Simulating listener gaze and evaluating its effect on human speakers

Laura Frädrich[1] ✉, Fabrizio Nunnari[2], Maria Staudte[1], and Alexis Heloir[2,3]

[1] Embodied Spoken Interaction Group, Saarland University, Saarbrücken, DE
[2] SLSI Group, German Research Center for Artificial Intelligence, Saarbrücken, DE
[3] LAMIH, UMR CNRS 8201 / Université de Valenciennes, F

Abstract. This paper presents an agent architecture designed as part of a multidisciplinary collaboration between embodied agents development and psycho-linguistic experimentation. This collaboration will lead to an empirical study involving an interactive human-like avatar following participants' gaze. Instead of adapting existing "off the shelf" embodied agents solutions, experimenters and developers collaboratively designed and implemented experiment's logic and the avatar's real time behavior from scratch in the Blender environment following an agile methodology. Frequent iterations and short implementation sprints allowed the experimenters to focus on the experiment and test many interaction scenarios in a short time.

1 Introduction

Gaze is a very important aspect of social communication as it is closely connected with comprehension, planning, and prediction processes [6]. It furthermore represents a strong cue for the speaker and listener's focus of attention [2]. While the influence of the speaker's gaze and utterances on the listener's gaze has been investigated to some extent [7], the influence of the listener's gaze on the speaker's behaviour is largely unexplored so far. The difficulty of precisely controlling a human being's gaze behaviour might have contributed to the sparse amount of related work. Indeed, in an experimental setup which intends to investigate the speaker's reaction to the listener's behaviour, the manipulated variable, i.e. the listener's gaze behaviour, has to be controllable in a way that minimises the interfering impact of any confounding variables. In human-human interaction, however, the possible sources of interference are manifold. Only recent advances in the development of embodied conversational agents provide a possible solution to overcome these difficulties. Employing an artificial agent as listener makes the experiment substantially more controllable, interrupting the recursive relation between speaker and listener behaviour. Also, even though interacting with a virtual agent is an unusual situation for most, people generally treat agent gaze similarly to human gaze [6]. Our goal is to examine whether listener gaze can be simulated by simple gaze-following (also imitating joint attention) and whether that affects speaker behaviour in terms of speech production and gaze behaviour.

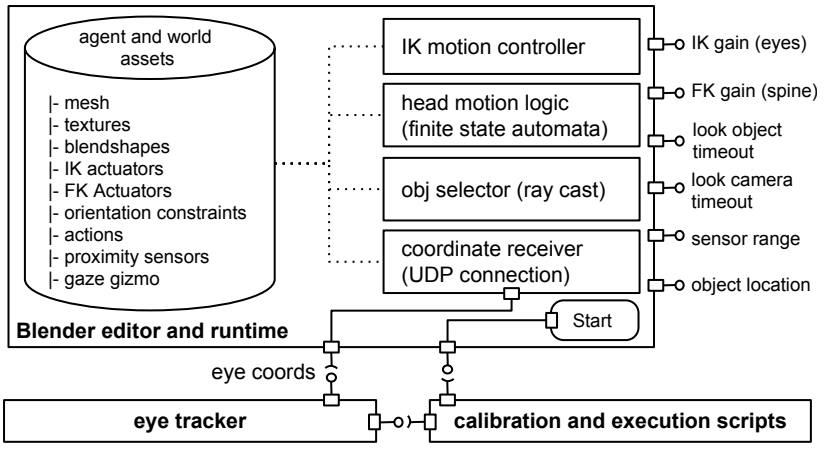

Fig. 1: Software architecture of the virtual agent.

2 Implementation Decisions and Architecture

The avatar-based interactive application supporting the experiment described in this paper capitalises on a previous work experiment [7]. The solution employed back then was comparable to available agent control frameworks [8,9,5,1] and offered a BML [4] interface to the experimenters to interactively control the avatar. This raised a number of issues, the most significant drawbacks being a cumbersome deployment process, a non-trivial compilation and packaging pipeline, and the impossibility for experimenters to interactively edit the scene layout or the assets. These issues hindered the collaboration flow between the multidisciplinary team as well as the experimental setup eventually used in the study.

In this experiment, the interactive agent framework was reimplemented from scratch around the Blender software[4]. Blender indeed provides an integrated full-fledged editing environment tightly coupled with a complete game engine which does not constrain which attributes of the game logic, assets, or real-time controllers can be exposed to the experimenters. Also, because all resources (assets, game logic, and input/output scripts) can be packed in a single .blend file, packaging and deployment becomes as trivial as sharing a single file on a convenient file-sharing service like Dropbox[5]. All assets used in the experiment are covered by the Creative Common (static objects) or the LGPL licence (Mesh, textures and weight maps of the Brad Character from ICT's Smartbody [8]).

Figure 1 depicts the architecture of the interactive setup. The eye tracker machine (bottom left) sends, via local area network, the coordinates of the screen point that the user is watching. The top box represents the content of the Blender

[4] `https://www.blender.org/` – 18 July 2017
[5] `http://www.dropbox.com/` – 18 July 2017

editor. It includes all the assets needed for the visualization as well as the invisible assets animating and controlling the agent.

A Blender scene makes use of four software modules: i) the receiver reads the eye coordinates via a UDP connection; ii) the object selector performs a ray cast of the eye coordinates in the 3D scenes to pick up a reference to the object which the user is currently watching; iii) the logic controller determines if the character has to play back an animation (e.g. nodding) or if it has to look at either a neutral location, the same object the user is looking at, or back at the camera; iv) finally, the motion controller applies the motion to the spine (via forward kinematics) and to the eyes (via inverse kinematics) of the character.

The right side of the Blender box shows the elements which are exposed to the experimenters so that they can autonomously fine-tune the experiment: i) the gain of both the eyes and the spine movements, which also influence the speed of the gazing behaviour; ii) the timeouts used by the state machine to change state, which are needed to tune the reaction speed of the avatar as well as to smooth the avatar's behavior when the subject's inferred gaze trajectory becomes noisy; iii) the range of the sensors used to intercept the ray cast; and iv) the location of the objects, to customise the layout of the scenes. The experimenters were also provided with two commodity scripts to start the calibration procedure and to run the trials.

3 Experimental set-up

We plan to use the avatar to address psycholinguistic research questions, in particular the influence of listener gaze on human-agent interaction. We intend to examine whether simple gaze-following by the avatar, which merely mirrors (after some filtering) the human participants' own gaze, might be enough to create the impression of an intelligent avatar that processes and understands spoken descriptions, and whether this affects the participant's behaviour in terms of speech production and gaze behaviour. In contrast to previous research concentrating, for instance, on the perception of virtual agents [3], we will exploit the actual, real-time gaze behaviour of the participants.

To this end, participants will sit in front of a computer screen displaying the content of the experiment. An eye tracking camera will be positioned directly below the screen. During the experiment, participants will view face and torso of the avatar (introduced as "Brad") with a set of six similar 3D objects arranged on a table in front of it, which they will be asked to describe verbally. The eye tracking technology will be used both as a diagnostic tool to measure the fixation locations as well as as input data for the experiment, endowing the virtual agent with the ability to follow the participant's gaze. Figure 2 depicts the data flow in this set-up.

References

1. Courgeon, M.: MultiModal Affective and Reactive Characters. Springer Lecture Notes in Artificial Intelligence (2011)

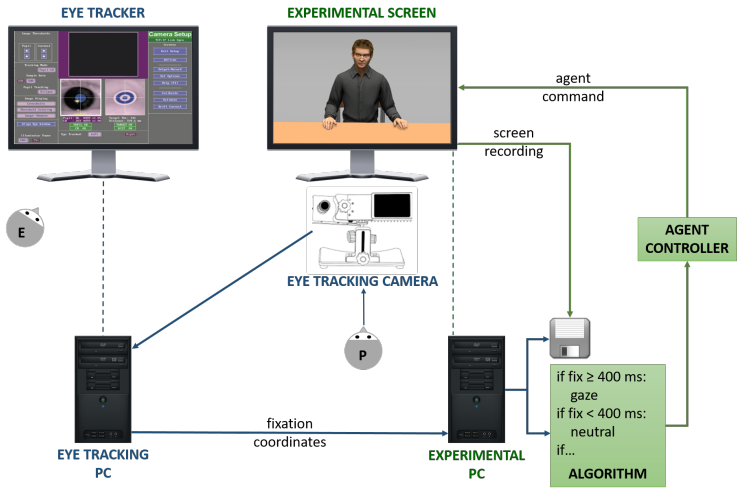

Fig. 2: Architecture of the experimental set-up. Data regarding the eye tracking are marked in blue; data relevant to the agent, in green. The camera records the participant's (P) fixations which are processed by the eye tracker PC. Calibration is done on the experimental PC screen.

2. Courgeon, M., Rautureau, G., Martin, J.C., Grynszpan, O.: Joint attention simulation using eye-tracking and virtual humans. IEEE Transactions on Affective Computing 5(3), 238–250 (2014)
3. Heylen, D., van Es, I., Nijholt, A., van Dijk, B.: Controlling the gaze of conversational agents. In: Advances in Natural Multimodal Dialogue Systems, pp. 245–262. Springer (2005)
4. Kopp, S., B, K., Marsella, S.S., Marshall, A., Pelachaud, C., Pirker, H., Thorisson, K.: Towards a common framework for multimodal generation in ECAs: The behavior markup language. In: Intelligent Virtual Agents 2006. pp. 205–217. Berlin: Springer-Verlag (2006)
5. Mancini, M., Pelachaud, C.: Dynamic behavior qualifiers for conversational agents. In: Intelligent Virtual Agents. pp. 112–124 (2007)
6. Staudte, M., Crocker, M.W.: Investigating joint attention mechanisms through spoken human–robot interaction. Cognition 120(2), 268–291 (2011)
7. Staudte, M., Crocker, M.W., Heloir, A., Kipp, M.: The influence of speaker gaze on listener comprehension: Contrasting visual versus intentional accounts. Cognition 133(1), 317–328 (Oct 2014), `http://linkinghub.elsevier.com/retrieve/pii/S0010027714001139`
8. Thiebaux, M., Marsella, S., Marshall, A.N., Kallmann, M.: Smartbody: Behavior realization for embodied conversational agents. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1. pp. 151–158 (2008)
9. Welbergen, H., Reidsma, D., Kopp, S.: An incremental multimodal realizer for behavior co-articulation and coordination. In: Intelligent Virtual Agents, vol. 7502, pp. 175–188. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)