Visual Attention and Eye Gaze during Multiparty Conversations with Distractions

Erdan Gu , Norman I. Badler

Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 19104-6389
{erdan, badler}@seas.upenn.edu

Abstract. Our objective is to develop a computational model to predict visual attention behavior for an embodied conversational agent. During interpersonal interaction, gaze provides signal feedback and directs conversation flow. Simultaneously, in a dynamic environment, gaze also directs attention to peripheral movements. An embodied conversational agent should therefore employ social gaze not only for interpersonal interaction but also to possess human attention attributes so that its eyes and facial expression portray and convey appropriate distraction and engagement behaviors.

1. Introduction

In order to build a plausible virtual human or embodied conversational agent (ECA), we must understand how it might be given a cognitive ability to perceive, react and interact with the environment [2]. Conventional ECA animation techniques fall short of providing agents with human-like responses to environmental stimuli and internal goals, principally because they endow the agent with perfect cognition. There are, however, many intricate shortcomings to real human perception. Our work seeks to address and rectify these problems by seeking insights from cognitive psychology to model aspects of human vision, memory and attention.

An ECA should also be equipped to perceive and express many non-linguistic social signals to communicate information in a shared environment. Eyes direct attention, expose the actual mood of the subject, and express a wide range of human expressions [14]. For example, the amount of eye opening can reflect various emotional states, the blinking rate decreases when a person is attentive to objects in the environment, and gaze provides an important cue to regulate conversations [15].

People focus on eyes to "read" insights into human behavior. Natural gaze behavior is critical to the realism and believability of an animated character. An ECA should employ social gaze for interpersonal interaction and also possess human attention attributes so that its eyes and facial expression convey appropriate distraction and attending behaviors. Our objective is to develop a computational model of multiple influences on eye gaze behavior for an ECA in a dynamic environment. Eye behaviors should be influenced by human-like imperfect cognitive ability, social aspects of interaction behaviors, as well as some internal cognitive states. Our work here makes

two contributions: constructing a social gaze model for multiparty conversation and observing its behavior and consequences under varying environmental distractions, conversation workload, and participant engagement.

The paper is organized as follows. Section 2 describes relevant studies on ECA gaze behavior in order to situate this work within the current state of the art. Section 3 presents a comprehensive eye movement model for conversational and emotive gaze. Section 4 concentrates on the turn-allocation strategy in multiparty conversation and associated gaze behaviors. Section 5 examines an experiment with varying external distractions and internal workload for the agent, who then exhibits appropriate gaze behavior. Section 6 concludes with a discussion and future work.

2. Background

There have been several attempts to model the role of gaze in ECAs. Gaze, combined with gesture, facial expression and body orientation all give information about what we are saying and thinking, and help (perhaps unconsciously) to communicate emotions. Eye movement is heavily related to information processing in the brain. Lee *et al*. [16] exploited an eye saccade statistical model during talking and listening based on empirical eye tracking data. In our work, we explore emotive gaze to expose mood and thought processes. We do not present here specific speech-relevant gaze behaviors which synchronize to verbal communicative acts but rather consider the correlation between eye motor control and general cognitive activity.

Directional gaze cues are frequently present to communicate the nature of the interpersonal relationship in face-to-face interactions [1]. It is estimated that 60% of conversation involves gaze and 30% involves mutual gaze [24]. Garau *et al*. [8] and Colburn *et al.* [7] analyze frequencies of mutual gaze to simulate patterns of eye gaze for the participants. Social gaze serves to regulate conversation flow. Cassell *et al.* [4] use eye gaze as a sign to open and close the communication channel. Novick *et al.* [22] observe two simple gaze patterns (mutual-break and mutual-hold) to account for much of the turn-taking behavior. So far, however, ECA simulations for face-face conversation are mainly dyadic and turn allocation using gaze signals is relatively simple. Multiparty turn-taking behavior is an open challenge and some attempts [28] [29] are based largely on the dyadic situation. Much of this work focuses on user-perceptual issues or has involved mediated communications rather than ECA simulation. Intuitively, a significant difference exists in gaze behaviors between dyadic and multiparty situations: at the minimum the latter must include mechanisms for turn-requests, acknowledgement, and attention capture. We address the role of gaze in turn-taking allocation strategy, appearance of awareness, and expression of the feedback signal.

Ideally, we would like to implement the ECAs such that they interact with their conversational partners and environment in the same way as real people do by having a limited visual resource. Suppressed or inappropriate eye movements damage the experienced effectiveness of an ECA. Gaze behavior should be emergent and responsive to a dynamic environment. Engagement is a key factor that underlies realistic human-like cognitive commitment. Sidner *et al*. [27] define it as “the process by

which two or more participants establish, maintain and end their perceived connection during interactions they jointly undertake." We construct a framework to decide engagement due to the demands of simultaneously executing interpersonal tasks and managing exogenous stimuli and, consequently, to predict gaze behavior.

In our recent work [10], we suggested a visual attention model that integrates both bottom-up and top-down filters, and combines 2D snapshots of the scene with 3D structural information. While it is commonly believed that an object requires only reasonable physical (perceivable) properties to be noticed in a scene, recent studies [17] [20] have found that people often miss very visible objects when they are preoccupied with an attentionally demanding task. Green [9] classifies the prominent inadequacies in visual processing into four categories: (sensory and cognitive) conspicuity, mental workload, expectation, and capacity. Based on this descriptive model, we formulated a computational framework that determines successful attention allocation and consequent inattention blindness [11]. In our preliminary investigations, we quantified our model with a computational experiment analogous to other inattention blindness studies [26] and examined the effects of Green's four factors on the subject's awareness level of the unattended object. Here we employ the same model for ECAs and examine some of the most important parameters. The ECA interactions are affected by each other as well as unexpected events in the external environment. The attention model of the ECA decides what should or should not be permitted into consciousness. The ECA may or may not be aware of peripheral movements according to different engagement levels. Our approach attempts to leverage multiple influential accounts from external visual stimuli and social interaction into a computational model that drives consistent ECA gaze animation.

3. Computational Model for Eye Motor Control

The human repertoire of eye motor control can be defined by saccade, fixation, smooth pursuit, squint and blink. There are parameters to describe these ocular movements [16] including gaze direction, magnitude, duration, the degree of eye open, blink, and so on. The magnitude defines the angle the eyeball rotates, while velocity differentiates smooth pursuit and saccade. Duration is the amount of time that the movement takes to execute. Our attention model affects eye motor control by specifying the gaze direction, the degree of eye open, and size of pupil relative to luminance.

There are many eye-related communicative functions. Here we focus on directional gaze patterns such as eye contact, mutual gaze, gaze aversion, line of regard, and fixation. Two participants use mutual gaze to look at each other, usually in the face region. Gaze contact means they look in each other's eyes. In gaze aversion, one participant looks away when others are looking toward her. Head rotation and nod or shake are always linked to eye movement [5]. Head and eyes continuously align with a moving target. Horizontal gaze shifts greater than 25° or vertical shifts greater than 10° produce combined head and eye movement [6]. Once the head is aligned with the target, the eyes re-center.

In addition, various eye movements accompany a wide range of human expressions. People generally partially close their eyes during unpleasant emotions to reduce

vision, but react to happiness by spreading. Table 1 summarizes eye movement patterns in different emotion expressions [14].

Type	Face/Eye Behavior Description	Eye Movement
Laughter	Submissive: apprehension around the eyes.	Downcast gaze; decreased eye contacts
	Smile: relaxed, teeth together but lips are barely parted.	Flat gaze
	Laughter: teeth often parted, partially covered by the lips.	Upraised and out-of-focus gaze; eyes wide open
Surprise	Sudden opening of the eyes followed by mixed emotions: pleasant, anger, shock	Fixation and up to mixed emotions
Fear	Similar to surprised	Eye fixation or aversion
Interest	Eyes wide open (object is close) or squint (great distance), fixed on the object	Fixation, scan with longer glances
Anger	Eyes wide open and fixed; face a rigid mask	Fixation
Contempt	Eyes are a little closed, wrinkles under the eyes, but fixed on insignificant object	Eyes looking sideways
Disgust	Upper eyelids may be partially closed, or raised slightly on one side.	Eye aversion

Table 1: Emotion state and corresponding eye movement patterns.

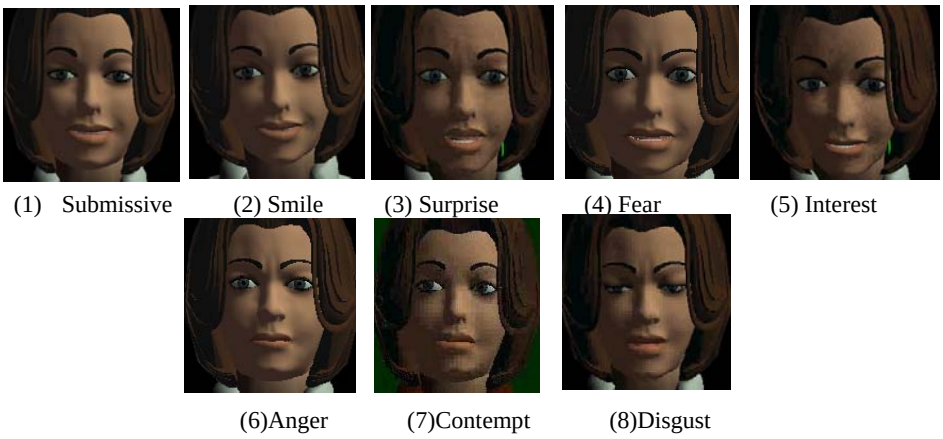


Fig. 1: Examples of eye movements accompanying a wide range of human expressions: (1)Submissive (2)Smile (3)Surprise (4)Fear (5)Interest (6)Anger (7)Contempt (8)Disgust

We are constructing a comprehensive eye model from low level eye motor control to high level gaze patterns exhibited by conversational gaze, emotional state, and visual attention. Conversational gaze as a turn-taking signal is elaborated below.

4. Gaze Roles in Turn-Taking

Gaze behaviors and visual contact signal and monitor the initiation, maintenance and termination of communicative messages [3]. Short mutual gaze (~1s.) is a powerful mechanism that induces arousal in the other participants [1]. Gaze diminishes when disavowing social contact. By avoiding eye gaze in an apparently natural way, an audience expresses an unwillingness to speak.

Conversation proceeds in turns. Two mutually exclusive states are posited for each participant: the speaker who claims the speaking turn and the audience who does not. Gaze provides turn-taking signals to regulate the flow of communication. Table 2 shows how gaze behaviors act to maintain and regulate multiparty conversations.

State	Signals	Gaze Behavior
Speaker	Turn yielding	Look toward listener
	Turn claiming suppression signal	Avert gaze contact from audience
	Within turn signal	Look toward audience
	No turn signal	Look away
Audiences	Back channel signal	Look toward speaker
	Turn claiming signal	Seek gaze contact from speaker
	Turn suppression signal	Avert gaze contact from speaker
	Turn claiming suppression signal	Look toward other aspiring audiences to prevent them speaking
	No response	Random

Table 2: Turn-taking and associated gaze behaviors

In dyadic conversation, at the completion of an utterance or thought unit the speaker gives a lengthy glance to the audience to yield a speaking turn. This gaze cue persists until the audience assumes the speaking role. The multiparty case requires a turn-allocation strategy. Inspired by Miller [19], we address the multi-party issue with two mechanisms: a *transition-space* where the speaker selects the next speaker and a *competition space* where the next turn is allocated by self-selection.

Transition Space (Fig. 2(2))
Speaker:
1: She gives a lengthy glance (turn yielding) to one of the audiences.
2.i: Receiving gaze contact (turn claiming) from the audience, the speaker relinquishes the floor.
2.ii: Receiving gaze aversion (turn suppression) from the audience, the speaker decides to keep transition-space to find another audience or go to competition space directly. If no one wants to speak, the speaker has the option of continuing or halting.
Audiences:
1: Audience who wants a turn will look toward speaker's eye to signal her desire to speak (turn claiming), and want to draw the attention of the speaker.

2: Audience receiving speaker gaze (turn yielding) uses quick gaze contact (turn claiming) to accept the turn or lengthy gaze aversion (turn suppression) to reject it.

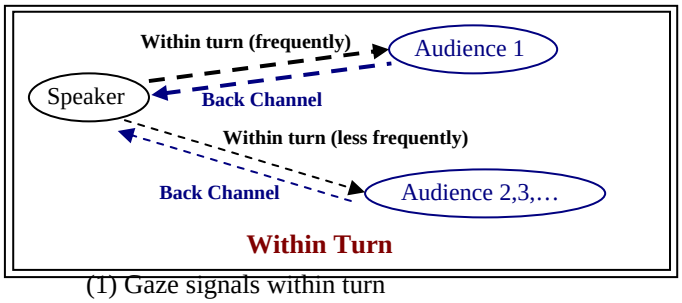
Competition Space (Fig. 2(3))

Speaker:

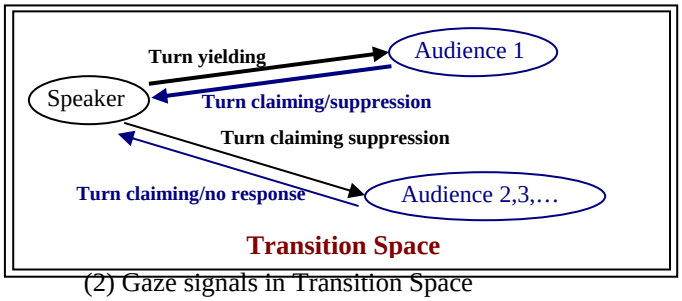
She scans all the audiences, serially sending a turn yielding signal (Fig. 3).

Audiences:

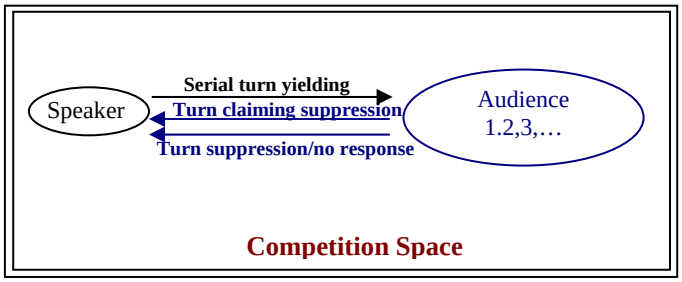
They may have eye interactions at that time. The aspiring audience looks towards the speaker to signal a desire to speak (turn claiming). After receiving visual contact from the speaker, she looks at all the other aspiring audiences to signal her taking the floor (turn claiming suppression). Non-aspiring audiences may follow the speaker's gaze direction or use random gaze (no response).



(1) Gaze signals within turn



(2) Gaze signals in Transition Space



(3) Gaze signals in Competition Space

Fig. 2: Diagram for turn taking allocation and employed conversational gaze signal

Turns begin and end smoothly, with short lapses of time in between. Occasionally an audience's turn-claim in the absence of a speaker's turn signal results in simultaneous turns [14] between audiences, even between audience and speaker. Favorable simultaneous turns will occur that show it is a comfortable and communicative circum-

stance. The general rule is that the first speaker continues and the others drop out. The dropouts lower gaze or avert gaze to signal giving up.

(1) (2) (3)

Fig. 3: A four-party conversation. (1) Full view image. (2)(3) Turn yielding gaze signal.

Within a turn, audiences spend more time looking toward the speaker (back channel) to signal attention and interest. They focus on the speaker's face area around the eyes. The speaker generally looks less often at audiences except to monitor their acceptance and understanding (within turn signal). The speaker glances during grammatical breaks, at the end of a thought unit or idea, and at the end of the utterance to obtain feedback. As Fig. 2(1) shows, the speaker usually assigns a longer glance to the audience to whom she would like pass the floor.

5. Engagement Level of Conversational Agent

Eye gaze is fundamental in showing interest levels between characters and as a means of anticipating events. When audiences looked at their partner less than normal, the audiences were rated as less attentive [24]. Thus, the duration and frequency of glances directed towards the speaker will be considered indicative of the audience's attentive level. Peters *et al.* [23][24] present an ECA model with the capability of visual perceiving another's level of interest based on direction of the eyes, head, body and locomotion. After being aware of such signals, the speaking agent has the option to continue or stop talking. Both speaker and audience are also influenced by what happens in the external environment. While attending to the conversational partner is the most basic form of signaling understanding by the agent [21], an audience whose eyes never waver from her partner, despite background events, appears lifeless. Therefore, an ECA with a realistic attention system can use perceptual information to project more realistic involvement in conversation.

We discuss two types of engagement behaviors: engagement cues from a conversational partner or herself, and those from the environment. We apply our attention framework to determine attention shifts between these two cues. The speaker determines the arousal or discouragement of talking by perceiving visual contact from the audiences or distractions from any peripheral movements. In the

remainder of this section we study these attention effects, particularly the transition from self/partner to the environment. In our system we can experimentally adjust several influence parameters, such as mental workload of participants and conspicuity of distraction.

5.1 Parameterized Experiment

Because human cognitive resources are limited, attention acts as a filter to examine sensory input quickly and limit cognitive processing. We endow the ECA with a human-like perceptual ability to automatically decide to maintain or halt the conversation. Sensory conspicuity refers to the bottom-up properties of an object, while cognitive conspicuity reflects the personal or social relevance it contains [30]. As tasks become more difficult they increase the mental workload of the subject and require more attention, increasing the likelihood of missing an unexpected event. Thus workload and conspicuity are related more to the visual system while expectation and capacity appear closer to other cognitive structures such as memory.

Our attention model relies on the cooperation of internally-driven top-down settings and external bottom-up inputs. The bottom-up input uses the "saliency" (sensory conspicuity features) of objects in the scene to filter perceptual information and compute an objective saliency map. Primary visual features consist of 2D and 3D visual cues relevant to the object, such as its size, depth (distance from the agent to the object), location in the agent's view image (how far from focus center to the object), color and movement speed. Simultaneously, top-down settings, such as expectation and face pop-out, determine the set of items that are contextually important. Known as the attentional set, this is a subjective feature pool of task-prominent properties maintained in memory. At any moment, focused attention only provides a spatio-temporal coherence map for one object [11]. This coherence map highlights the object calculated to be the most important at that moment in the scene, and thus can be used to drive the ECA's gaze.

The appearance and movement of an unexpected object in the scene were varied in order to affect sensory and cognitive conspicuity level. The inherent physical salience value of the unexpected object could be *high*, *medium*, or *low*. We used three objects: one falling red cube outside a window, one big green cube moving on the table, and one man who suddenly appears outside the window. The possible field of vision of the agent is considered. In the third object case, face pop-out detection reveals a man in the agent's visual field; since faces as socially relevant features are meaningful to a person they are more likely to capture attention. As Fig. 4 shows, the speaker exhibits different responses to different peripheral movements.

Fig. 4: Adjustment of conspicuity level by varying different distractions. (1) Full view of four party conversation. (2) Red falling cube with low conspicuity level goes unnoticed. (3) Green floating cube grabs attention and causes speaker engagement shift from the partner to the external stimuli; she does smooth pursuit to track the movement. (4) The speaker is surprised since the man's face makes the speaker immediately consciousness of him.

(1) (2) (3)

Fig. 5: Adjustment of mental workload level by adding more parties. (1) Conversation with three participants. (2) Conversation with four participants. (1) Conversation with five participants.

In the second variation, mental workload could be *high*, *medium*, or *low*, determined by the intensity level of the conversation (Fig. 5). Difficulty increases as parties are added to the interaction. The speaker's mental workload will be high when she wants to maintain an active atmosphere with more than four participants. Simultaneously, more frequent turn exchanges with more participants enhance the arousal of the speaker to maintain the conversation. The interest level of the audience, reflected in the frequency of back channel signals, also augments their involvement. They all occupy considerable attention for the participants and reduce the probability of attention shift. In the highest workload case, we place five participants and four turn exchanges in a 2-minute conversation. The speaker pays no attention to any unexpected objects: not even the human face pop-out although it falls into her line of vision.

6. Conclusion

Our contribution lies in building convincing computational models of human gaze behavior grounded in cognitive psychological principles. To interact with humans in a shared environment, an ECA must posses an analog of human visual attention, visual limitations, and non-linguistic social signals. This model can improve social acceptability and interpersonal interactions between people and animated human agents in diverse applications. These applications include tutoring, teaching, training, web agents, movie special effects, and game characters.

In the future, we aim to further integrate the internal state of the ECA such as emotion, personality and mental states with eye gaze, head motion and facial animation. Appropriate eye movements increase the realism of an agent's engagement behavior. Computational eye gaze models will allow us to explore other inattentional blindness factors, such as expectation and capacity. In addition, experimentally supported quantification and model validation engaging human and synthetic participants in shared spaces is required. Human subjects should be asked to empirically evaluate the naturalness and effectiveness of the animated nonverbal behavior of the ECAs during real-time interactions.

Acknowledgements

This work is partially supported by NSF IIS-0200983 and NASA 03-OBPR-01-0000-0147. Opinions expressed here are those of the authors and not of the sponsoring agencies. Thanks to Catherine Pelachaud for use of the Greta agent and to Jan M. Allbeck for her assistance.

References

1. Argyle, M. and Cook, M. (1976) *Gaze and Mutual Gaze*. Cambridge University Press, London.
2. Badler, N., Chi, D. and Chopra S. (1999) Virtual human animation based on movement observation and cognitive behavior models. In *Proc. Computer Animation*, Geneva, Switzerland, IEEE Computer Society Press, pp. 128-137
3. Cassell, J. and Thorisson, K. (2000) The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence Journal*, 13(4-5):519-538**.**
4. Cassell, J. and Vilhjalmsson, H. (1999), Fully embodied conversational avatars: Making communicative behaviors autonomous. *Autonomous Agents and Multi-Agent Systems*, 2(1): 45-64. .
5. Chopra-Khullar, S. and Badler, N. (2001) Where to look? Automating attending behaviors of virtual human characters. *Autonomous Agents and Multi-agent Systems* 4, pp. 9-23.
6. Chopra-Khullar, S. (1999) Where to look? Automating certain visual attending behaviors of human characters. Ph.D Dissertation, University of Pennsylvania.
7. Colburn, A., Cohen, M. and Drucker, S. (2000) The role of eye gaze in avatar mediated conversational interfaces.. MSR-TR-2000-81. Microsoft Research, 2000.
8. Garau, M., Slater, M., Bee, S. and Sasse, M. (2001) The impact of eye gaze on communication using humaniod avatars. *Proc. ACM SIGCHI*, pp. 309-316.
9. Green, G. (2004) Inattentional blindness and conspicuity. Retrieved November 10, http://www.visualexpert.com/Resources/inattentionalblindness.html
10. Gu, E., (2005) Multiple Influences on Gaze and Attention Behavior for Embodied Agent, Doctoral Dissertation Proposal, Nov., 2005, Computer and Information Science Department, Univeristy of Pennysylavania.
11. Gu, E., Stocker, C. and Badler, N. (2005) Do you see what eyes see? Implementing inattention blindness. *Proc. Intelligent Virtual Agents* 2005, LNAI 3661, pp 178-190.
12. Gu, E., Wang, J. and Badler, N. (2005). Generating sequence of eye fixations using decision-theoretic bottom-up attention model. 3rd International Workshop on Attention and Performance in Computational Vision, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop, San Diego, pp92.
13. Itti, L. (2003) Visual attention. *The Handbook of Brain Theory and Neural Networks*, Cambridge, Michael A. Arbib (Editor), MIT Press. pp. 1196–1201.
14. Knapp, L. and Hall, A. (1996) The effects of eye behavior on human communication. *Nonverbal communication in human interaction*. Harcourt College Pub., 4th edition, Chapter 10, pp. 369-380.
15. Kendon, A. (1967) Some functions of gaze direction in social interaction. *Acta Psychologica*, 32, pp. 1–25.
16. Lee, S., Badler, J. and Badler, N. (2002) Eyes alive, *ACM Transactions on Graphics* 21(3), pp. 637-644.

17. Mack, A. and Rock, I. (1998) *Inattentional Blindness*. Cambridge, MA, MIT Press.
18. Matsusaka, Y., Fujie, S. and Kobayashi, T. (2001) Modeling of conversational strategy for the robot participating in the group conversation. *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, pp. 2173-2176.
19. Miller, E. (1999) Turn-taking and relevance in conversation. For the course, *Ways of Speaking*, at the University of Pennsylvania, May 1999.
20. Most, S., Scholl, B., Clifford, E. and Simons, D. (2005) What you see is what you set: Sustained inattentional blindness and the capture of awareness. *Psychological Review* 112, pp. 217-242.
21. Nakano, Y. and Nishida, T. (2005) Awareness of perceived world and conversational engagement by conversational agents. AISB Symposium: Conversational Informatics for Supporting Social Intelligence & Interaction, England.
22. Novick, D., Hansen, B. and Ward, K. (1996) Coordinating turn-taking with gaze. *Proc. of ICSLP-96*, Philadelphia, PA, pp. 1888-1891.
23. Peters, C. (2005) Direction of attention perception for conversation initiation in virtual environments. *Proc. Intelligent Virtual Agents*, pp. 215-228.
24. Pelachaud, C., Peters, C., Mancini, M., Bevacqua, E. and Poggi, I. (2005) A model of attention and interest using gaze behavior. *Proc. Intelligent Virtual Agents*, pp. 229-240.
25. Slater, M., Pertaub, D. and Steed, A. (1999) Public Speaking in Virtual Reality: Facing and Audience of Avatars, IEEE Computer Graphics and Applications, 19(2), March/April 1999, p6-9
26. Simons, D. and Chabris, C. (1999) Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception* 28, pp. 1059-1074.
27. Sidner, C., Lee, C. and Lesh, N. (2003) Engagement rules for human-robot collaborative interactions. *Proc. IEEE International Conference on Systems, Man & Cybernetics (CSMC)*, Vol. 4, pp. 3957-3962.
28. Vertegaal, R., Der Veer, G. and Vons, H. (2000) Effects of gaze on multiparty mediated communication. *Proc. Graphics Interface*. Morgan-Kaufmann Publishers, Montreal, Canada: Canadian Human-Computer Communications Society, pp. 95–102.
29. Vertegaal, R., Slagter, R., Der Veer, G. and Nijholt, A. (2001) Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. *ACM CHI Conference on Human Factors in Computing Systems*, pp. 301–308.
30. Wolfe J. (1999). "Inattentional amnesia", in Fleeting Memories. In *Cognition of Brief Visual Stimuli*. Cambridge, MA, MIT Press, pp. 71-94.