

Using Virtual Humans to Study Embodied Social Interaction

Chen Yu, Hui Zhang and Linda B. Smith

Department of Psychological and Brain Sciences, and Program in Cognitive Science
Indiana University, Bloomington, 47405, IN USA

Abstract

An important step to build machines that can learn from social interactions with human users is to understand the nature of learning-oriented interactions. To do so, a central problem is to find a way to decouple the social interaction between two agents (e.g. a human supervisor and a machine learner), so that we can systematically manipulate and control the dynamic flow of the interaction to create and examine various interactive learning conditions. In this paper, we build a set of virtual humans as language learners possessing different social-cognitive skills, and ask real people to teach them object names. Using multisensory recording devices, we measured how well real people interact with virtual humans and how they shape their behaviors to adapt to different social-cognitive skills that virtual humans possess. Multimodal data were analyzed to shed light on both perceptual and behavioral aspects of human users in interaction. These results can be used both to guide building artificially intelligent system and to provide useful insights on human-human communication and child language learning.

Introduction

The ways that humans acquire knowledge are quite different with current machine learning (ML) approaches. Most ML systems first collect data with (or without) teaching labels from users and the environment, and then rely on implementing efficient mathematical algorithms and applying them onto pre-collected data to infer knowledge. The methodology largely assumes that a learner (e.g. a machine) passively receives information from a teacher (e.g. a human supervisor) in a one-way flow. In contrast, humans are situated in social contexts and acquire knowledge through our everyday social interactions with others. For example, in child language learning, parents dynamically adjust their behaviors based on their understanding of a child's mental state. Thus, language teachers provide "on-demand" information in real time learning. Meanwhile, the child also actively generates actions to interact with the physical environment and to shape the teachers' responses and acquire just-in-need data for his learning. Thus, current machine learning studies focus on one aspect of learning – what kind of **learning device** can perform effective computations on pre-collected data, but ignore an equally important aspect of learning — the **learning environment** that a learner is situated in. An important way in which human learning does not resemble passive machine learning approaches is that there is an active social partner in the learning environment.

The present study attempts to systematically investigate the role of social interaction in automatic language learning in machines. A typical scenario of language learning is like this: a language teacher provides spoken names of things in the physical world. Meanwhile, a language learner perceives and processes information collected from the learning environment and the social partner to build his vocabulary. In this

kind of interaction, both the learner and the teacher dynamically adjust their behaviors based on the responsive actions from the other agent. Without interfering with the interaction, we can control neither the learner's responses nor the teacher's actions. Without systematically manipulating some variables in the interaction, we cannot perform quantitative analyses of the role of social cues in language learning.

We argue that the key to solve the above puzzles is to decouple the social interactions between two agents without interfering with the interaction itself. To achieve this goal, we propose and implement a new paradigm based on virtual reality techniques. The central idea is to use virtual humans as well-controlled agents to interact with real users. In this way, we can pre-program virtual humans to generate different kinds of behaviors and demonstrate different kinds of social capabilities. Then we can use them as a tool to systematically manipulate the learning environment in social interactions and measure the adaptive behaviors of real humans. Following this general idea, the present work builds a set of virtual learners who demonstrate different kinds of social understanding (e.g. following the eye gaze of real teachers) when real teachers are asked to interact with them and teach them object names. The questions we seek to answer are (1) how well real humans perceive behaviors and social skills of virtual humans; (2) whether and if so, in what ways real humans shape their behaviors based on their observation of virtual learners' states; and (3) what kind of learning environment real teachers provide through social interactions.

A Virtual Reality Platform

As shown in Figure 1, the virtual humans we developed possess a set of social capabilities embodied by sensorimotor primitive actions, such as facial expressions, pointing at objects by hand, gazing and following a real user's attention. We design different virtual humans with different social capabilities and use them to measure and analyze the behaviors and responses of real humans when they interact with virtual humans. One important issue in our design is the "behavioral realism" of the virtual agents, which means that virtual humans should act and respond like real humans, or in other words, they should be believable in terms of both the physical actions of virtual humans themselves, and their social interactions with real humans (MacDorman et al., 2005). In our design, we use Boston Dynamics's DI-Guy libraries to animate lifelike human characters that can be created and readily programmed to generate realistic human-like behaviors in the virtual world, including **gazing** and **pointing** at an object or a person, **walking**, and **moving lips** to synchronize with speech while **speaking**. In addition, the virtual human can generate 7 different kinds of **facial expression**, such as smile, trust, sad, mad and distrust. All these combine to result in smooth and lifelike behaviors being generated automatically.

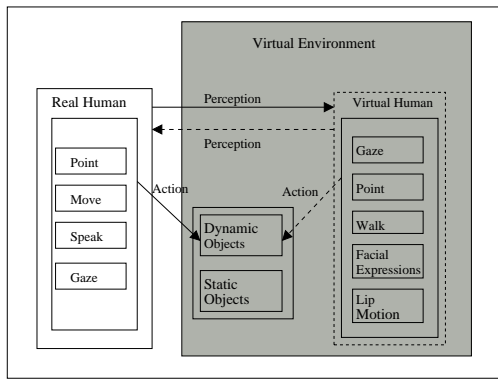


Figure 1: A real person and a virtual human interact in a virtual environment. We control the actions of the virtual person and measure the behavioral responses of a real person.

Experiment

As shown in Figure 2, we recruited 26 college students who were asked to teach virtual foreigners the names of several everyday objects. They were allowed to point to, gaze at and move those objects through a touch screen. There was no constraint about what they have to say or what they have to do. There were three conditions in this experiment wherein three virtual agents demonstrated different levels of engagement in interaction - engaged in 10%, 50% or 90% of total interaction time. When a virtual human is fully engaged in interaction, she would share visual attention with a real teacher and generating positive facial expressions (e.g. smile, trust, etc.). While she is not engaged, she would look at somewhere else with negative facial expressions (e.g. sad, conniving, etc.). The objects attended by a real person are detected based on where he is looking as well as his actions on those objects through the touch screen. The attentional information is then sent to the virtual human so that she can switch her attention to the right objects in real time when she is in the engaged state. We recorded real people's behaviors in interaction including their pointing and moving actions, speech acts and eye gaze. Moreover, they were asked to complete questionnaires at the end of the experiment.

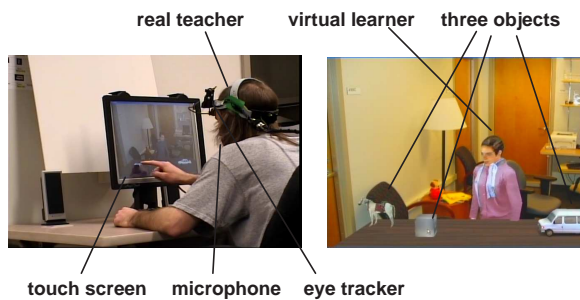


Figure 2: Left: a participant wearing an eye tracker and a microphone interacts with the virtual human in a virtual environment through a touch screen. Right: the VR scene consists of a virtual human and three objects on a table in each trial.

A 5-point Likert scale was used for a set of 10 questions in our questionnaire. Those questions focus on different aspects of participants' perception of the social-cognitive skills of three virtual humans: (1) **Joint attention and eye contact** measures how much the participants felt that eye movements of virtual humans were natural, social and friendly. (2) **Social intelligence/engagement** measures how much the participants felt that virtual learners were engaged during interaction.

(3) **Overall intelligence** measures participants' estimates of virtual learners' intelligence. (4) **Gaze time estimation** asks participants were also asked to estimate the amount of time (on a scale of 0 to 100 percent) that virtual humans paid attention to their behaviors.

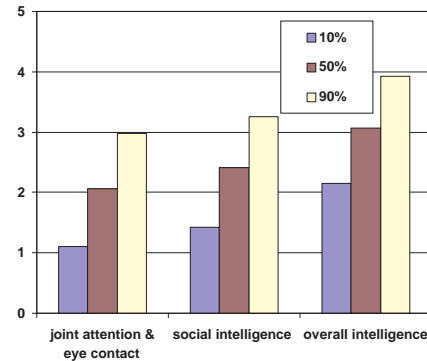


Figure 3: Participants' evaluations of three virtual humans.

Table 1: The estimated engagement times of virtual humans

	10%	50%	90%
gaze	M= 22.50%	M=54.37%	M= 86%
time	SD= 22.10%	SD= 23.89%	SD=16.1%

Figure 3 shows a comparison of the results of three virtual humans with different engagement levels. Clearly, participants were aware of social behaviors of virtual humans and provided quite consistent estimates of their social sensitivities. Thus, the significant differences between three conditions are not surprising. We note that even when the virtual human almost fully engaged in interaction by following the real person's actions in 90% of the total time, most people were still not satisfied with the virtual human's social behaviors. Table 1 shows the estimated times that virtual humans pay attention to participants' behaviors. Although the means of two out of three estimated times are close to 50% and 90% separately. Surprisingly, participants provided quite different estimates in all of three conditions.

Conclusion

Compared with using real robots, virtual humans are easy to implement and use because we can neglect low-level technical problems, such as motor control of joint angles, which perfectly matches our research purpose - studying high-level social-cognitive skills in language learning. Our results show that real people treat virtual humans as social partners and are willing to interact with them. We report our first steps to systematically and quantitatively study social cues in language learning, suggesting that learning words through social interaction will have the best chance to approaching human capacity. In addition, we argue that the new proposed paradigm itself is promising to study the roles of social cues in both human language learning and machine intelligence.

References

MacDorman, K., Minato, T., Shimada, M., Itakura, S., Cowley, S., & Ishiguro, H. (2005). Assessing human likeness by eye contact in an android testbed. In *Proc of the 27th annual meeting of the cognitive science society*.