Suh J, Bennett CC, Weiss B, Yoon E, Jeong J, Chae Y (2021) “Development of Speech Dialogue Systems for Social AI in Cooperative Game Environments.” *IEEE Region 10 Symposium (TENSYMP)*. In Press.

Development of Speech Dialogue Systems for Social AI in Cooperative Game Environments

Jaeyoung Suh[1], Casey C. Bennett[1*], Benjamin Weiss[2], Eunseo Yoon[1], Jihong Jeong[1], Yejin Chae[1]

[1] Department of Intelligence Computing
Hanyang University
Seoul, Republic of Korea

[2] Quality and Usability Lab
Technische Universität Berlin
Berlin, Germany

***Abstract*— There is increasing interest in developing more human-like artificial intelligence (AI) capable of natural social interaction. Previous research has suggested ideas about what it means to be “life-like” AI, and some studies have attempted to test these hypotheses using game environments. In this paper, we introduce the development of the Speech Dialogue System for a “Social AI”, which communicates and interacts autonomously with a human player in cooperative game environments (in this case a social survival game called “Don’t Starve Together”). Based on our hypothesis that the AI should contain specific components to be perceived as more human-like, we conducted a series of pilot tests to develop the Social AI using a *data-driven approach*. After finishing the pilot tests, we identified six components to add or revise, based on participant interactions and feedback. These components mainly include features of the Speech Dialogue System that pertain to the interplay of AI behavior with contextual factors of the social environment (“the game state”). In future work, we intend to improve the Social AI based on these findings. The research here highlights the use of cooperative game environments for data-driven development of speech dialogue systems for artificial agents.**

Keywords*— *artificial intelligence; human-computer interaction; social cognition; social games; speech system; virtual avatar

I. Introduction

There is a growing interest in creating artificial intelligence (AI) that can emulate natural human social behavior to produce better interactive technology. There are a few open questions, however, around which components of social interaction are relevant to producing the social fluidity necessary for humans to perceive an interaction as “life-like”. Previous work has focused on understanding how such fluidity arises from the construct of social presence, a sense of being there with a “real person” in artificial environments [2]. Such previous research has shown that both behavioral factors related to the artificial agent itself as well as contextual factors beyond the agent (i.e. interaction context) play a critical role in how people perceive interactions with interactive technology [3]. These questions also tie back to Dennett’s work on intentional stance as it relates to attributions of agency in artificial agents, i.e. an agent that is perceived to have its own self-driven goals and intentions (averse to a machine) [4].

Here, we seek to explore the above open questions during interaction between a human and a “Social AI” in the form of a virtual avatar capable of autonomous speech based on its perceptions of the social environment. Our approach is modeled on previous research understanding basic principles of social interaction with robotic faces [5, 6]. In this paper, we describe current progress of making the Social AI to play video games with human players, specifically the “Don’t Starve Together” game (https://www.klei.com/games/dont-starve-together). The “Don’t Starve Together” game is a social survival game where players need to collect resources, make tools, fight monsters, and cooperate with each other to survive longer. As such, the game provides an ideal environment to experiment with interactive behavior during cooperative goal-oriented tasks [7].

We conducted a series of pilot studies using this game paradigm during interaction between human players and the Social AI as co-player (henceforth referred to as our “social environment”) similar to the previous work [1]. The aim of these pilot studies was two-fold: 1) to understand how humans interact naturally in this social environment, and 2) to use that information to adopt a data-driven approach for development of the Social AI in order to begin to elucidate components of the social interaction that affect social fluidity in this environment. We hypothesize that for the agent to be perceived as a “real person”, its social actions are important such as leaving resources to the one in need, responding appropriately to player actions, but that also simple auditory/visual cues may play a critical role [2]. Below, we describe both the development of the Social AI as well the pilot studies.

II. Experiments

In this section, we describe two pilot study experiments which were conducted in sequence (henceforth referred to as “1st Pre-Test” and “2nd Pre-Test”). For each of these experiments, we detail how they were conducted, what we learned, and how the results were utilized as part of the Social AI development. Each study was conducted using lab

[*] Corresponding Author

personnel (n=6 for the first pre-test, and n=8 for the second pre-test), comprised of 6 males and 2 females. Protocols were developed for each pre-test (described below) with the aim of emulating naturalistic game-playing behavior.

A. 1st Pre-Test

For the first pre-test, we played human versus human, without the Social AI. The focus was on developing an analytical understanding of the social environment, such as the flow of the game and relevant player interactions, as well as to collect needed data about what triggers those interactions, such as resources, tools, and player status. The steps of the first pre-test were as follows. First, we purchased "Don't Starve Together" from Steam and added each other as "Steam friends". Second, inside the game, we made a secure room for friends only to prevent any interruption in the test, allowing for uninterrupted 30-minute one-on-one gameplay sessions. After that, we set up a Zoom meeting to allow audio-visual communication with each other while playing the game. We then used OBS Studio (https://obsproject.com/) to record the entire screen during the gameplay, including the game window itself as well as the Zoom window of simultaneous social interactions. An example of this can be seen in Figure 1.

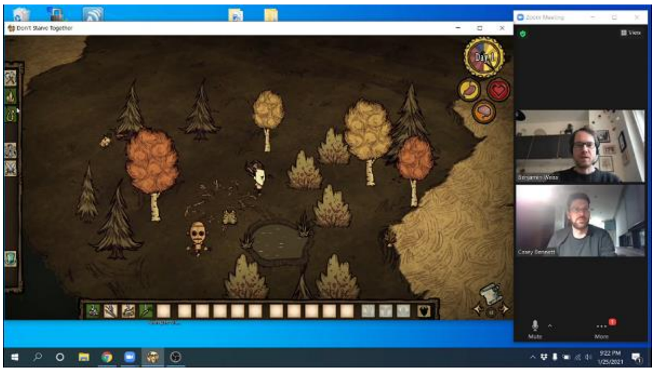

Fig. 1. Gameplay example of "Don't Starve Together" during the first pre-test

B. 1st Pre-Test Results

After the first pre-test, we analyzed the recorded video and made annotations of player utterances and the immediate situation in which the communication occurred. The situations comprise various aspects related to the game, such as events changing the status of resources, monsters in the vicinity, and activities like making tools, etc. Based on these annotations, a hierarchy diagram was derived to specify each situation and the related speech samples. The hierarchy was derived by four separate coders, who first categorized the utterances independently, then worked during a focus group to align those categories into a hierarchy.

In order to use this hierarchy diagram for creating a first Social AI, two major parts followed. One was the game modification part (henceforth referred to as the "Game Mod") and the other was the Speech Dialogue System part. For the Game Mod part, we extracted required data based on the hierarchy diagram in order to define the social game state relevant for triggering spoken interaction. These game state definitions were then used to create the Game Mod. This Game Mod had two primary capabilities. One was a customizable game setup to enable testing the interaction between the human player and Social AI in various social environment scenarios. The other was the "game data writing" functionality, which allowed for game data related to our above game state definitions as well as in-game interactions to be written continuously in real-time as an external file during game play. For the Speech Dialogue System part, we linked the written game data to the Social AI, capable of reacting to in-game events through autonomously generated speech. We used locally-installed (Window or Mac) voice packages as part of the Text-to-Speech (TTS) module, with the audio output redirected to an internal "virtual" microphone jack, then used the Loomie application (https://www.loomielive.com/) as a visual avatar capable of moving its lips synchronously with the speech.

C. 2nd Pre-Test

For the second pre-test, we focused on evaluating the prototypical Social AI and the Game Mod mechanics. During the second pre-test, a human player played the game with the Social AI (represented by the virtual avatar). The experiment was setup using a wizard-of-oz design, where the avatar was capable of autonomous speech based on in-game events but the actual in-game character actions were controlled by a human confederate, as we had not yet implemented AI mechanics for in-game character behavior. Human players were instructed to try to talk and act normally as if they were playing with the other human player. The experimental protocol was similar to the first pre-test, but with the following additions. First, we started the game on the Social AI (represented as the virtual avatar) side, applying the Game Mod that we created based on the first pre-test. After that, we set up a Zoom meeting same as the first pre-test, except in this case the human players appeared on-screen alongside the Social AI in the form of the Loomie virtual avatar. Next, we ran the Text-to-Speech (TTS) module, then recorded the entire screen during gameplay including both the game window as well as the Zoom window of simultaneous human-avatar interactions. Each trial lasted approximately 30 minutes. An example of this can be seen in Figure 2.

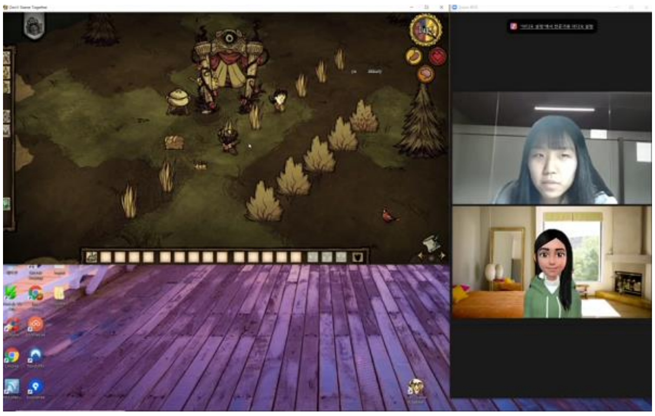

Fig. 2. Gameplay example of Social AI Interaction during the second pre-test. Social AI is represented as the virtual avatar on the lower right-hand side of screen.

At the end of each participant trial, we collected qualitative data about the participant experience using a questionnaire, which included questions about when they felt like the Social AI speech matched the gameplay (or did not), things they found annoying, issues with the Game Mod and/or social environment, among others. That data was then analyzed about common patterns in the experience, to identify things to augment the Social AI and the Game Mod for future experiments.

D. 2nd Pre-Test Results

During the second pre-test, several types of data were collected. For each trial, we obtained gameplay video (as shown in Figure 2), written game data from the Game Mod about the Social AI, and participants' questionnaires about their experience. Analyzing the questionnaires, we first extracted lists of frequent statements, which were then summarized into categories of missing "components" related to the social environment and behavior of the Social AI. Participants reported six common aspects in which the AI could be more "life-like" in the future.

1. There is a need to add reactions to the player's communication with the Social AI. Currently, the Social AI ignores the player's talk because it produces utterances based on the extracted game data only. Therefore, automatic speech recognition (ASR) will be implemented. This feature was already included in the speech hierarchy but not yet realized for the second pre-test.

2. Participants reported a desire for the Social AI to speak in a more friendly and natural manner, so the player could feel like they are playing with another human player. Some of this may entail simple re-phrasing of existing speech utterances, but the rest may require introduction of *sentence inflection* to create variability and unpredictability in how the utterances are spoken.

3. The Social AI should be aware of its own past communication, so it does not repeat certain statements unnaturally often or quickly. These repetition delays are similar to Inhibition of Return mechanisms seen in human attention systems [9].

4. Events involving sharing objects during gameplay are currently not included in the social game state. Therefore, no appropriate speech is produced relating to such events. However, participants noticed the Social AI only commented on indirect interactions, not these "direct" interactions.

5. The Game Mod needs some fixing. Currently, only the data from the viewpoint of the Social AI is recorded. If we could extract the player's data appropriately, the Social AI could speak utterances based on the player's data. (i.e. player's perspective). This kind of "mentalizing" is akin to Theory of Mind approaches in social robotics [10]. An example would be like the Social AI could *suggest* an action if the player does not move at the same position for a while.

6. The video recordings need to be analyzed and compared to the written game data, in order to allow the Social AI to give plans rather than just speaking about the current game status. Participants noted that the Social AI only talks about the current game state, not future events.

Related to the above, we are currently undertaking work to annotate the videos for use in machine learning models to predict game events and make plans based on the written game data. Similarly, facial expression analysis work on the human player during interactions with the Social AI is currently underway, which could be used to create more multi-modal interactions involving non-verbal cues to augment the Speech Dialogue System.

III. DISCUSSION

In this paper, we describe preliminary results from pilot tests of the Speech Dialogue System for the Social AI developed for cooperative game environments. We made and tested a first version of the Social AI which presently has rudimentary autonomous speech interaction capabilities. As mentioned in the introduction, the long-term goal is to explore factors affecting perceptions of social fluidity in human-agent interactions related to intentionality attribution [2,3], and how we can use such knowledge to emulate natural social behavior in cooperative social environments [1,7]. However, before this Social AI can be used in experiments studying components of artificial intentionality, several improvements need to be carried out. For the Speech Dialogue System, six components were identified to be revised or added based on the results (see Section 2.D).

These components have not been identified with the traditional aim of optimizing usability or user experience of a spoken service [11], but with the goal of intentionality attribution in an autonomous agent. While these aims do not exclude each other, our approach resulted in a specific view on appropriateness and agency, rather than effectiveness or efficiency. Therefore, these components include introducing appropriate numbers and timing of statements, responding to the player's talk, and suggesting future plans. Many of these components are directly tied to the cooperative nature of the social environment as well, which underscores the interplay of the AI behavior and contextual factors [12]. Indeed, the speech dialog capabilities are deeply interlinked to the characters' cooperative actions and the game session evolvement – and this is a pre-requisite for creating a successful spoken agent [13]. This situatedness demands an empirical (in our case data-driven) design approach chosen, which is best practice for designing successful voice interaction [14]. We successfully completed the two pre-tests and having systemized typical communication at the game, the next step aims to improve these aspects by adding additional functionality, which will be tested in larger-scale human interaction experiments. Moreover, we aim to define personality aspects for the Social AI in the form of creating archetypal personas [8]. This would not only support consistent wording but would also allow to systematically identify further verbal activities that are only required for the target AI's personality, such as communication preferences (e.g., announcing own movements, engaging in more social banter) or politeness phrases.

The above ongoing research provides insight into the development of the "Social AI" (using the Text-to-Speech (TTS) module and Loomie virtual avatar) through a *data-driven approach* to explore how humans interact in cooperative social environments such as video games, and then applying those findings to an artificial agent. It also highlights how that same process can be used to create customizable social environments, to explore a broad range of hypotheses related to how contextual factors relate to people's perceptions of interactive technology.

ACKNOWLEDGMENT

This research was supported through funding by a grant from the National Research Foundation of Korea (NRF) (Grant number: 2021R1G1A1003801). We would also like to thank Cheda Stanojevic and Sawyer Collins from Indiana University for their assistance in this work.

REFERENCES

[1] Stephanie Tulk Jesso, William G. Kennedy and Eva Wiese, "Behavioral Cues of Humanness in Complex Environments: How People Engage With Human and Artificially Intelligent Agents in a Multiplayer Videogame," Frontiers in Robotics and AI, vol. 7, article 531805, 2020.

[2] Catherine S. Oh, Jeremy N. Bailenson and Gregory F. Welch, "A Systematic Review of Social Presence: Definition, Antecedents, and Implications," Frontiers in Robotics and AI, vol. 5, article 114, 2018.

[3] Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan, "What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach," In ACM Conference on Human Factors in Computing Systems (CHI), Yokohama, Japan, 2021.

[4] Daniel C. Dennett, "Intentional systems," The Journal of Philosophy, vol. 68, no. 4, pp. 87–106, 1971.

[5] Casey C. Bennett and Selma Šabanović, "Deriving minimal features for human-like facial expressions in robotic faces," International Journal of Social Robotics, 6(3), pp. 367-381, 2014.

[6] Angelika Hönemann, Casey C. Bennett, Petra Wagner, and Selma Šabanović, "Audio-visual synthesized attitudes presented by the German speaking robot SMiRAE," In AVSP2019 International Conference on Auditory-Visual Speech Processing, Melbourne, Australia. 2019.

[7] Filipa Correia, Patricia Alves-Oliveira, Tiago Ribeiro, Francisco Melo, and Ana Paiva, "A social robot as a card game player," In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 13, No. 1), 2017.

[8] Alan Cooper, "The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity," SAMS, 1999.

[9] Orit Nafcha, Simond Shamay-Tsoory, Shai Gabay, "The sociality of social inhibition of return," Cognition, vol. 195, article 104108, 2020.

[10] Jaime Banks, "Theory of mind in social robots: replication of five established human tests," International Journal of Social Robotics, pp. 1-12, 2019.

[11] Sebastian Möller, "Perceptual quality dimensions of spoken dialogue systems: a review and new experimental results," In: Proceedings of the of Forum Acusticum, Budapest, pp. 2681–2686, 2005.

[12] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz et al. "Mental models of ai agents in a cooperative game setting," Proceedings of the Conference on Human Factors in Computing Systems (CHI), pp. 1-12, 2020.

[13] Roger K. Moore, "From talking and listening robots to intelligent communicative machines," In J. Marko Witz (ed.), *Robots That Talk and Listen*. Boston, MA: De Gruyter, 2015.

[14] Cathy Pearl, *Designing Voice User Interfaces*, O'Reilly, Beijing, 2017.