This paper is a preprint version of an article published by Springer-Verlag. The original publication is available at http://link.springer.com/chapter/10.1007%2F978-3-540-85845-4_36.

Toward Alignment With a Virtual Human - Achieving Joint Attention

Nadine Pfeiffer-Leßmann and Ipke Wachsmuth

Artificial Intelligence Group, Faculty of Technology
Bielefeld University
33594 Bielefeld, Germany
{nlessman,ipke}@techfak.uni-bielefeld.de

Abstract. Alignment forms the basis of successful communication. It can be seen as the most efficient means for action coordination of cooperating agents, covering adaptation processes operating without an explicit exchange of information states. One critical condition of alignment consists of *joint attention*. We present work on equipping a virtual human with the capability of reaching joint attention with its human interlocutor. On the one hand, mechanisms to detect the human's focus of attention are employed. On the other hand, basic cognitive as well as intentional processes underlying the phenomenon of joint attention are incorporated in our agent's cognitive architecture. In this context, a dynamic working memory and a partner model accounting for theory of mind and intentionality are crucial constituents.

Key words: joint attention, alignment, virtual humans, theory of mind

1 Introduction

In order to build a believable virtual human, we must understand how to model its cognitive abilities to engage in natural, successful face-to-face communication. According to Pickering and Garrod [1] successful communication is based on efficient action coordination and adaptation mechanisms realizing a close connection between the interlocutors. These *alignment processes* are joined processes between the interactants allowing them to sufficiently reconstruct the meaning of the interaction. One central condition of these joint process consists of *joint attention*. Joint attention facilitates interaction processes and supports inferences about people's current and future activities, both overt and covert. It is a foundational skill in human social interaction and cognition and can be defined as simultaneously allocating attention to a target as a consequence of attending to each other's attentional states [2]. However, to distinguish *joint attention* from *joint perception*, Tomasello stresses the intentional aspect of joint attention by demanding that the interlocutors have to deliberatively focus on the same entity while being mutually aware of this [3].

We investigate *joint attention* in an interaction scenario with the virtual human *Max* [4]. The human interlocutor meets the embodied conversational

agent face-to-face in a CAVE-like virtual environment. This scenario allows for the inclusion of verbal as well as non-verbal communication channels (e.g. gaze and gestures) for both the human interlocutor as well as for the virtual agent.

After discussing the psychological background of joint attention in the next section, current research on implementing attentional behavior in virtual humans and robots will be presented. The essential interplay of intentionality and attention will be covered in the requirements section. Based on these insights, our own approach of modeling joint attention in the virtual human *Max* will be presented in (Sect 5).

2 Psychological Background

Attention has been characterized as consisting of an increased awareness with respect to internal as well as external aspects such as perceptions, conceptions, and behaviors. This awareness can be invoked by involuntary as well as deliberate processes [5]. Attention is therefore not a unitary process but a complex phenomenon. Attention can be defined as intentionally directed perception [3]. Its purpose lies in the allowance and maintenance of goal-directed behavior. Cohen et al. [6] follow [7] in assuming three attentional subsystems: an *anterior attentional system* concerned with cognitive control and action selection, a *posterior attentional system* associated with orienting and perceptual attention, and an *arousal system* covering alertness phenomena.

During interactions, human attention is modulated by the observation of gaze direction and by inferences derived from observations. Recent experiments suggest that interacting agents pay as much attention to each other's intentions as they do to each other's observable acts [8]. Hobson additionally underlines that agents need the capacity of joining one another by sharing an experience and registering an intersubjective linkage. To reach joint attention, the agents need to be aware of of the other's focus of attention as well as of the process of sharing attention itself [9]. Considering these deliberative aspects of joint attention, the attentional focus of *cognitive control* has to be taken into account. Cognitive control and attention can be seen as emergent properties of information representation in working memory [10]. In line with this research, Cowan claims that the focus of attention is controlled conjointly by voluntary processes (*central executive system*) and involuntary processes (*the attention orienting systems*) [11]. Oberauer adopts these ideas seeing working memory as a concentric structure with its parts being characterized by an increased state of accessibility for cognitive processes: (1) The activated part of long-term memory holds information over brief periods of time. (2) The region of direct access serves to keep a limited number of chunks available for ongoing cognitive processes. (3) The *focus of attention* itself holds at any time the one chunk being selected for the next cognitive operation to be applied upon [12].

One of the most comprehensive mod1els of joint attention comes from Baron-Cohen's work on autism [13]. He postulates a tiered model containing four modules including an intentionality detector, an eye-direction detector, a shared at-

tention module, and a theory of mind module. Emphasis is put on the theory of mind as an endpoint and meta-representation as a process, but thereby some key relations, especially between attention and intentions are not described [14].

3 Related Work

In the area of virtual humans, researchers have mainly focused on modeling the perceptional attention focus as well as convincing gaze behavior [15] [16]. These computational models can be seen as prerequisites for joint attentional mechanisms. However, aspects of conjoining the attentional foci of the interlocutors are not covered.

A number of researchers in cognitive science and cognitive robotics use developmental insights as a basis for modeling joint attention. They show how a robot can acquire aspects of joint attention by supervised and unsupervised learning [2], [17], [18]. However, the aspect of intentionality and explicit representation of the other's mental state are not accounted for in these approaches. Work on a listener-robot explicitly addresses the issue of *joint attention* [19]. But in this robot, joint attention is modeled as an unconscious mechanism. The robot's behaviors are not subject to deliberate decisions but are implemented by if-then rules based on the redundancy of the interlocutor's attentional behavior and the communication mode. Breazeal et al. [20] work on a robot which is capable of rich social interactions and is provided with joint attention capabilities modeled as a collaborative process. The robot has an attention system which determines its attentional focus by calculating saliency values for all perceived objects. Additionally, the attention system is used to monitor and represent the human's focus of attention by attaching saliency tags to the respective objects. The robot's attention following and directing skills can be accompanied by conversational policies along with gestures and shifts of gaze accouting for repair, elaboration, and confirmation of the shared referential focus.

Hence, besides that the robotics community has recently demonstrated an increasing interest for modeling joint attention, most of the existing models focus only on partial and isolated elements of joint attention phenomena. They cover surface behaviors like simultaneous looking or simple coordinated behavior, but do not address the deeper, more cognitive aspects of joint attention [21].

4 Requirements

Following [21], we understand joint attention as an active bilateral process which involves attention alternation, and can only be fully understood by assuming that it is realized by intentional agents. To reach joint attention, the agents must be aware of the coordination mechanisms of understanding, monitoring and directing the intentions underlying the interlocutor's attentional behavior. To this end, the agent needs to be able to (r1) track the attentional behavior of other agents by gaze monitoring and has to (r2) derive the candidate objects the interlocutor may be focusing on from observation of the interlocutor's behavior and the

situational context. Furthermore, the agent has to (r3) recognize, whether the attentional direction cues of the interlocutor are put out intentionally or not. This aspect can be covered by keeping a model of the interlocutor's mental state with respect to his focus of attention. The agent has to (r4) react instantly, as simultaneity plays a crucial role in joint attention. When in response to an attentional direction cue, the agent deliberately draws its focus of attention on the referred object, it should (r5) use an adequate overt behavior which can be observed by its interlocutor.

To manipulate the attentional behavior of its interlocutor, the agent should also be able to engage in proto-declarative pointing, the ability to point in order to comment or remark on the world to another person. This behavior can be applied when gaze alone does not suffice. Also verbal references can be constituted to draw the interlocutor's attention on a specific object.

5 Modeling Joint Attention in a Virtual Human

Reconciling the requirements and research on joint attention, we propose the following model of joint attention (see Fig. 1). In order to reach joint attention, three main aspects need to be considered. Firstly, the agent's *mental state* serves as the origin of the attention mechanisms allowing for intentionally guided behavior. To reach joint attention, information covering a *partner model* which accounts for the interlocutor's *focus of attention* is acquired. This attentional focus has to be inferred from the interlocutor's overt behavior. Additionally, the environmental context is taken into account in order to embrace situatedness. This is achieved by bottom-up *activation processes* marking relevant objects as salient in the current situation. These processes can be seen as the second main aspect of joint attention. Thirdly, the agent itself needs to display appropriate overt behaviors to accentuate its *focus of attention* and to manipulate the *interlocutor's mental state*.

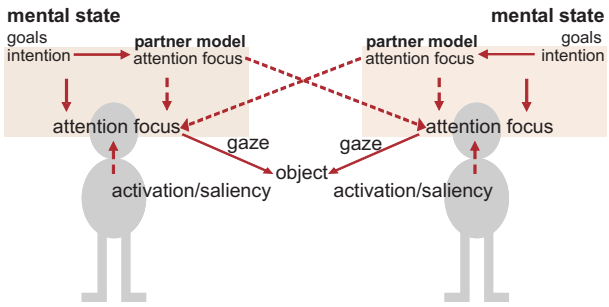

Fig. 1. Model of joint attention

5.1 Cognitive Architecture

Instead of modeling a separate module solely concerned with producing aligned gaze behavior of joint attention, we want to model the mechanisms behind the phenomenon of joint attention having their roots in the agent's cognition. We extend the *cognitive architecture* of our virtual agent [4]. In CASEC (Cognitive Architecture for a Situated Embodied Cooperator), ideas of basic cognitive mechanisms are integrated with explicit representations of mental states. The CASEC architecture adopts the BDI (*belief-desire-intention*) paradigm of rational agents [22] but additionally incorporates a dynamic memory model being inspired by work of Cowan and Oberauer (see Sect 2) to account for basic cognitive processes. Instead of static sets of beliefs in which entities have to be deliberately added and removed, the dynamic model employs automatic activation and decay processes. The activation values of the entities kept in working memory represent their saliency in the current context. They can be influenced by events, internal processing, and by the decision of the agent. In addition to *automatic processes* increasing the activation values whenever the object gets in the agent's gaze focus, *deliberative mechanisms* increase the activation values whenever the object is subject to internal processing. That is, in contrast to attention modeling in form of a stack on which objects are deliberatively pushed, we model the mechanisms behind attention so that an attention focus emerges out of the agent's goals, its behavior, and its interaction with the environment.

As attention is not seen as a unitary process, several attentional mechanisms are modeled in our architecture. Figure 2 outlines the agent's working memory. The rhombi represent relational chunks written into working memory functioning as explicit representations of the agent's *beliefs*. They descend from visual and auditive perception processes residing in the *phonological loop* and the *visual spatial memory*. Additionally, relevant chunks can be retrieved out of *long-term semantic memory* and are represented together with the agent's current conclusions. Activation impulses are sent by the agent's perception mechanisms. Whenever a new object is perceived or refocused, information about it covering its unique ID and its relevant attributes are written into working memory. The content of working memory is represented by use of relations which consist of a relation name e.g. *color, is_a, or inst_of* together with the respective attributes. Each of the relation entities has its own activation value. But not only the salience of the object's perceptual attributes makes them appear as attractors for the focus of attention. Additionally, when the agent perceives verbal input referring to an object, the respective activation values will be increased. Also when the agent notices that another agents focuses on a specific object, activation impulses are initiated. By this means, the content of working memory and thereby the candidate set of the agent's cognitive processes is influenced by its interlocutor providing the basis for *attention alignment processes*.

5.2 Attention Detection

As an indicator for the human interlocutor's focus of attention, we track the human's gaze as a basic manifestation of joint attention is gaze following. For

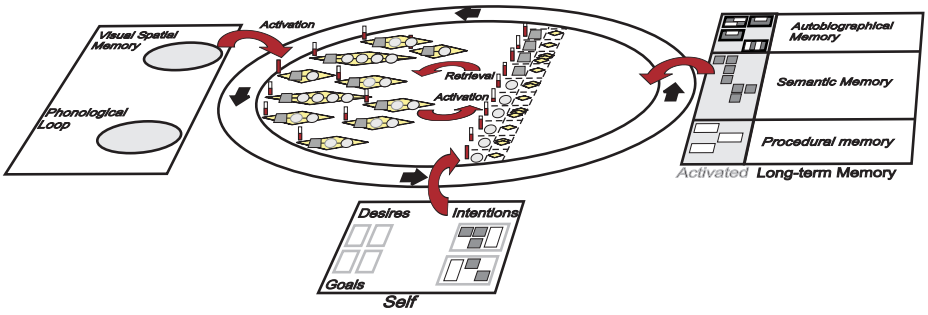

Fig. 2. Model of dynamic working memory

the first requirement (r1) (see Sect 4) tracking the interlocutor's focus of attention, an eye-tracker is employed. Additionally, the human's head direction is tracked by an infrared camera tracking system providing the position of the eyes relative to the world coordinate system. To (r2) derive the objects lying in the in the human's focus of attention, a cone of 2.5 degrees is used. In addition to the boundary of the cone, activation values of the agent's working memory are incorporated in the calculation of the candidate set. By this means, we do not only take the humans's line of view into account but also the situational context. The determination of the selected object is done using histogram calculations. An object is detected as being in the human's focus of attention, when it has been focused at least for a sum of 400ms in a 600ms time frame. In addition to this heuristic, the use of other communication channels such as pointing gestures or verbal expressions are interpreted as *intentional direction cues*. If attentional direction cues of a certain intensity are detected, the resulting belief is written into working memory.

5.3 Partner Model

The incorporated partner model covers beliefs about the conversational state and the conversational role of the interactant. Additionally, the interlocutor's attentional focus is represented as beliefs in form of *"attention_focus $interlocutor $object"* being updated dynamically and thereby leading to new beliefs or increasing a belief's activation respectively. However, as only intentional focusing is reckoned as an invitation to *joint attention*, we follow [19] with respect of (r3) detecting the speaker's intention: The intensity of the speaker's overt behaviors are interpreted as indicators for the interlocutor's intention. The intensity is calculated by the communication channel used and the redundancy of the interlocutor's attentional direction cues. For the agent to ascribe the desire to reach joint attention to its interlocutor, we use the following heuristic: An object has be the focus of attention for several times with additional short glances addressing the agent inbetween (triadic relation). Otherwise the interlocutor may just be focusing on the object with respect to other than interactional aspects. When

the activation value of an *attention-focus-belief* passes a threshold and the interlocutor has shown interactive glances, the agent believes *"achieve $interlocutor attention_focus self $object"* leading to an (iv) instantaneous activation of a *conclude*-plan. Thus the agent becomes aware of the interlocutor's intention and decides how to respond e.g. gazing at the same object to reach joint attention.

5.4 Attention Manipulation

While the attention detection mechanisms can be seen as prerequisites for engaging in joint attention, the agent also employs (r5) pro-active mechanisms to manipulate the interlocutor's focus of attention. Being an embodied conversational agent, the agent deploys different attentional direction cues e.g. intentional gaze, deictic gestures, and verbal expressions [4]. The agent's mental state is modeled using the BDI paradigm. Besides the explicitly represented beliefs, intentional states are represented as explicit goal states together with the means to achieve them. In case of the agent's goal to reach joint attention, the agent represents an explicit goal *"ACHIEVE attention_focus $interlocutor $object"* and pursues a plan in order to draw the interlocutor's attention toward the same object it is focusing on. When deciding which mechanism to use, the agent relies upon its plan library, usually adopting the plan with the least involved effort (e.g. gaze). If no suitable reaction of the interlocutor is perceived, more obvious attentional direction cues are applied. The plans are only carried out for as long as the goal state has not been achieved. As soon as the agent believes *"attention_focus $interlocutor $object"*, joint attention is achieved and the agent turns to the next goal of its agenda. Also, when the top-level goal which led to the instantiation of the *joint-attention-goal* is achieved or dropped, the *joint-attention-goal* is automatically abandoned. In case of not achieving joint attention, the failure part of the plan catches leading to a verbal expression making the agent's goal explicit.

6 Conclusion and Future Work

We presented work on equipping our virtual human Max with capabilities of joint attention. To this end, perception and detection mechanisms have been proposed and the processing in the agent's cognitive architecture has been presented. In future research, we want to evaluate how the model of joint attention is approved by naive human interactants. Additionally, we want to explore how parameters of joint attention mechanisms (e.g. timing, explicitness, intensity) can be tuned to adapt to the interactant's reactions applying for alignment processes.

Acknowledgments. This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center SFB 673. This paper is a preprint version of an article published by Springer-Verlag. The original publication is available at http://link.springer.com/chapter/10.1007%2F978-3-540-85845-4_36

References

1. Pickering, M.J., Garrod, S.: Alignment as the basis for successful communication. Research on Language and Computation **4**(2) (2006) 203–228
2. Deak, G.O., Fasel, I., Movellan, J.: The emergence of shared attention: Using robots to test developmental theories. In: Proc. of the First Intl. Workshop on Epigenetic Robotics, Lund University Cognitive Studies, 85. (2001) 95–104
3. Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences **28** (2005) 675–691
4. Leßmann, N., Kopp, S., Wachsmuth, I.: Situated interaction with a virtual human - perception, action, and cognition. In Rickheit, G., Wachsmuth, I., eds.: Situated Communication. Mouton de Gruyter, Berlin (2006) 287–323
5. Brinck, I.: The objects of attention. In: Proc. of ESPP2003, Torino. (2003) 1–4
6. Cohen, J., Aston-Jones, G., Gilzenrat, M.: A systems-level perspective on attention and cognitive control. In Posner, M., ed.: Cognitive Neuroscience of Attention. Guilford Publications (2004) 71–90
7. Posner, M., Petersen, S.: The attention system of the human brain. Annu. Rev. Neurosci. **13** (1990) 25–42
8. Nuku, P., Bekkering, H.: Joint attention: Inferring what others perceive (and don't perceive). Conscious and Cognition **17**(1) (2008) 339–349
9. Hobson, R.P.: What puts the jointness into joint attention? In Eilan, N., Hoerl, C., McCormack, T., Roessler, J., eds.: Joint attention: communication and other minds. Oxford University Press (2005) 185–204
10. Courtney, S.M.: Attention and cognitive control as emergent properties of information representation in working memory. Cognitive, Affective, & Behavioral Neuroscience **4**(4) (2004) 501–516
11. Cowan, N.: An embedded-processes model of working memory. In Miyake, A., Shah, P., eds.: Models of Working Memory, Mechanisms of Active Maintenance and Executive Control. Cambridge University Press (1999) 62–102
12. Oberauer, K.: Access to information in working memory: Exploring the focus of attention. J. of Exp. Psych.: Learning, Memory, and Cognition **28** (2002) 411–421
13. Baron-Cohen, S.: The eye-direction detector (edd) and the shared attention mechanism (sam): two cases for evolutionary psychology. In Moore, C., Dunham, P., eds.: Joint Attention: Its origins and role in development. L. Erlbaum (1994) 41–61
14. Tomasello, M.: Joint attention as social cognition. In Moore, C., Dunham, P., eds.: Joint Attention: Its origin and role in development. L. Erlbaum (1995) 103–128
15. Kim, Y., Hill, R.W., Traum, D.R.: Controlling the focus of perceptual attention in embodied conversational agents. In: Proceedings AAMAS. (2005) 1097–1098
16. Gu, E., Badler, N.I.: Visual attention and eye gaze during multiparty conversations with distractions. LNCS Intelligent Virtual Agents **4133** (2006) 193–204
17. Nagai, Y., Hosoda, K., Morita, A., Asada., M.: A constructive model for the development of joint attention. ConnectionScience **15**(4) (2003) 211–229
18. Doniec, M., Sun, G., Scassellati, B.: Active learning of joint attention. In: IEEE/RSJ International Conference on Humanoid Robotics. (2006)
19. Ogasawara, Y., Okamoto, M., Nakano, Y., Nishida, T.: Establishing natural communication environment between a human and a listener robot. In: AISB Symposium on Conversational Informatics. (2005) 42–51
20. Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., Chilongo, D.: Tutelage and collaboration for humanoid robots. International Journal of Humanoid Robots **1(2)** (2004) 315–348

21. Kaplan, F., Hafner, V.: The challenges of joint attention. Interaction Studies **7(2)** (2006) 135–169
22. Rao, A., Georgeff, M.: Modeling rational behavior within a BDI-architecture. Proc. Intl. Conf. on Principles of Knowledge Repr. and Planning (1991) 473–484