

Real-Time Adaptive Behaviors in Multimodal Human-Avatar Interactions

Hui Zhang, Damian Fricker, Thomas G. Smith, Chen Yu
Indiana University, Bloomington
{huizhang, dfricker, thgsmith, chenyu}@indiana.edu

ABSTRACT

Multimodal interaction in everyday life seems so effortless. However, a closer look reveals that such interaction is indeed complex and comprises multiple levels of coordination, from high-level linguistic exchanges to low-level couplings of momentary bodily movements both within an agent and across multiple interacting agents. A better understanding of how these multimodal behaviors are coordinated can provide insightful principles to guide the development of intelligent multimodal interfaces. In light of this, we propose and implement a research framework in which human participants interact with a virtual agent in a virtual environment. Our platform allows the virtual agent to keep track of the user's gaze and hand movements in real time, and adjust his own behaviors accordingly. An experiment is designed and conducted to investigate adaptive user behaviors in a human-agent joint attention task. Multimodal data streams are collected in the study including speech, eye gaze, hand and head movements from both the human user and the virtual agent, which are then analyzed to discover various behavioral patterns. Those patterns show that human participants are highly sensitive to momentary multimodal behaviors generated by the virtual agent and they rapidly adapt their behaviors accordingly. Our results suggest the importance of studying and understanding real-time adaptive behaviors in human-computer multimodal interactions.

Categories and Subject Descriptors

H.1.2 [User/Machine Systems]: Human factors and Human information processing

General Terms

Design, Experimentation, Human Factors.

Keywords

Embodied agent, Virtual human, Multimodal interaction, Visualization.

1. INTRODUCTION

Human perceptual and cognitive systems are by nature multimodal. We make contact with the physical world through a vast array of sensory systems -- vision, audition, touch, smell, to name a few. Moreover, our sensorimotor experiences are closely coupled. What we will see next depends on how we currently shift our gaze (and position of the whole body) in the physical

environment and what current actions we may take in this environment. Meanwhile, what action we will take next also depends on what we see at the present moment, which provides sensory information for our motor control system. This multimodal perspective can also be easily extended into the case of human-human communication and social interaction. The body and its momentary actions are crucial to social collaboration by serving as outward signs, observable by social partners, and are tightly tied to our own internal cognitive state.

How can we build an intelligent agent (a physical robot or a virtual avatar) that can interact with human users to emulate real-time smooth fluidity of collaborative human behaviors as everyday conversation or joint action? This challenge requires intelligent agents to meet with the human user's expectations and sensitivities to the real-time behaviors generated by virtual agents and perceive them in the similar way just as the user interacts with other humans. In light of this, the goal of the present study is to investigate how human participants adjust their multimodal behaviors in real time based on their perception of real-time behaviors from the other interacting agent. Toward this goal, we develop a multimodal human-avatar interaction platform and conduct an empirical study using this new platform to collect fine-grained multimodal behavioral data. Next we analyze such data to discover interesting patterns which can then be used to shed lights on fundamental principles in multimodal human-agent interactions. After reviewing related work, the present paper will first describe our multimodal interaction platform, followed by a description of our experimental design and setup. We will then present multimodal patterns discovered from human-avatar interaction and further discuss the insights gained from this multimodal data analysis that can be used to guide the design of better multimodal interfaces in human-computer interaction.

2. Related Work

A rich literature on multimodal human-computer interactions already exists, researchers and scientists have taken various approaches including recording and analyzing user behaviors in different natural and semi-natural situations, and Wizard of Oz studies (see e.g., [1,2]). Lately there have been several systems that tried to improve the smoothness of human-computer interactions through predicting the right time for feedbacks. Ward and Tsukahara [5], describe a pause-duration model based on the best fit to speech acts. Gratch et al. [6] describe a recent experiment on multimodal, yet purely nonverbal agent feedback and its effects on the speaker; their work analyzes the speaker's head moves and body postures captured through a camera, and implements a pitch cue algorithm to determine the right moment for giving feedbacks by head nods and gaze.

Other representative efforts include studying real-time behaviors in human-computer interaction to get insights into how to build better interfaces. In [7], an avatar generated reactive gaze behavior that is based on the user's current state in an interview

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI-MLMI'10, November 8–12, 2010, Beijing, China.

Copyright 2010 ACM 978-1-4503-0414-6/10/...\$10.00.

scenario. In [8], sequential probabilistic models were used to select multimodal features from a speaker (e.g. prosody, gaze and spoken words) to predict visual back-channel cues (e.g. head nods). [3] built an engagement estimation algorithm based on analyzing gaze transition patterns from users. [4] developed a real-time gaze model for embodied conversational agents that generated spontaneous gaze movements based on the agent’s internal state of cognitive processing. There has also been growing interest in more micro-analytic studies of real-time multimodal behavioral data in both human-human communication and human-agent interface (see, e.g., [14, 15]).

All those studies point to a critical issue to design better agents that can interact with participants in human-like ways, that is, they need to not only generate appropriate behaviors but also execute those actions at the right moment and with the right timing. For example, head nodding at the right moment may reflect a listener’s understanding as a back-channel feedback signal. In contrast, nodding at the unexpected moment may cause the speaker’s confusion in reading/accessing the listener’s attentional state. Similarly, a nodding action with abnormal timing may cause interruptions in communication. Although there is no doubt about the importance of this direction, how to study real-time adaptive behaviors remains a challenge and there are few studies attempting to systematically address this topic. With the advances of modern multi-modal sensory equipment, interactive computer graphics and data mining techniques, we develop a highly integrated multimodal interaction platform that allows us to collect and analyze fine-grained real-time multimodal data in human-agent interactions.

3. A Multimodal Experimental Platform

Our proposed work is specifically concerned with systematically studying the exact timing of real-time interactions between humans and virtual agents. To achieve this goal, we implement a research framework for studying and evaluating different important aspects of multi-modal real-time interactions between humans and virtual agents, including establishment of joint attention via eye gaze coordination (an example application we will demonstrate by a pilot study described below), coupling of eye gaze, gestures, and utterances between virtual speaker and human listener in natural dialogues, and mechanisms for coordinating joint activities via verbal and nonverbal cues.

Our approach for studying multimodal human-avatar interaction is well exemplified by Cassel’s Study-Model-Build-Test development cycle [9]. Specifically, we have three primary goals of building and using such a framework:

- 1) to test and evaluate moment-by-moment interactive behavioral patterns in human-agent interaction;
- 2) to develop, test and evaluate cognitive models that can emulate those patterns;
- 3) to develop, test and design new human-agent multimodal interfaces which include the appearance of the virtual agent, the control strategy as well as real-time adaptive human-like behaviors.

More importantly, we expect to use this platform to discover fundamental principles in human-agent interaction which can be easily extended to various scenarios in human-computer interactions. Therefore two critical requirements for our framework are that it be able to collect, in an unprecedented way, fine-grained multi-modal sensorimotor data that can be used for

discovering coupled behavioral patterns embedded in multiple data streams from both the virtual agent and the human user, and that the virtual agent can monitor the user’s behaviors moment by moment, allowing the agent to infer the user’s cognitive state (e.g. engagement and intention) and react to it in real time.

We are thus motivated to develop a framework which allows human participants to interact with a virtual agent in a virtual environment through multimodal interaction. There are three key components (as shown in Figure 1) in this research platform:

- an virtual experimental environment;
- a virtual agent control system;
- multi-modal sensory equipment;

In the following, we provide details of each of the three components.

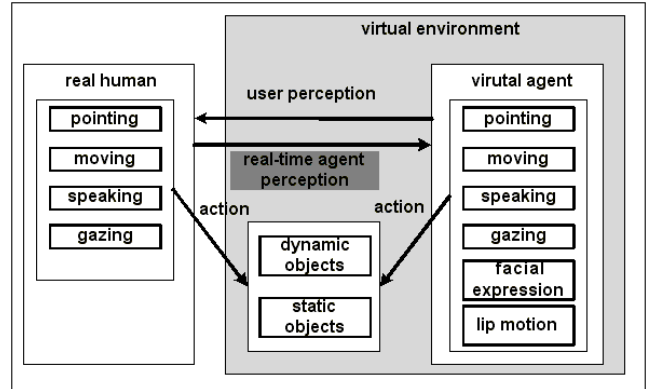


Figure 1: Overview of our research platform to investigate multimodal human-avatar interactions. A real person and a virtual human interact with each other in a virtual environment. We control the actions of the virtual person and measure the behavioral responses of a real person. A critical component in the current setup is that the virtual agent can perceive the human user’s behaviors in real time and generate spontaneous responsive actions accordingly.

3.1 The Virtual Experimental Environment

This virtual environment consists of a virtual living room with everyday furniture, e.g. chairs and tables. This virtual scene is rendered on a computer screen with a virtual agent standing behind a table so that she can have a face-to-face interaction with the human user. There are a set of virtual objects on the table in the virtual living room that both the virtual agent and the human user can move and manipulate. The virtual human’s manual actions toward those virtual objects are implemented through VR techniques and the real person’s actions on the virtual objects are performed through a touch-screen which is covered on the computer monitor. There are several joint tasks that can be carried out in this virtual environment. For example, the real person can be a language teacher while the virtual agent can be a language learner. Thus, the communication task is for the real person to attract the virtual agent’s attention and then teach the agent object names so the virtual agent can learn the human language through social interaction. For another example, the virtual agent and the real user can collaborate on putting pieces together in a jigsaw puzzle game. In this collaborative task, they can use speech and gesture to communicate and refer to pieces that the other agent can easily reach.

3.2 The Virtual Agent

3.2.1 Building Human-like Virtual Agent

In our implementation, we use Boston Dynamics DI-Guy libraries to animate virtual agents 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 in a specific 3D location, walking to a 3D location, 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 behaviors being generated automatically. A critical component in the virtual human is to perceive the human user's behavior in real time and react appropriately within the right time span. As shown in Figure 1, this human-like skill is implemented by a combined mechanism including real-time eye tracking and motion tracking on the human side, real-time data transfer in the platform and real-time action control on the virtual human's side.

3.2.2 Generation of Virtual Agent's Multimodal behaviors

Our basic implementation of the virtual agent's eye gaze uses a stochastic model, as illustrated in Figure 3. Transitions between virtual agent's "Looking-At" state and "Looking-Away" state are triggered by real-time behaviors from the participant as the agent's control system can access those data in real-time interaction. Figure 2 shows four examples of the virtual agent's behaviors: a) The avatar shares visual attention with the user; b) The avatar is looking at an object that the real person is not attending; c) The avatar is looking at somewhere randomly in the 3D virtual environment; and d) The avatar is looking at the real person's face.

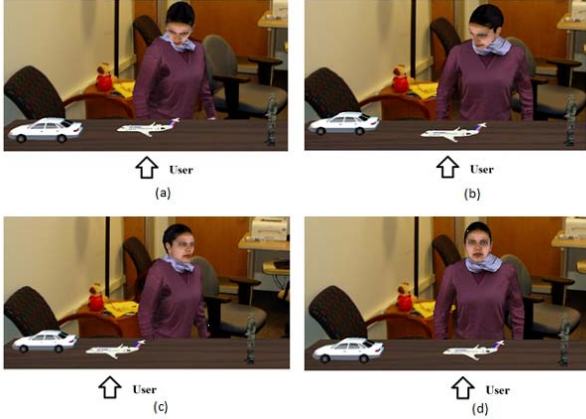


Figure 2: Eye gaze behaviors for virtual agents in our experimental design. The arrows point to the object that a human participant is attending. The virtual agent can access this information in real time by monitoring both the human's actions on the touch-screen panel and the human's eye gaze. With such momentary information, the virtual agent may decide to generate four different behaviors: a) joint attention: attending to the same object that the human participant attends; 2) not joint attention: the virtual agent looks at one of the other objects that the real person is NOT attending; 3) disengaged: the virtual agent just randomly looks around; and 4) face-to-face: the virtual agent looks at the face of the real person.

3.3 Multi-modal Sensory Equipment

During human-avatar interaction, our platform collects fine-grained behavioral data from both the virtual agent and the human user (see Figure 3). On the human side, a Tobii 1750 eye tracker is used to monitor the user's eye movements at the frequency of 50Hz. The user's manual actions on virtual objects through the touch-screen are also recorded with timing information on a dedicated computer. Meanwhile, the system also records the user's speech in the interaction. More recently, we added a video camera pointing to the face of the user and a faceAPI package from **SeeingMachine** (www.seeingmachine.com) is deployed and integrated into the whole system to record 38 3D face landmarks plus head rotation and orientation at the frequency of 15Hz.

On the virtual human side, our VR program not only renders a virtual scene with a virtual agent but also records gaze and manual actions generated by the virtual agent and his facial expressions (30 FPS). In addition, we also keep track of the locations of objects on the computer screen. As a result, we gather multimodal multi-stream temporal data streams from human-avatar interactions. All of the data streams are synchronized via system-wide timestamps.

4. Experiment

The overall goal of this research platform is to build a better human-agent communication system and understand multimodal agent-agent interaction. Joint visual attention has been well documented as an important indicator in smooth human-human communication. In light of this, our first study focuses on joint attention between a human user and a virtual agent. More specifically, given the real-time control mechanism implemented in our platform, we ask how a human agent reacts to different situations wherein the virtual agent may or may not pay attention to and follow the human agent's visual attention. In practice, we

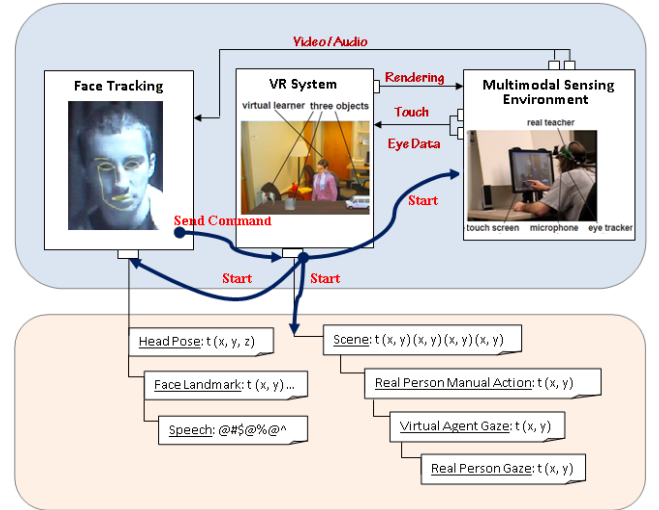


Figure 3: Multimodal Data Recording. Top: A real person and a virtual human are engaged in a joint task with a set of virtual objects in a virtual environment. The platform tracks the user's gaze and hand movements in real time and feed the information to the virtual agent's action control system to establish real-time perception-action loops with real and virtual agents. Bottom: multiple data streams are recorded from human-avatar interactions which are used to discover fine-grained behavioral patterns and infer more principles.

employed a word learning task where the human participants were asked to teach the agent the names of a set of objects. We selected this task for five reasons: (1) it has an explicit goal that allows participants to naturally engage with the agent in interactions while being constrained enough to make real-time processing on the agent’s actions feasible, which in turn allows for adaptive agent behavior; (2) it has been used successfully in a variety of developmental studies investigating multi-modal human-human interactions (e.g., between parents and their children [10]); (3) it allows us to investigate the fine-grained temporal patterns and relationships between human eye gaze and human speech as part of the larger joint attention processes; (4) beyond modeling human interactions, the task itself has its own merits as it can help shed light on how agents might acquire new knowledge through human-agent social interaction [11]; and (5) it can ultimately be used to develop cognitive models of temporal interaction patterns that in an unprecedented way capture the time course of human-human interactions (cp. to [3]).

The experimental paradigm is noteworthy in that it is:

- **Multimodal:** Participants and the agent interact through speech and visual cues (including perceivable information from vision, speech, and eye gaze).
- **Interactive and adaptive:** The agent can follow what the human is visually attending to (based on real-time tracking of human eye gaze) and thus provide visual feedback to human subjects who can (and will) adjust their behavior in response to the agent’s response.
- **Real-time:** The agent’s actions are generated in real time as participants switch their visual attention moment by moment.
- **Naturalistic:** There are no constraints on what participants should or should not do or say in the task.

Table 1: Five experimental conditions with different engagement levels and attentional states.

condition	description
90%	90% of time engaging +10% on one of the other objects
50%/random	50% engaging + 50% on random locations
50%/object	50% engaging + 50% on one of the other objects
10%/random	10% engaging + 90% on random locations
10%/object	10% engaging + 90% on one of the other objects

In the present experiment, we manipulated three engaged levels of the virtual agent: 90%, 50% or 10% engaged. As we described earlier, if the agent is engaged in the interaction, she will show her interests toward the object that the human participant is manually holding and manipulating. In the 90% engaged condition, the agent is engaged 90% of time. Similarly, the agent

is engaged 50% or 10% of time respectively in the other two conditions. In those two less engaged conditions, at those moments that the agent is not engaged, we designed two sub-conditions, in the 50%/object condition, the agent showed her interest to one of the other objects that the real participant was not attending. In the 50%/random condition, the agent randomly looked at some spatial location without paying attention to any of the three objects on the table. Table 1 summarizes 5 experimental conditions.

Will participants in the less engaged conditions pay overall more attention to the virtual agent compared with when they are in the 90% condition? Moreover, in addition to a comparison of overall eye movement patterns in those five conditions, the more interesting research question is in what ways the behaviors of the participants may differ in those conditions at a fine-grained level. For example, subjects might spend more time attending to the virtual agent with longer eye fixations in the less engaged conditions. They also might spend more time in attracting the agent’s attention before naming objects, and therefore, generate fewer naming utterances. Alternatively, they might more frequently monitor the agent’s attention with shorter eye fixations and generate more naming utterances in order to attract the agent’s attention through the auditory channel. Furthermore, open questions can be asked about the details of eye fixations in conjunction with naming events: will subjects look more at the agent or more at the named object? Will their eye movement patterns change over the course of naming speech production? Moreover, a direction comparison between the random and object conditions with the same level of engagement (either 50% or 10%) would inform us to what degree participants will be sensitive to where the virtual agent attend when the agent is not engaged; and how the factor of the overall level of engagement and the factor of where to look when not engaged may interact.

To answer those questions, we collected and analyzed fine-grained multimodal behavioral data that would allow us to discover the time course of sensorimotor patterns.

4.1 Participants

25 undergraduate students at Indiana University participated in the study (2 of them were excluded due to technical problems with their eye tracking).

4.2 Procedure

As shown in Figure 3, participants were asked to sit in front of a Tobii eye tracking monitor and the experiment started with eye gaze calibration in which they were asked to sequentially look at 5 calibration points displayed on the screen. Next, the experimenter calibrated the eye tracker using the Tobii eye-tracker software. A logitech webcam with a microphone was attached on the computer monitor, pointing to the participant, to record the subject’s speech and as well as capture the participant’s face image.

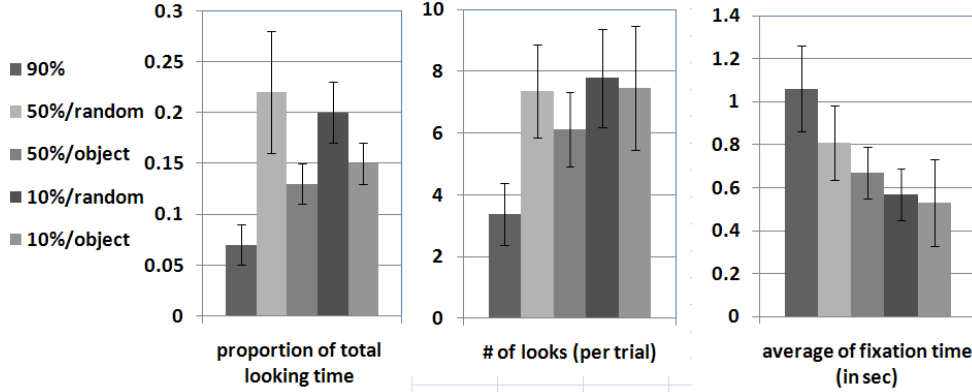


Figure 4: Gaze patterns on the virtual agent’s face. Across the five conditions, human participants gazed at the face of the virtual agent to monitor the agent’s visual attention. They did so more frequently when the agent was less engaged in the interaction. Three measures of the total proportion of looking time, the number of looks, and the average fixation time reveal different strategies that human participants applied in different situations.

There were three trials in each condition and in total 15 trials for all of the five experimental conditions for each subject to participate. The order of those trials was randomized. Within each trial, participants saw three virtual objects on a virtual table and were instructed to teach a virtual agent those object names. Subjects were allowed to use their hands to move, point and manipulate those virtual objects on the 2D computer screen through a touch panel to attract the agent’s attention, and then teach the agent however they wanted and were encouraged to be creative. The subject was allowed one minute to teach the agent the names of the three objects in a trial. The participant could switch from the current trial to the next one by pressing the spacebar if s/he thought that the virtual agent already learned all of the three object names. Otherwise, after the minute was up, the program automatically switched to the next trial. There was no difference across five conditions in terms of the amount of time participant spent in each trial (Mean of the condition with maximal length = 59.58 seconds, mean of the condition with minimal length = 58.10 seconds).

5. Data and Data Processing

Figure 3 illustrates multimodal data collected from human-agent interaction that will be used in our data analyses. In the following, we briefly overviewed data processing of raw multimodal data.

a) **Visual data:** Since the interaction environment was rendered using computer graphics techniques. We recorded detailed information such as the location of each of three objects, the location of the virtual agent, on the computer screen.

b) **Gaze data:** For both eye gaze data from the human user and the virtual agent, we computed eye fixations using a velocity-based method to convert continuous eye movement data into a set of fixation events (see details in [8]). For each fixation, we superimposed (x,y) coordinates of eye gaze onto the screen image sequence to identify the corresponding Region-of-Interest (ROIs) moment by moment.

c) **Speech:** We implemented an endpoint detection algorithm based on speech silence to segment a speech stream into several spoken utterances, each of which may contain one or multiple spoken words. We then transcribed speech into text. The final result is a temporal stream of utterances, each of which is coded with onset and offset timestamps and a sequence of spoken words in the utterance.

d) **Hand action data:** The human user’s hand actions on objects through the touch screen were analyzed in two different ways. First, we extracted a speed profile of manual actions. Second, we categorized the actions on objects into slow vs. fast movement events. Slow movement events are defined as the human participant moved the objects slowly and within a small area on the screen (< 5 pixels/second). The fast movement events are characteristic of a fast moving from one location to another distant location (>200 pixels/second).

As the results of data processing, we derived multiple time series from multimodal data, including where human participants gaze at moment by moment, where the virtual gazes at, what actions participants generate, what they say and so on. In the following subsections, we first report analyses of eye movement data from participants and then report analyses of their speech and hand movements. Finally, we integrated data streams from different modalities, gaze data, hand movements and speech, and from both the human participant and the virtual agent, and extracted dynamic time-course patterns around those naming moments. All of the following statistics and analyses were based on the whole dataset of 23 subjects and across five experimental conditions, containing about 690,000 images, 1,242,000 gaze data points, 10,156 human eye fixations, 2,536 naming events, and 6,500 hand actions.

6. Results

6.1 Analyses of Gaze data

As shown in Figure 4 (left), participants spent less time on the virtual agent’s face when the agent was 90% engaged in the interaction. In contrast, they spent much more time when the agent is not engaged, suggesting that they kept track of the agent’s attention and noticed that the agent was not quite engaged in both 50% and 10% conditions. Interestingly, there is no difference between 50% and 10% conditions, suggesting that as far as the agent demonstrated distracting behaviors, participants were sensitive to that but meanwhile they would spend a certain proportion of their attention on the agent’s face – the amount of time that is probably enough for them to continuously monitor the agent’s attention. Thus, there was no need to spend more time even the agent was in the 10% conditions. Interestingly, with the same level of engagement, random looks from the agent appeared to be more distracting and made participants look more at the

agent’s face (22% in the 50%/random condition and 20% in the 10%/random condition) compared with looking at the other object (13% in the 50%/object condition and 15% in the 10%/object condition). This result revealed that participants not only noticed when the agent was not paying attention but also in what ways they were not attending -- whether the agent was looking away or attending to other objects on the table, which again demonstrates the participant’s sensitivities to the agent’s behaviors. Next, we measured both the number of looks (Figure 4 middle) and the average fixation time (Figure 4 right). Participants generated fewer but longer fixations in the 90% condition. Instead, in all of the other four conditions, they produced significantly more looks on the virtual agent’s face and each gaze fixation is relatively shorter. This result suggests a strategy of frequently checking the agent’s face to continuously monitor the agent’s attentional state.

Table 2: A comparison of gaze data across 5 conditions when the virtual attend is attending or not attending

total looking	90%	50%/ random	50%/ object	10% random	10%/ object
attending	0.05	0.06	0.06	0.07	0.05
not attending	0.17	0.21	0.19	0.21	0.19

Further, for each experimental condition, we extracted those moments that the virtual agent was attending to the object that the human participant manipulated through the touch-screen panel and those moments that the virtual agent was not attending to what the real user attended to. As shown in Table 2, the proportions of looking time across those conditions show a significant difference between those two kinds of moments within each condition, and meanwhile there was no difference across conditions. Thus, across the five conditions, participants produced similar gaze patterns on the agent’s face at those moments that the agent was attending and as well as those moments that the virtual agent was not attending. The overall engagement level of the agent didn’t influence the participant’s gaze behavior, but instead, their gaze behavior was determined by moment-by-moment attentional state of the virtual agent.

Overall, the present results derived from human gaze suggest that humans were sensitive to the differences in the agent’s behaviors

in the 5 experimental conditions, and they therefore adjusted their behaviors accordingly. Moreover, their gaze behavior was mostly influenced by the momentary attentional state of the virtual agent but not by the overall engagement level of the agent over time.

6.2 Analyses of Hand Actions

In the experiment, human participants actively hold, point and move virtual objects to attract the virtual agent’s attention through the touch-screen panel. On average, they generated approximately 12 distinct actions per trial ($M_{90\%}=12.21$; $M_{50\%/random}=13.21$; $M_{50\%/object}=12.37$; $M_{10\%/random}=12.30$; $M_{10\%/object}=13.26$) with no significant difference across five experimental conditions. And they switched to another object about 4-5 times within a trial ($M_{90\%}=4.51$; $M_{50\%/random}=4.03$; $M_{50\%/object}=4.54$; $M_{10\%/random}=4.59$; $M_{10\%/object}=4.68$) with similar average durations of individual actions ($M_{90\%}=4.12\text{sec}$; $M_{50\%/random}=3.72\text{sec}$; $M_{50\%/object}=3.76\text{sec}$; $M_{10\%/random}=3.95\text{sec}$; $M_{10\%/object}=3.92$). All this suggests similar hand actions across five experimental conditions. However, a closer look of their hand actions also reveals different action types that they conducted. We computed the proportion of time with rapid and fast hand actions on objects and found that in the 3 less engaged conditions (50%/random, 10%/random, 10%/objects), participants spend more time on those rapid actions by moving objects dramatically from one location to another location ($M_{50\%/random}=15.18\%$; $M_{10\%/random}=17.30\%$; $M_{10\%/object}=16.54\%$) compared with the other two more engaged conditions ($M_{90\%}=10.31\%$; $M_{50\%/object}=9.91\%$). Interestingly, with the same 50% engaged level, the virtual agent’s random looking made participants generate more dramatic hand actions on objects with the attempt to attract the virtual agent’s attention. This again indicated that human participants were not only sensitive to at what moments the virtual agent was attending or not attending, but also where the virtual agent was attending.

6.3 Analyses of Temporal Dynamics of Multimodal Data

We focused on the moments right before and after the human participant was holding an object to measure the time-course of eye movement as a way to integrate gaze and hand data. The approach is based on what has been used in psycholinguistic studies to capture temporal profiles across a related class of

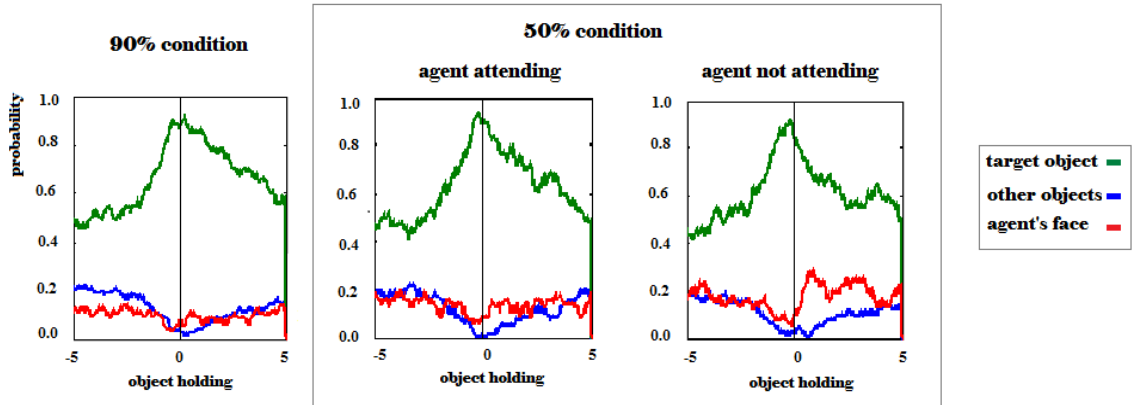


Figure 5: The proportion of time that participants were looking at the target object(green), the agent’s face(red) and the other two objects (blue). We selected three representative conditions, 90%(left), 50%/random(middle) when the virtual agent was attending to the target object held by participants, 50%/random(right) when the virtual agent was not attending to the target object.

events [12] (in our case, the relevant class of events is an object holding event). Such profiles enable one to discern potentially important temporal moments within a trajectory and compare temporal trends across trajectories. Figure 5 shows the average proportion of time across all the hand holding instances (which can be viewed a probability profile) that participants looked at the agent’s face (red), the target object (green), and the other two objects (blue). Each trajectory in a plot shows the probability that participants looked at one of three identities 5 sec before and after the holding moment. The left plot is from the 90% condition and the other two are from the 50%/random condition in which the virtual might or might not attend to the target object held by participants. Accordingly, we zoomed into that condition and divided all of the holding instances in that condition into two groups based on whether the virtual agent was attending (middle) to or not attending (right) to the target object. There were several interesting patterns: 1) participants increased their looking time on the target object and this looking behavior reached the peak before the holding action, suggesting a coordination of eye and hand movements to guide the reaching action. 2) At approximately the same moment with the decrease of looking at the target object, there was an increase of looking at the agent’s face in all of three plots indicating that humans checked whether the agent will be attending to that target object; 3) When the virtual agent was attending, the plot from the 90% condition (left) looks very similar with the one from the 50%/random condition (middle) ; 4) however, when the virtual agent was not attending (right), participants spent much more time on monitoring the agent’s attention while holding the target object. Those patterns illustrate dynamical behaviors of participants in real-time interaction. Most interestingly, the difference between their eye and hand coordination when the virtual agent was attending vs. not attending in the 50%/random condition supported the same conclusion from previous data analyses (e.g. Table 2), that is, participants dynamically adjust their behaviors based on their perception of real-time behaviors of the virtual agent but not on the overall engagement level of the agent.

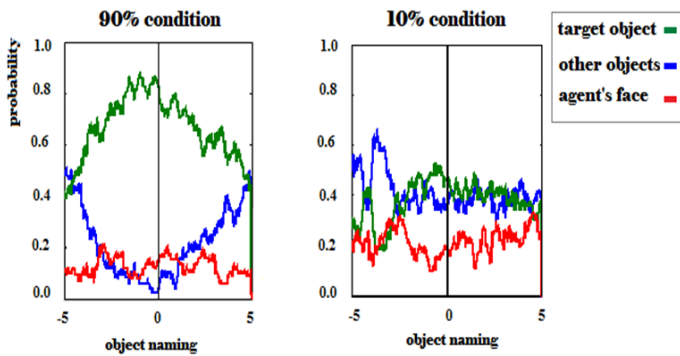


Figure 6: The proportion of time that participants were looking at the target object(green), the agent’s face(red) and the other two objects(blue). We selected two representative conditions, 90%(left) and 10%(right).

6.4 Analyses of Temporal Dynamics of Naming events

Since the experiment was designed to be a language learning task and participants were instructed to act as a teacher to teach a virtual agent object names, we next focus on naming utterances in speech when they mentioned object names. Across five conditions, participants mentioned those object names almost equally frequently within a learning trial ($M_{90\%}=9.31$; $M_{50\%/random}=10.32$; $M_{50\%/object}=8.68$; $M_{10\%/random}=9.25$; $M_{10\%/object}=10.15$). Thus, there is no significant difference in their speech naming acts alone. Next, we correlated those speech acts with the virtual agent’s attentional state and found that the approximately same number of naming events was distributed differently across experimental conditions. For example, more naming acts were generated when the virtual agent was looking straight toward participants. participants in both the 10%/random (36% of naming events) and 10%/object (33% of naming events) conditions, compared with the other three more engaged conditions ($M=26\%$). Since the virtual agent was most often not paying attention to the target object, participants in those conditions might realize that and therefore created more naming events at the moments that the virtual agent gazed at themselves by treating those moments as better learning moments compared with the moments that the agent just either randomly looked around or looked at one of the other objects.

Further, we extracted probabilistic profiles of participants’ gaze ROIs right before and after naming utterances start. Figure 6 shows those trajectories with a temporal window of 5 seconds before the onset of the naming utterance and 5 seconds after the onset of that naming utterance. In the 90% condition (left), participants paid more attention to an object around the moments they uttered the name of the object. This particular gaze pattern between target and other objects is in line with the results from psycholinguistic studies on the coupling between speech and gaze [13]. In the 10% condition, participants looked more at the virtual agent (red) which is expected as they wanted to maintain and monitor joint attention in face-to-face interaction. Surprisingly, the rest of their attention was more or less equally distributed between target and other objects. This suggests that participants perceived the agent’s random behaviors and attempted to adjust their own behaviors. By doing so, they failed to demonstrate typical behaviors that are well-documented in psycholinguistic studies on speech and simultaneous eye gaze. This is of importance for human-computer interactions and requires further investigation as current multimodal interfaces are likely to violate subtle temporal patterns and thus the time course of attention that humans expect. This finding also serves as justification for pursuing a temporally fine-grained multi-modal analysis of human joint attention processes in human-computer interactions.

7. Conclusion and Future Work

In multimodal human-avatar interaction, dependencies and interaction patterns between two interacting agents are bi-directional, i.e., the human user shapes the experiences and behaviors of the virtual agent through his own bodily actions and sensory-motor experiences, and the virtual agent likewise directly influences the sensorimotor experiences and actions of the human user. To understand the nature and fundamental principles in such multimodal interaction, we developed a real-time multimodal human-agent interaction system allowing us to collect and

analyze fine-grained behavioral data. Using this research platform, we discovered the following major findings from the joint attention experiment described earlier: 1) Human participants are sensitive to the virtual agent's behaviors and use them to infer the virtual agent's attentional state; 2) Accordingly, they adjust their own behaviors in various ways (e.g. more frequent looks on the virtual agent's face and more dramatic hand actions on objects). Importantly, in such multimodal interactions, their adaptive behaviors are not reflected by what they say, but instead they naturally generate more subtle non-verbal bodily cues to adjust their coordination with the virtual agent; 3) Their adaptive behaviors are not based on the overall engagement level of the virtual agent but momentary real-time behaviors of the virtual agent; 4) Time-course analyses of their eye movement patterns reveal momentary dynamic changes of their adaptive behaviors produced by real-time perception-action interactions between participants and the virtual agent. In summary, their behaviors seem to be composed of coordinated adjustments that happen on time scales of fractions of seconds and that are highly sensitive to the task context and to changing circumstances. All this suggests the importance of understanding real-time multimodal interaction in order to create smooth human-computer coordination and build better multimodal interfaces. Indeed, the experiment and the results reported here are the first steps toward this goal using the multimodal human-agent interaction platform we developed, among many other potential studies. For example, another application of this framework is to determine the timing of back-channel feedback (from eye gaze, to gestures, to body postures, to verbal acknowledgments), which is critical for establishing common ground in conversations. This could include questions about how head movements, gestures and bodily postures are related to natural language comprehension, as well as general questions about the functional role of non-linguistic aspects of communication contributing to natural language understanding. Thus, with this real-time interactive system, we can collect multimodal behavioral data in different contexts, allowing us to systematically study the time-course of multimodal behaviors. The results from such research will provide insightful principles to guide the design of human-computer interaction. Moreover, those fine-grained patterns and behaviors can also be directly implemented in an intelligent virtual agent who will demonstrate human-like sensitivities to various non-verbal bodily cues in natural interactions.

8. REFERENCES

- [1] Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. 2003. Predicting human interruptibility with sensors: a Wizard of Oz feasibility study. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Ft. Lauderdale, Florida, USA, April 05 - 10, 2003). CHI '03. ACM, New York, NY, 257-264.
- [2] Lee, J., Chai, J., Reitsma, P. S., Hodgins, J. K., and Pollard, N. S. 2002. Interactive control of avatars animated with human motion data. *ACM Trans. Graph.* 21, 3 (Jul. 2002), 491-500.
- [3] Ishii, R., Nakano, Y.I.: Estimating user's conversational engagement based on gaze behaviors. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 200-207.
- [4] Lee, J., Marsella, S., Traum, D., Gratch, J., Lance, B.: The rickel gaze model: A window on the mind of a virtual human. In: IVA '07: Proceedings of the 7th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2007) 296-303.
- [5] Ward, N.: Prosodic features which cue backchannel responses in English and Japanese. *Pragmatics* 32, 1177-1207 (2000)
- [6] Marsella, S.C., Morency, L.-P., Gratch, J., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R.J.: Virtual Rapport. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS(LNAI), vol. 4133, pp. 14-27. Springer, Heidelberg (2006)
- [7] Kipp, M., Gebhard, P.: Igaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 191-199.
- [8] Morency, L.P., Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 176-190.
- [9] Cassell, J.: Body Language: Lessons from the Near-Human. In: Riskin, J. (ed.) *The Sistine Gap: History and Philosophy of Artificial Life*, University of Chicago Press, Chicago.
- [10] Thorisson, K.R.: Modeling multimodal communication as a complex system. In: Wachsmuth, I., Knoblich, G. (eds.) *ZiF Research Group International Workshop. LNCS (LNAI)*, vol. 4930, pp. 143-168. Springer, Heidelberg (2008)
- [11] Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., & Stocksmeier, T. (2008): Modeling Embodied Feedback With Virtual Humans. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling embodied communication in agents and virtual humans*, LNAI 4930. Springer-Verlag.
- [12] Ou, J., Oh, L. M., Fussell, S. R., Blum, T., and Yang, J. 2005. Analyzing and predicting focus of attention in remote collaborative tasks. In Proceedings of the 7th international Conference on Multimodal interfaces (Toronto, Italy, October 04 - 06, 2005). ICMI '05. ACM, New York, NY, 116-123.
- [13] Z. Griffin, "Why look? Reasons for eye movements related to language production," *The interface of language, vision, and action: Eye movements and the visual world*, pp. 213-247, 2004
- [14] Yu, C., Zhong, Y., Smith, T., Park, I., Huang, W.: Visual data mining of multimedia data for social and behavioral studies. *Information Visualization* 8 (2009) 56-70.
- [15] Yu, C., Scheutz, M., Schermerhorn, P.: Investigating multimodal real-time patterns of joint attention in an hri word learning task. In: HRI '10: Proceeding of the 5th ACM/IEEE international conference on Human-agent interaction, New York, NY, USA, ACM (2010) 309-316.