Who, Me? How Virtual Agents Can Shape Conversational Footing in Virtual Reality

Tomislav Pejsa ✉, Michael Gleicher, and Bilge Mutlu

University of Wisconsin–Madison, USA
tpejsa@cs.wisc.edu

Abstract. The nonverbal behaviors of conversational partners reflect their conversational *footing*, signaling who in the group are the speakers, addressees, bystanders, and overhearers. Many applications of virtual reality (VR) will involve multiparty conversations with virtual agents and avatars of others where appropriate signaling of footing will be critical. In this paper, we introduce computational models of gaze and spatial orientation that a virtual agent can use to signal specific footing configurations. An evaluation of these models through a user study found that participants conformed to conversational roles signaled by the agent and contributed to the conversation more as addressees than as bystanders. We observed these effects in immersive VR, but not on a 2D display, suggesting an increased sensitivity to virtual agents' footing cues in VR-based interfaces.

Keywords: embodied conversational agents, virtual reality, gaze, orientation

1 Introduction

Many envisioned applications of virtual reality (VR) in games and social media involve multiparty conversations among avatar-mediated humans and virtual agents. In order to achieve natural and effective interactions, agents and avatars must produce humanlike nonverbal signals. In face-to-face social interactions, humans use nonverbal signals, such as spatial orientation and gaze, to regulate who is allowed to speak and to coordinate the production of speech utterances. Such signals help prevent misunderstandings, awkward silences, and people talking over one another. Conversational participants' nonverbal signals establish their roles—also known as *footing*— which determine their conversational behavior. Clear conversational roles are vital for smooth, effective multiparty interactions. However, there is a lack of computational models of footing-signaling behaviors for virtual agents as well as of studies that assess whether agents can use such behaviors to effectively shape the roles of human participants.

In this paper, we focus on two nonverbal signals of footing—spatial orientation and eye gaze—and introduce computational models of these behaviors. Building on prior work that has studied how humanlike robots can shape footing using their gaze [22], we develop a gaze model that generalizes to a wider variety of conversational scenarios and supplement it with a model that enables the agent to reconfigure the spatial configuration of the interaction. Our models are based upon the key insight that shifts in both spatial orientation and gaze can be realized as parametric variations of the same basic movement, allowing their integration in a single animation controller.

J. Beskow et al. (Eds.): IVA 2017, LNAI 10498, pp. 347-359, 2017.
DOI 10.1007/978-3-319-67401-8_45

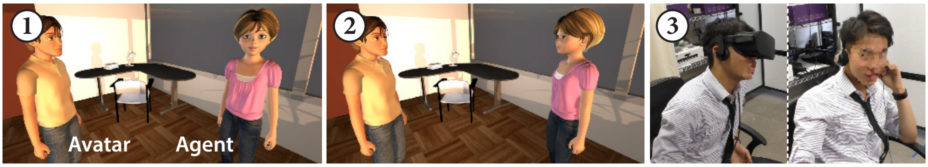


Fig. 1: (1–2) A virtual agent engaging in interaction with two participants (participant's view). The agent (1) puts participants on equal footing by distributing its gaze and body orientation evenly or (2) excludes a participant by looking and facing away from them. (3) Participants conform to their roles in virtual reality, but not on a 2D display.

We evaluate the effectiveness of the models in a study with human participants. Study results show that a virtual agent with appropriately designed gaze and spatial orientation cues can influence the footing of human participants, but only when using a VR display. We attribute this finding to wide field of view, stereo, and natural viewpoint control afforded by modern VR displays, which may enhance the effects of agent behaviors on the users of these systems.

2 Related Work

Our work focuses on two types of social signals in multiparty interaction: spatial orientation and gaze. Below, we summarize prior work on these signals from the social sciences and from research on virtual reality and embodied conversational agents (ECAs).

Human Communication — Conversational participants use nonverbal behaviors, particularly gaze and body orientation, to establish their conversational roles—what Goffman [13] has termed "footing." Participants use gaze to clarify who is being addressed [26], display attentiveness [15], and coordinate conversational turn-taking [18]. Speakers and addressees are the core participants, who make the majority of conversational contributions and spend most of the time gazing toward each other [22]. By contrast, bystanders make few conversational contributions and receive little gaze [22]. Spatial orientation is another footing cue; the core participants position and orient themselves in an "F-formation" [17], a spatial arrangement that creates a space between them to which they have equal, direct, and exclusive access and which excludes bystanders. When another participant joins the conversation, the core participants reorient themselves to include the newcomer in the F-formation.

Avatars in VR — A number of studies have investigated the effects of avatars' gaze and spatial positioning in VR. Avatars displaying gaze that matches their speech are attributed stronger presence and more positive traits [11]. When participants' eye gaze is accurately reproduced on their avatars, their gaze patterns match those observed in face-to-face conversations [27] and they also produce less speech [6], suggesting increased nonverbal communication. Studies [e.g., 7,29] have shown that participants in avatar-based communication tend to display compensatory interpersonal distance and gaze behaviors predicted by the Equilibrium Theory [4], even in a non-immersive setting [31]. Avatar spatial positioning produces similar effects; participants maintain greater distance from avatars who face them [8]. While these studies suggest that people are sensitive to gaze patterns and spatial orientation in VR, no prior work, to our knowledge, has assessed their ability to shape conversational footing in multiparty interactions.

Embodied Conversational Agents — Researchers have worked to endow virtual agents and robots with computational models of human conversational behaviors in order to increase their communicative capabilities. Well-designed gaze mechanisms on virtual agents have been shown to facilitate more efficient turn-taking [1,10,14], better management of engagement [9], and better recall of information [2]. Pedica et al. [23] have introduced a framework for automated generation of spatial positioning behaviors in virtual agents and found that interactions where agents employ such behaviors are viewed as more believable [24].

While no prior work has studied the ability of virtual agents to shape the footing of human participants, researchers have studied footing in the context of human-robot interaction. Mutlu et al. [22] have shown that participants conform to conversational roles signaled by a robot's gaze cues, while Kuzuoka et al. [19] have shown that a robot can reconfigure the conversational formation by reorienting its own body. These findings provide strong motivation for endowing virtual agents with equivalent capabilities.

3 Footing Behavior Models

Spatial orientation and eye gaze are key nonverbal cues that shape the footing of conversational participants. In this section, we describe the gaze controller used to synthesize gaze and body orientation shifts that comprise our footing behaviors. We then introduce the two models for synthesis of these behaviors. Finally, we give an overview of the models' prototype implementation in an embodied, multiparty dialog system.

3.1 Animating Gaze and Spatial Orientation Shifts

People shift their attention toward targets in their environment—objects, information, or other people—by performing coordinated movements of the eyes and head toward the target. In larger attention shifts, they may also shift their torso, while keeping their feet planted on the floor, or completely turn their body toward the target. Studies in

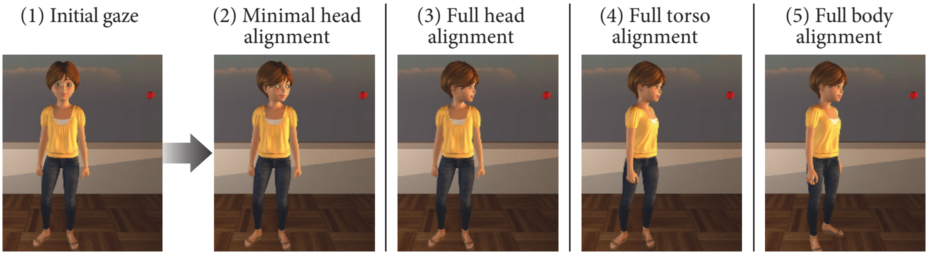


Fig. 2: Examples of synthesized gaze and body-orientation shifts: (1) eye contact with the observer, (2–5) gaze shifts with varying head, torso, and whole-body alignments.

neurophysiology [16,21,28] have found that eye, head, torso, and feet movements in attention shifts occur in tight coordination with one another and they display similar kinematic properties.

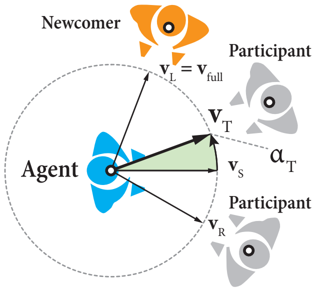
We have implemented a gaze animation controller for virtual agents that can synthesize coordinated movements of the eyes, head, torso, and whole body based on earlier work by Pejsa et al. [25]. In this model, the controller synthesizes a gaze movement toward a target in the environment by employing a set of kinematic laws derived from neurophysiological measurements of human gaze. A high-level model of conversational behavior provides the following parameters of each gaze shift to the controller: target position, $\mathbf{p}_T$, and head and torso alignment parameters, α_H and α_T. The alignment parameters control how much the head and torso participate in the gaze shift (Figure 2). For example, by setting $\alpha_H = 0$ and $\alpha_T = 0$, we can control the agent to gaze at the target out of the corner of its eye (Figure 2.2).

In this work, we extend the model proposed by Pejsa et al. [25] by adding support for whole-body orientation shifts that are required to change the agent's spatial orientation. We introduce a new parameter, α_B, which specifies how much the agent's lower body should turn toward the target. Setting $\alpha_B = 1$ results in the agent turning its whole body toward the target (Figure 2.5). We integrate the gaze controller with a custom turning controller, which replants the feet over the course of the gaze shift, resulting in a new spatial orientation of the agent. This extended model enables us to control the eye-gaze and body-orientation shifts of the virtual agent in a coordinated way and to signal the conversational footing of the agent.

3.2 Spatial-Reorientation Model

When two people interact, they typically face each other directly, creating a "vis-à-vis" arrangement, or stand at a 90° angle, forming an "L-shape" configuration [17]. Conversation between more than two participants generally occurs in a circular formation. Kendon [17] has coined the term "F-formation" to refer to these spatial arrangements of interacting participants. In order to correctly establish conversational footing, a virtual agent must maintain an F-formation with other participants. Specifically, when a new addressee approaches, the agent must turn toward the newcomer to reconfigure the F-formation. When a participant leaves, it may need to reorient itself toward the remaining participants. Below, we describe a model of body orientation, which achieves correct F-formation and utilizes our gaze controller to synthesize the required body movements.

When the first participant approaches, the agent performs a gaze shift toward the participant with the head, torso, and whole-body alignment parameters all set to 1 ($\alpha_H = \alpha_T = \alpha_B = 1$), facing the participant head-on. If the interaction already involves other participants, the agent must evenly distribute its body orientation among all the participants. To do so, we set $\alpha_H = 1$ and $\alpha_B = 1$ as before, whereas α_T is set such that the agent is oriented toward the midpoint between the leftmost and rightmost participant. To compute α_T, we project all direction vectors defined in Figure 3 onto the ground plane. The agent must realign its body such that its torso facing direction is $\mathbf{v}_T = slerp(\mathbf{v}_L, \mathbf{v}_R, 0.5)$ where *slerp* denotes spherical linear interpolation between two



$\mathbf{v}_S$	Current torso direction
$\mathbf{v}_T$	Target torso direction
$\mathbf{v}_{full}$	Torso direction for full alignment with newcomer
$\mathbf{v}_L$	Direction vector to the leftmost participant
$\mathbf{v}_R$	Direction vector to the rightmost participant
α_T	Realignment needed to accommodate newcomer

Fig. 3: Computing the torso alignment parameter α_T needed for the agent to reconfigure the F-formation when a new participant has joined the interaction.

direction vectors. The torso alignment α_T needed to achieve the facing direction $\mathbf{v}_T$ is:

$$\alpha_T = \frac{\angle(\mathbf{v}_S, \mathbf{v}_T)}{\angle(\mathbf{v}_S, \mathbf{v}_{\text{full}})} \tag{1}$$

This mechanism can be used to reestablish the F-formation when the leftmost or rightmost participant has departed or moved. In that case, using an updated α_T, the agent shifts its body orientation toward the participant at the opposite end of the formation.

3.3 Eye-Gaze Model

Conversational footing is also reflected in participants' gaze behavior. Speakers use gaze to indicate the addressees of the utterance or to release the floor to them, while the addressees gaze toward the speaker to display attentiveness. As a result, speakers and addressees gaze toward each other most of the time and only infrequently toward bystanders. Mutlu et al. [22] have calculated the gaze distributions of human speakers engaging in interactions involving addressees and bystanders. According to their data, speakers spend 26% of the time looking at each addressee's face (making eye contact) and avert their gaze toward the addressees' torsos and the environment other times in order to regulate intimacy [3]. When a second addressee is present, the amount of gaze toward each addressee's face remains around 26%, likely because switching gaze among the addressees now also achieves the purpose of intimacy regulation.

We build our footing gaze model based on the distributions reported by Mutlu et al. [22]. Our model defines a discrete probability distribution over the set of potential gaze targets, which includes the faces and torsos of all the addressees and bystanders as well as the environment. The distribution is characterized by the probability mass function, $p_T = p(T, N_A, N_B)$ (Table 1). The function p_T specifies the probability of looking toward the candidate target T given the current footing configuration, defined by the number of addressees, N_A, and the number of bystanders, N_B. In addition to the spatial distribution of the agent's gaze, our model specifies temporal durations of gaze fixations, shown in Table 1, defined as *gamma* distributions by prior work [22]. While the exponential distribution is commonly used to model events such as gaze shifts that occur at a constant average rate, we find the gamma distribution to more accurately

Table 1: Spatial probability distribution of the speaker's gaze and the agent's gaze fixation lengths (in seconds) toward possible targets in the given configuration of conversational roles. N_A is the number of addressees, while N_B is the number of bystanders.

	Spatial probability distributions		*Gaze fixation lengths*	
Gaze Target	**Footing Config.**	**Gaze Prob.**	**Footing Config.**	**Fixation Length**
Addressee face	$N_A = 1$	26%	$N_A = 1, N_B = 0$	*Gamma*(1.65, 0.56)
	$N_A \geq 2$	$54\%/N_A$	$N_A = 1, N_B = 1$	*Gamma*(0.74, 1.55)
			$N_A \geq 2$	*Gamma*(1.48, 1.10)
Addressee torso	$N_A = 1$	48%	$N_A = 1, N_B = 0$	*Gamma*(1.92, 0.84)
	$N_A \geq 2$	$16\%/N_A$	$N_A = 1, N_B = 1$	*Gamma*(1.72, 1.20)
			$N_A \geq 2$	*Gamma*(1.92, 0.52)
Bystander face	$N_B = 1$	5%	$N_B \geq 1$	*Gamma*(2.19, 0.44)
	$N_B \geq 2$	$8\%/N_B$		
Bystander torso	$N_B = 1$	3%	$N_B \geq 1$	*Gamma*(1.76, 0.57)
	$N_B \geq 2$	$5\%/N_B$		
Environment	$N_A = 1, N_B = 0$	26%	$N_A = 1, N_B = 0$	*Gamma*(0.90, 1.14)
	$N_A = 1, N_B = 1$	18%	$N_A = 1, N_B = 1$	*Gamma*(1.84, 0.59)
	$N_A = 1, N_B \geq 2$	13%	$N_A \geq 2$	*Gamma*(2.23, 0.41)
	$N_A \geq 2, N_B = 0$	30%		
	$N_A \geq 2, N_B = 1$	24%		
	$N_A \geq 2, N_B \geq 2$	17%		

represent human gaze, as it assigns a low probability to short fixations that are unlikely due to human motor limitations.

To illustrate the operation of our model, let us consider a scenario where the agent is speaking with two addressees ($N_A = 2$) named Alice and Bob, with two bystanders present ($N_B = 2$). To shift the agent's gaze, we draw from the spatial probability distributions to determine the target of the next gaze shift. According to Table 1, the probability of looking toward Alice's face is $54\%/2 = 27\%$ (Row 2). If Alice's face is the desired target, we supply the target to the gaze controller, which performs a gaze shift toward it. We hold the agent' gaze there for a duration determined by drawing from the distribution $Gamma(k = 1.48, \Phi = 1.10)$ (Table 1, Row 3). Alternatively, if the environment is the desired target, resulting in a gaze aversion, the direction of this shift can be computed using a supplemental model of conversational gaze aversion [e.g., 1,20].

Because our goal was to support a wide range of footing configurations, we extrapolated the data provided by Mutlu et al. [22] to derive probability distributions for configurations of three or more addressees and two or more bystanders.

3.4 System Design & Implementation

We implemented the footing behavior models within a high-level *behavior controller*, which controls the agent's gaze behavior and spatial orientation based on current dialog state. When an addressee joins or leaves the interaction, the behavior controller triggers a body orientation shift to reconfigure the conversational formation using the mechanism described in Section 3.2. While the agent is speaking or releasing the floor to addressees, the controller triggers gaze shifts based on the probabilistic model introduced in Section 3.3. Gaze shifts and body orientation shifts are synthesized by our gaze animation controller (Section 3.1) and rendered on the virtual embodiment.

Our prototype system is implemented in the Unity game engine. System components such as the behavior controller and gaze controller are implemented as Unity C# scripts. The system uses Microsoft Speech SDK to detect and recognize users' speech utterances and to synthesize the agent's speech. The visemes generated by the Speech SDK are used to animate the agent's lip movements.

4 Evaluation

To evaluate the models introduced above, we conducted a study with human participants aimed at answering two research questions: "Can a virtual agent use our models to shape the footing of participants in multiparty interactions in virtual reality?" and "Does the display type (e.g., VR or on-screen) influence these effects?" In the study, human participants engaged in a short conversation with a virtual agent and a simulated, avatar-embodied confederate that lasted 10 minutes. The virtual agent displayed gaze behaviors and spatial-orientation shifts that either included the participant as an addressee or excluded them as a bystander. We measured whether or not participants conformed to the conversational role signaled by our behavior models, for example, when assigned the role of bystander, by conversing less with the agent.

While our models are independent of input and display type, we expected their effects to be more salient in a virtual-reality setting. A VR application using a head-mounted display provides different affordances than a desktop display, such as a wide field of view and natural control of viewpoint using head tracking, which may strengthen the perception of social cues. Therefore, we expected people to be more sensitive to footing-signaling behaviors displayed by agents within an immersive VR environment. To test this prediction, our study also manipulated *display type*, comparing a VR headset (Oculus Rift CV1) and a desktop display.

4.1 Hypotheses

Based on prior work, we developed and tested the following hypotheses:

H.1 Participants will demonstrate conversational behavior that conforms to the footing signaled by the agent. Specifically, participants in addressee roles will speak more.

H.2 Participants in addressee roles will feel more groupness and closeness with the agent, as well as evaluate the agent more positively than those in bystander roles.

H.3 The agent's footing cues will have a stronger effect on the participants' conversational behavior (H.1) in VR than when using a 2D display.

H.4 The agent's footing cues will have a stronger effect on perceptions of the agent and feelings of closeness and groupness (H.2) in VR than in a 2D display.

H.1 is consistent with findings that conversing partners orient themselves in an F-formation [17] and that people look toward the addressees of their utterances [18]. H.2 is based on findings that people report negative feelings about a group and its members when being ignored or excluded [12].

H.3–4 are based on the premise that the improved affordances of a modern VR display will heighten awareness of the agent's nonverbal signals. Immersive VR blocks out external visual stimuli and creates a better sense of space due to stereoptic vision. Moreover, the Rift's high field of view (110°) affords a better view of the agent, confederate, and environment than the low-FOV camera settings typically utilized on 2D displays. Finally, the head tracking capabilities allow more intuitive control over the viewpoint than the traditional mouse-look interface, making it more intuitive and quicker for participants to reorient their viewpoint toward the agent during the interaction.

4.2 Study Design

The study followed a mixed, 2 × 2 factorial design, manipulating *agent behavior* (between-participants) and *task setting* (within-participants). Agent behavior was either *exclusive* or *inclusive*. In the exclusive condition, the agent displayed nonverbal behaviors that excluded the participant from the interaction, treating the participant as a bystander. It oriented its body toward the confederate and gazed toward the confederate much more. In the inclusive condition, the agent displayed nonverbal behaviors that included the participant in the interaction as an addressee. It distributed its body orientation evenly between the participant and the confederate and gazed toward them equally. Figure 4 illustrates these agent behaviors.

The conditions of the other independent variable, task setting, were either *2D display* or *VR*. Both conditions were designed to approximate the expected usage of interactive, virtual agent systems. In the 2D display condition, the participant experienced the interaction on a 27" Dell monitor, at 2560 × 1440 resolution and a field of view of 50°, and used the mouse to control the viewpoint. In the VR condition, the participant wore a VR headset (Oculus Rift CV1). The participant saw the scene at the resolution of 1080 × 1200 per eye and a 110° field of view. Built-in head-orientation tracking and the external positional tracker allowed the participant to control the viewpoint by moving the head.

Because the study had a within-participants factor (*task setting*), we implemented two versions of the task, described in the next section, to reduce transfer effects. The

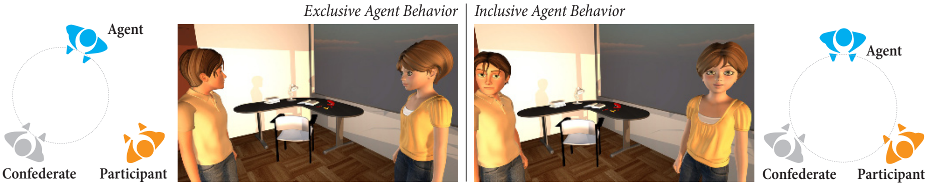


Fig. 4: Conditions of the *agent behavior* independent variable. Graphs show the conversational formations, and screenshots show the participant's view of the scene.

participants were assigned to conditions in a stratified order, counterbalanced with respect to task setting (*2D display* or *VR*) and task version (*Task 1* or *Task 2*).

4.3 Task & Procedure

The study task was a three-party, casual conversation between a virtual agent, the participant, and a "simulated confederate," which was a human-voiced agent producing prerecorded utterances in a virtual room. The participants were told that the confederate was a real human in another room. The task started with the participant standing at the entrance of the room with a view of the agent and confederate facing each other in a vis-à-vis formation. The participant was prompted to approach them by clicking a button. Depending on the agent behavior condition, the agent then either continued facing the confederate (*exclusive*) or reoriented herself toward the participant (*inclusive*). The agent then asked the participant and confederate casual, interview-style questions about themselves, such as "Where are you from?" or "What is your favorite movie about?" Most questions were implicitly addressed at both parties, giving them the choice to answer them. We expected participants to speak more if the agent used inclusive cues.

Implementation — The task was implemented in Unity, and the task logic and measurements were implemented as C# scripts. The confederate's lip movements were animated using Oculus Lip Sync. The character models for the agent and the confederate were imported from DAZ.[1] Both models had a looping, idle body motion applied to enhance the naturalness of their behavior.

Procedure — Following informed consent, participants were seated at a table with a PC in the study room. They received verbal task instructions and printed instructions to serve as a reminder. Participants then put on an audio headset (*2D display*) or Oculus Rift (*VR*). The experimenter launched the task application and left the room. Upon task completion, participants filled out a questionnaire. Next, participants performed a second trial of the task and filled out another questionnaire. Finally, participants received $5 USD as compensation. The procedure took approximately 30 minutes.

Participants — We recruited 32 participants (17 female and 15 male) through an online student job website and through in-person solicitation from the University of Wisconsin–Madison campus. All participants were students, and 27 of them were native English speakers.

4.4 Measures

The experiment involved two behavioral and several subjective measures. The behavioral measures were designed to capture the level of participation by the participant in the interaction to test H.1 and included the *number of speaking turns* taken over the course of the interaction and *total speaking time* in seconds. To alleviate acclimation effects, we excluded responses to the first five questions (out of twenty-five total).

The subjective measures were collected using a questionnaire consisting of seven-point scale items. To measure the agent's likeability, we asked participants to rate the agent on nine traits such as likeability, cuteness, and friendliness. From this data,

[1] DAZ Productions: http://www.daz3d.com/

we constructed a scale consisting of two factors: *likeability* (two items, Cronbach's $\alpha = 0.824$) and *attractiveness* (three items, Cronbach's $\alpha = 0.929$). Feelings of *closeness* were measured using a four-item scale (Cronbach's $\alpha = 0.823$) adapted from Aron et al. [5], who asked participants to indicate their agreement with statements such as "The agent paid attention to me." *Groupness* was measured using a seven-item scale adapted from Williams et al. [30] (Cronbach's $\alpha = 0.827$), which included statements such as "I felt ignored or excluded by the group." The questionnaire also included a check for the agent-behavior manipulation, implemented as a two-item scale, including "The agent faced me during the interaction" and "The agent faced the other participant."

4.5 Results

Our analysis began with averaging the two manipulation-check items and performing a one-way analysis of variance (ANOVA). We found a significant effect of agent behavior on the check, $F(1,61) = 5.743, p = .0196$.

Next, we analyzed our behavioral measures with a two-way, mixed-design ANOVA. We found a marginal effect of agent behavior on the *number of speaking turns* ($F(1,30) = 3.757, p = .062$) and no effect on the *total speaking time* ($F(1,30) = 1.159, p = .290$), providing partial support for H.1.

We found no interaction between agent behavior and task setting on the *number of speaking turns* ($F(1,30) = 2.217, p = .147$). However, we did find a marginal interaction between agent behavior and task setting on the *total speaking time* ($F(1,30) = 3.637, p = .066$). In the *VR* setting, participants took a significantly higher number of turns in the *inclusive* condition than in the *exclusive* condition ($M = 5.7$ versus $M = 8.4$), $F(1,55) = 5.970, p = .0178$). No such effect of agent behavior was found in the *2D display* setting, $F(1,55) = 0.465, p = .498$. An equivalent set of comparisons for speaking time found a marginal effect of agent behavior in the *VR* setting ($F(1,59) = 3.472, p = .0387$). No such effect of agent behavior was found in the *2D display* setting ($F(1,59) = 0.287, p = .594$). All the pairwise comparisons used a Bonferroni-corrected alpha level of .025. These findings provide support for H.3.

To analyze the subjective measures, we performed a two-way, mixed-design ANOVA. We found no effects of agent behavior on any of our subjective measures: *likeability*

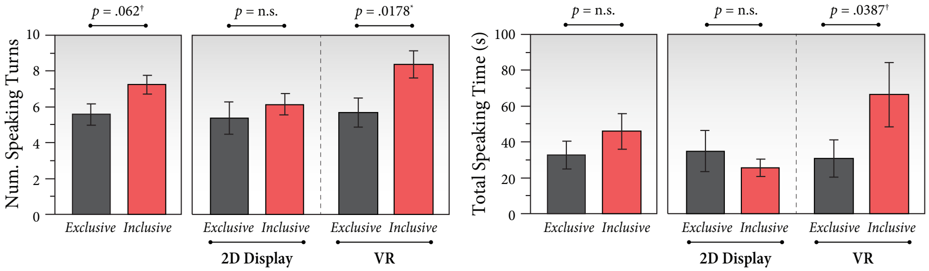


Fig. 5: Results from behavioral measures. Number of speaking turns (left) and total speaking time (right) by agent behavior (*exclusive* vs. *inclusive*) and task setting (*2D display* vs. *VR*). (∗) and (†) denote significant and marginal effects, respectively.

($F(1,30) = 0.363, p = .551$), *attractiveness* ($F(1,30) = 1.719, p = .200$), *closeness* ($F(1,30) = 0.606, p = .443$), or *groupness* ($F(1,30) = 0.140, p = .711$). We also found no significant interactions between agent behavior and task setting on any of the subjective measures: *likeability* ($F(1,30) = 1.233, p = .276$), *attractiveness* ($F(1,30) = 0.417, p = .243$), *closeness* ($F(1,30) = 0.092, p = .764$), *groupness* ($F(1,30) = 0.086$, $p = .771$). H.2 and H.4 were not supported by these results.

4.6 Discussion

The study results suggest that our models enable virtual agents to use gaze and spatial orientation to shape the conversational roles of human users, but only in an immersive VR setting. We found no evidence that these cues improve user experience. Some of the effects of our manipulations were not as strong as expected, possibly due to limitations of the dialog system and the agent's behaviors, reducing the overall realism and fluency of the interaction. Speech recognition lag, coupled with some participants' tendency to pause between utterances, occasionally caused our system to interpret speech gaps as floor releases and to cut participants off. This behavior of the system might have discouraged participants from speaking, limiting speaking to minimally sufficient responses and reducing the variability in the speaking-time measurement. Furthermore, the effects of the agent's footing signals might have been confounded by the minimal body animation of the agent and the confederate.

5 Limitations and Future Work

The behavior models introduced in this work, while effective at influencing conversational behavior in VR, are still rudimentary compared to behaviors observed in real-world multiparty interactions. These interactions are characterized by much greater variation in spatial arrangement of participants. Supporting such variation in virtual agents will require more sophisticated behavior models for spatial positioning and orienting. Moreover, gaze behaviors in human interactions demonstrate complex contingencies that are not adequately described by first-order statistical models. Variables such as interpersonal distance, discourse structure, personality, sex, and many others influence human gaze. Supporting such complexity will require more advanced models.

While our study suggests that a virtual agent's footing cues are only effective in immersive VR, it is unclear which aspects of VR support their effectiveness. We speculate that the high field of view, stereopsis, and head tracking all contribute, but further research is needed to understand the individual effects of these features and whether or not they affect social signals more generally. Future work may show that immersive, head-worn VR is not required to achieve believable multiparty interactions; for example, high FOV can be achieved with an ultra-wide curved display, while head tracking can be performed with an encumbrance-free device such as TrackIR.[2]

[2] TrackIR, http://www.naturalpoint.com/trackir/

6 Conclusion

In this paper, we have introduced computational models of gaze and spatial reorientation for virtual agents that enable them to signal conversational footing in multiparty interactions. An experimental evaluation has shown that participants interacting with the virtual agent conform to the conversational role signaled by the agent using our mechanisms, making more conversational contributions if they are treated as addressees than if they are treated as bystanders. However, this effect was observed only when participants interacted with the agent in a modern VR display and not a conventional 2D display. We speculate that this effect results from the immersion and more natural inputs that VR technology affords.

Designers can use the proposed models to give agents the ability to more effectively manage multiparty interactions. This work also provides impetus for further research on behavioral mechanisms that allow such interactions to proceed smoothly and effectively. As the new generation of VR devices becomes more widely adopted, more nuanced multiparty interactions could become an integral part of social experiences from online games to virtual worlds. The current work represents a stepping stone toward understanding and implementing such interactions.

Acknowledgements

This research was supported by the National Science Foundation awards 1017952 and 1208632. We thank Faye Golden for her help in conducting the experiment.

References

1. Andrist, S., Mutlu, B., Gleicher, M.: Conversational gaze aversion for virtual agents. In: Proc. IVA'13, vol. 8108, pp. 249–262 (2013)
2. Andrist, S., Pejsa, T., Mutlu, B., Gleicher, M.: Designing effective gaze mechanisms for virtual agents. In: Proc. CHI'12. pp. 705–714 (2012)
3. Argyle, M., Cook, M.: Gaze and mutual gaze. Cambridge University Press (1976)
4. Argyle, M., Dean, J.: Eye-contact, distance and affiliation. Sociometry 28, 289–304 (1965)
5. Aron, A., Aron, E.N., Smollan, D.: Inclusion of other in the self scale and the structure of interpersonal closeness. Journal of Personality and Social Psychology 63(4), 596 (1992)
6. Bailenson, J.N., Beall, A.C., Blascovich, J.: Gaze and task performance in shared virtual environments. The Journal of Visualization and Computer Animation 13(5), 313–320 (2002)
7. Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence 10(6), 583–598 (2001)
8. Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: Interpersonal distance in immersive virtual environments. Personality and Social Psychology Bulletin 29(7), 819–833 (2003)
9. Bohus, D., Horvitz, E.: Models for multiparty engagement in open-world dialog. In: Proc. SIGDIAL'09. pp. 225–234 (2009)
10. Bohus, D., Horvitz, E.: Multiparty turn taking in situated dialog: Study, lessons, and directions. In: Proc. SIGDIAL'11. pp. 98–109 (2011)
11. Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., Sasse, M.A.: The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In: Proc. CHI'03. pp. 529–536 (2003)
12. Geller, D.M., Goodstein, L., Silver, M., Sternberg, W.C.: On being ignored: The effects of the violation of implicit rules of social interaction. Sociometry pp. 541–556 (1974)

13. Goffman, E.: Footing. Semiotica 25(1–2), 1–30 (1979)
14. Heylen, D., Van Es, I., Van Dijk, E., Nijholt, A., Van Dijk, B.: Experimenting with the gaze of a conversational agent. In: Proc. CLASS'05. pp. 93–100 (2002)
15. Heylen, D.K.J.: Head gestures, gaze and the principles of conversational structure. International Journal of Humanoid Robotics 3(3), 241–267 (2006)
16. Hollands, M.A., Ziavra, N.V., Bronstein, A.M.: A new paradigm to investigate the roles of head and eye movements in the coordination of whole-body movements. Experimental Brain Research 154(2), 261–266 (2004)
17. Kendon, A.: Conducting interaction: Patterns of behavior in focused encounters. Cambridge University Press (1990)
18. Kendon, A.: Some functions of gaze-direction in social interaction. Acta psychologica 26, 22–63 (1967)
19. Kuzuoka, H., Suzuki, Y., Yamashita, J., Yamazaki, K.: Reconfiguring spatial formation arrangement by robot body orientation. In: Proc. HRI'10. pp. 285–292 (2010)
20. Lee, S.P., Badler, J.B., Badler, N.I.: Eyes alive. ACM ToG 21, 637–644 (2002)
21. McCluskey, M.K., Cullen, K.E.: Eye, head, and body coordination during large gaze shifts in rhesus monkeys: Movement kinematics and the influence of posture. Journal of Neurophysiology 97(4), 2976–2991 (2007)
22. Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., Ishiguro, H.: Conversational gaze mechanisms for humanlike robots. ACM TiiS 1(2), 12:1–12:33 (2012)
23. Pedica, C., Vilhjálmsson, H.H.: Spontaneous avatar behavior for human territoriality. Applied Artificial Intelligence 24(6), 575–593 (2010)
24. Pedica, C., Vilhjálmsson, H.H., Lárusdóttir, M.K.: Avatars in conversation: The importance of simulating territorial behavior. In: Proc. IVA'10. pp. 336–342 (2010)
25. Pejsa, T., Andrist, S., Gleicher, M., Mutlu, B.: Gaze and attention management for embodied conversational agents. ACM TiiS 5(1), 3:1–3:34 (2015)
26. Schegloff, E.A.: Sequencing in conversational openings. American anthropologist 70(6), 1075–1095 (1968)
27. Steptoe, W., Wolff, R., Murgia, A., Guimaraes, E., Rae, J., Sharkey, P., Roberts, D., Steed, A.: Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments. In: Proc. CSCW'08. pp. 197–200 (2008)
28. Uemura, T., Arai, Y., Shimazaki, C.: Eye-head coordination during lateral gaze in normal subjects. Acta Oto-Laryngologica 90(3–4), 191–198 (1980)
29. Wieser, M.J., Pauli, P., Grosseibl, M., Molzow, I., Mühlberger, A.: Virtual social interactions in social anxiety-the impact of sex, gaze, and interpersonal distance. Cyberpsychology, Behavior, and Social Networking 13(5), 547–554 (2010)
30. Williams, K.D., Cheung, C.K.T., Choi, W.: Cyberostracism: effects of being ignored over the internet. Journal of Personality and Social Psychology 79(5), 748 (2000)
31. Yee, N., Bailenson, J.N., Urbanek, M., Chang, F., Merget, D.: The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. CyberPsychology & Behavior 10(1), 115–121 (2007)