodied Conversational Agents. MIT press, Cambridge (2000)
4. Novick, D., Vicario, J., Santaella, B., Gris, I.: Empirical analysis of playability vs. usability in a computer game. In: Marcus, A. (ed.) DUXU 2014, Part II. LNCS, vol. 8518, pp. 720–731. Springer, Heidelberg (2014)
5. Desurvire, H., Caplan, M., Toth, J.A.: Using heuristics to evaluate the playability of games. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems, pp. 1509–1512. ACM (2004)
6. Pinelle, D., Wong, N., Stach, T.: Heuristic evaluation for games: usability principles for video game design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1453–1462. ACM (2008)
7. Fierley, R., Engl, S.: User experience methods and games: lessons learned. In: proceedings of the 24th BCS Interaction Specialist Group Conference, pp. 204–210. British Computer Society, Swinton (2010)
8. Nacke, L.: From playability to a hierarchical game usability model. In: 2009 Conference on Future Play on@ GDC Canada, pp. 11–12 (2009)
9. Fabricatore, C., Nussbaum, M., Rosas, R.: Playability in action videogames: a qualitative design model. Hum. Comput. Interact. **17**(4), 311–368 (2002)
10. Fabricatore, C., Nussbaum, M., Rosas, R.: Playability in action videogames: a qualitative design model. Hum. Comput. Interact. **17**(4), 311–368 (2002)
11. Lebling, P.D., Blank, M.S., Anderson, T.A.: Special feature zork: a computerized fantasy simulation game. Computer **4**, 51–59 (1979)
12. Novick, D., Gris, I., Rivera, D.A., Camacho, A., Rayon, A., Gutierrez, M.: The UTEP AGENT System. In: Proceedings of the 17th ACM International Conference on Multimodal Interaction, Seattle, WA, 9–13 November 2015
13. Nielsen, J., Molich, R.: Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 249–256 (1990)
14. University Communications [of the Univ. Texas at El Paso]: UTEP Computer Science Department develops award-winning interactive agent system, 21 January 2016. http://engineering.utep.edu/announcement012116.htm