An Architecture for Biologically Grounded Real-time Reflexive Behavior

Ulysses Bernardet[1]✉, Mathieu Chollet[2], Steve DiPaola[1], and Stefan Scherer[2]

1 School of Interactive Arts and Technology
Simon Fraser University Vancouver, Canada,
{ubernard, sdipaola}@sfu.ca
2 University of Southern California
Institute for Creative Technologies, Los Angeles, California
{mchollet, scherer}@ict.usc.edu

Abstract. In this paper, we present a reflexive behavior architecture, that is geared towards the application in the control of the non-verbal behavior of the virtual humans in a public speaking training system. The model is organized along the distinction between behavior triggers that are internal (endogenous) to the agent, and those that origin in the environment (exogenous). The endogenous subsystem controls gaze behavior, triggers self-adaptors, and shifts between different postures, while the exogenous system controls the reaction towards auditory stimuli with different temporal and valence characteristics. We evaluate the different components empirically by letting participants compare the output of the proposed system to valid alternative variations.

Keywords: Reactive behavior · reflexive behavior · cognitive architecture idle attention · virtual character

1 Introduction

In this paper, we present a pre-cognitive, reflexive architecture that is based on the distinction between reflexive behavior that has external and internal triggers. Examples of such reflexive behavior are orientation and startle responses to sounds in the environment or posture shift based on effort, respectively. We adopt a systems-based approach, and our aim is to provide a mechanistic model that is grounded in plausible psychological mechanisms. The presented architecture is geared towards the application in the control of the nonverbal behavior of the virtual humans that constitute the audience in a public speaking training system [1]. The development of the reflexive architecture is motivated by the argument that the efficacy of virtual humans as stand-ins for biological humans e.g. in training and therapy hinges on the social co-presence of the agent. Part and parcel of this is that the agent displays a realistic level of sensory-behavioral contingency, meaning that the character is showing behavior that is contingent on events that happen in the environment and within the agent [2]. This contingency is achieved by equipping the character with the ability to respond rapidly and plausibly to events that happen in the environment, be it the virtual space shared with the user in the case of full immersion, or events in the real world, in the case of a mixed-reality setup. In the context of public speaking training, the reflexive behavior system should increase the overall

adfa, p. 1, 2016.

plausibility of the behavior of the virtual humans, hence increasing the quality of the implicit feedback offered to the trainee. Furthermore, as the behavior control is parameterized, it can be easily integrated with implicit feedback by means of audience behavior, e.g. by controlling the level of restlessness of the spectators.

1.1 Related work

Behavior that is not a direct response to an event in the environment partially overlaps with what is called “idle” behavior in the domain of virtual characters. In the context of the model presented here, we refer to this behavior as endogenous reflexive behavior that we conceptualize as a response to internal triggers. The domains that are under the control of this system are self-adaptors, posture shifting, and gaze. The term "*self-adaptors*" refers to a class of self-touching behaviors that have no clear communicative function, and hence also occur when a person is alone. Self-adaptors are under weak intentional control and are thought to originally have served the purpose of satisfying some bodily need such as grooming. A number of studies empirically investigate self-adaptor usage e.g. in the context of counseling sessions [3], and the effect they have on the perception of a virtual character [4]. However, most architectures of nonverbal behavior, if they do include self-adaptors at all, do so by coupling them to verbal communication. The bulk of research on *posture* is related to gait and standing, and postural asymmetries related to pathologies such as stroke and Parkinson’s. Muscle fatigue plays a major role and is best investigate in standing postures. In the domain of conversational agents most, work on posture shifting is related to the structure and content of the discourse [5].

In our model, we refer to behavior that is triggered by events outside of the agent as “exogenous”. Attention as a mechanism that filters and prioritizes stimuli perceived by an organism plays a key role in this behavior and a number of attention models have been proposed for virtual characters. Most of these models, however, do not elaborate on the behavioral consequences of the attention process. In real as well as in virtual humans *gaze* serves a number of functions, including signaling of interest and emotional state, as well as regulation of conversations through the management of turn-taking [6]. Correspondingly, a number of works have investigated, mechanics of and models for gaze shifting [e.g. 23]. Gaze behavior independent of non-verbal or verbal exchange has been empirically investigated and modeled by [8], while [9] propose a model for attention towards specific objects in the environment of the agent.

Some virtual character architectures do include idle behavior [10], with the work of [11] on passive listening agents being probably the closest to the architecture presented here. Yet, in most virtual character architectures nonverbal behavior is coupled to symbolic expression.

2 Reflexive behavior architecture

The model presented here operates in an approximated continuous time domain with continuous internal variables (as opposed to a finite-state machine or a look-up table). Where possible we recur to known neurobiological processes such as habituation, refractory periods, leaky integrators etc. The rationale behind this approach is to gear the

model towards an eventual grounding in the neurobiological substrate. A pragmatic reason why we need to model underlying processes is that we want to develop a system that is capable of generating different behaviors with a minimal set of parameters. This is of particular interest when wanting to implement multiple discernible characters e.g. for a heterogeneous audience. By having a system with only a few, meaningful parameters, we can easily create a wide range of individualized characters without having to deal with an unmanageable number of parameters. Note that in the remainder of the description of the system we will indicate which are the parameters that can be tuned.

2.1 Architecture overview

At the most abstract level, the system can be divided into five components: At the "Input stage" the user is generating the inputs into the reflexive system and controls the playback of a spatialized sound in the virtual environment. The behavior control model itself comprises a slow and rapid exogenous and one endogenous reflexive behavior subsystem. These three subsystems independently send control signals to the virtual characters. In the current realization of the model, the endogenous subsystem controls gaze behavior, triggers self-adaptor actions, and shifts between different postures. The exogenous subsystem controls gaze behavior as well and additionally triggers different facial expressions. The endogenous subsystem is the default system that controls behavior in the absence of external events. As soon as an event in the environment occurs, that exogenous reflexive behavior takes control, shunting all endogenous behavior. This is achieved through a state of the slow exogenous subsystem that represent interest in the event, and a "post startle inhibition period" in the rapid response subsystem. Note that in the current version of the model, the exogenous reflexive behavior only includes auditory input, and the auditory stimulus is generated within the system itself and rendered with the virtual environment (as opposed to sensed from the real-world).

2.2 Endogenous reflexive behavior

Endogenous behaviors refer to actions that are driven by internal e.g. proprioceptive, signals. This class of behaviors comprises self-adaptors such as scratching, posture, and gaze behavior. Clearly, all these behaviors do have functions that go beyond mere reflexive action e.g. in communication, in the context of the architecture presented here, however, we, explicitly do not include these factors.

Self-adaptor behavior: Self-adaptors are behaviors of touching of one's hand, face or body to scratch, rub, groom, or caress it. Self-adaptor behavior can be motivated externally, or arise from internal motivations such as psychological discomfort, or as a displacement activity. Our functional view of self-adaptor behavior assumes that there are specific triggers and associated action that are being performed. At the core of the self-adaptor control stand two Poisson processes, that produce binary events with delays that follow a Poisson distribution (for diagram see https://figshare.com/articles/Endogenous_Self-adaptor_subsystem/3381547). One process generates events for adaptor targeting the head, the second one triggers self-adapting behavior on the extremities. We give the head self-adaptor a slight priority by implementing a "lateral inhibition" of the extremities self-adaptors, i.e. if both are triggered at exactly the same time, only the head action will be executed. Since the two pulse trains that trigger the actions are stochastic,

actions can potentially be triggered in rapid succession. To prevent this unrealistic behavior, we use a refractory period mechanism that suppresses triggers that are too close together. In the current implementation, the adaptor locations can trigger a set of two possible actions, i.e. neck rubbing/head scratching and hand rubbing/finger rubbing). Note that this choice was partially defined by the available animations, and does not present an inherent limitation of the system. In total the self-adaptor system has three tunable parameters: The two Poisson λ parameters that control the shape of the probability distribution of the occurrence for extremities and head self-adaptors, and the length refractory period.

Posture shifting based on fatigue: The second component of the endogenous system we will describe is the shifting between different postures. In the context of our model we are primarily interested in the somatic aspect of posture, and more specifically the motivation for switching from one posture to another. The key mechanism for posture switching is the accumulation of the effort that a posture requires maintaining. The effort of the current posture is integrated over time, and once the threshold is reached, a new posture is selected and the integrator is reset (for diagram see https://figshare.com/articles/Endogenous_Posture_control_subsystem/3381544). At this moment we manually define the effort each posture requires, but the model is explicitly constructed such that a realistic computation of actual strain on the joint and muscles can be added. The tunable parameters of the posture control system are the actual postures themselves, and their associated effort.

Gaze control: The gaze control system is loosely based on the system described in [12]. Similarly, we implement a process that is oscillating between mutual and non-mutual gaze. One key difference is that we draw the dwell times for mutual and non-mutual gaze from Poisson distributions. These distributions approximate the fitting function presented in [12], with that advantage of an easily tunable parameter in the form of the λ of the Poisson distribution. The gaze control process begins by drawing a random number from a Poisson distribution (for diagram see https://figshare.com/articles/Endogenous_Gaze_control_subsystem/3381550). This number then defines the duration for which the agent is looking at the speaker (this is implemented a linear decay function that triggers an event at zero-crossing). This event simultaneously starts the delay process for the non-mutual gaze and triggers gazing at a random location. Once the waiting time for the non-mutual gaze has expired, a new cycle of mutual gaze is initialized. The non-mutual gaze direction is drawing its horizontal and vertical saccade amplitude from a normal distribution with the location of the speaker as the mean. This allows having gaze which is more widely spread e.g. in the horizontal than the vertical plane. The parameters that can be tuned to control the gaze behavior are the λ parameters for the mutual and non-mutual Poisson distribution of the dwell time, and the variances for the horizontal and the vertical saccade amplitude. Additionally, the "Extent" parameter allows tuning which joints are involved in the gaze behavior (ranging from eyes only to eyes/neck/chest/back).

2.3 Exogenous reflexive behavior

We refer to behaviors that are a direct consequence of an event in the environment as exogenous (e.g., an acoustic distractor within the virtual or real space). Functionally this

reflexive behavior often subserves the acquisition of further information and, depending on the nature of the stimulus, the avoidance of harm. For the latter reason the exogenous reflexive system is generally more concerned with aversive than with appetitive stimuli, and in many cases, the behavior is accompanied by a brief, autonomic expression of affect. The exogenous reflexive behavior control is split into one circuit that deals with stimuli that are sudden, short, and strong, and one that controls the behavior towards sustained and slower onset stimuli. Both circuits are running in parallel, but due do their different sensitivity, most stimuli will only activate one or the other.

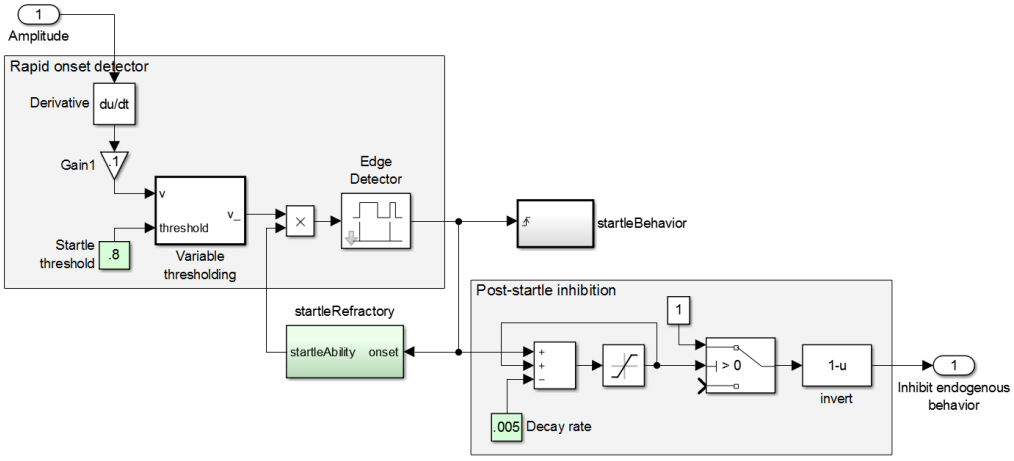


Fig. 1. Rapid response component of the exogenous reflexive system.

Rapid response system: The only input the rapid response system receives is the amplitude time course of the stimulus (distance scaling is taken care of in the "Input stage" block). The first signal processing stage is to detect that the stimulus is of rapid onset (Fig. 1 "Rapid onset detector"). This detection is achieved by first calculating the derivative of the stimulus, then applying a threshold, and finally detecting the binary edge. A computation of the derivative is required because the edge detector block by itself requires a signal that is raising from 0 to 1 in a single step, which is not realistic for even the most sudden signal that we would naturally encounter. The output of the edge detector will trigger a startle response comprising of a startle animation combined with the expression of surprise (Fig. 1 "startleBehavior"). The sensitivity of the system can be tuned using the "Startle threshold" parameter. We can assume that to startle is a fairly singular event, meaning that it should not occur repeatedly within a short amount of time. To implement this process, we use the mechanism of a refractory period (Fig. 1 "startleRefractory") that generates a shunting signal of a specific amount of time, effectively preventing startle behavior from occurring during that period. Internally the refractory mechanism is realized as a leaky integrator. Since startling is a somewhat disruptive event, we want to prevent the system from going back to normal operation for some time. This is realized with a "Post-startle inhibition", that produces a signal which

will inhibit the endogenous reflexive system for a specific amount of time. This inhibition process is as well implemented in the form a leaky integrator. Both the "startleRefractory" and the "Post-startle inhibition" have a tunable time constant parameters.

Slow response system: The slow response system has as inputs the position of the speaker (or camera) and the agent, as well as the amplitude time course and valence of the stimulus (Fig. 2). The positional inputs determine where that agent will attend to, while the amplitude time course and the valence, influence the dynamic response of the system.

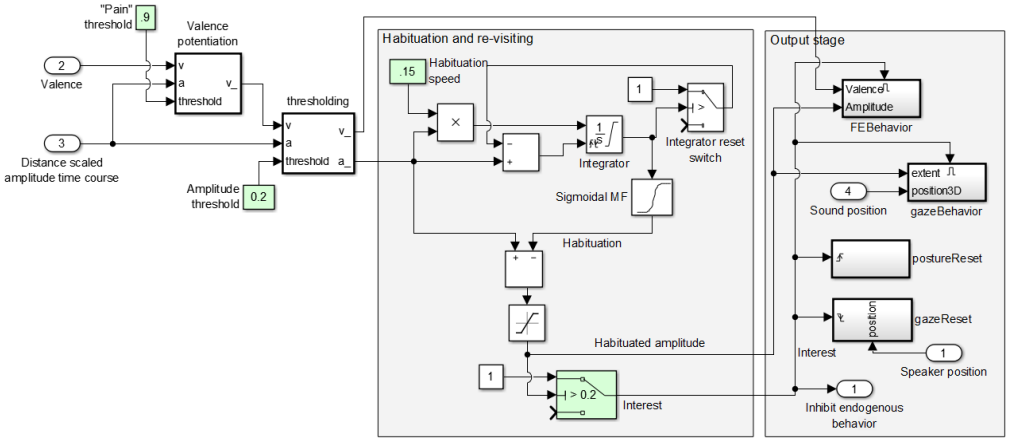


Fig. 2. Slow response component of the exogenous reflexive behavior subsystem.

The first processing step of the amplitude signal is a simple thresholding operation that ensures that low amplitude signals are discarded. At the core of the slow response system is the circuit for "Habituation and re-visiting". Within this block, the amplitude is integrated over time and smoothed with a sigmoid function (Fig. 2 "Sigmoid MF") to create a "Habituation" signal. This signal is then subtracted from the amplitude time course to yield a "Habituated amplitude". A drop of the input amplitude to zero immediately resets the integrator, and hence the habituation. A "Habituated amplitude" above a given threshold yields a binary "Interest" signal. At the output stage an "Interest" above zero enables both, the affective facial expression (Fig. 2 "FEbehavior") as well as the gazing in the direction for the sound (Fig. 2 "gazeBehavior"). The *valence* of the expression of affect is directly proportional to the valence' of the stimulus. While an onset of "interest" triggers the upright posture (Fig. 2 "postureReset"), a drop of "interest" to zero leads to gaze reset (Fig. 2 "gazeReset"), i.e. the agent gazes back at the speaker. In contrast to the binary "Interest" signal, the "Habituated amplitude" is a graded signal that varies in strength over time. It is this signal that controls the *amplitude* of the facial expression and the extent of joints – ranging from head only to head-neck-chest-back – involved in the gazing behavior. Hence the facial expression will be weaker, and the gazing will be less pronounced the weaker or further away a sound is. With the circuit described thus far the agent will gaze at the location of the sound source and display a facial expression as long as the agent has not habituated to the signal, or the single has

not dropped to zero. We assume that it is plausible that an agent will eventually re-visit a sustained input signal, not completely ignore it indefinitely. We implement this re-visiting behavior by resetting the integrator via the "Integrator reset switch" once it reaches a saturation threshold. The effect of this reset is that the system treats the input as novel, with the consequence of the agent exhibiting re-visiting behavior. The slow response subsystem has a total of three tunable parameters: Amplitude threshold, habituation speed, and Interest threshold.

2.4 Implementation

The high-level control of the behavior of the virtual humans is implemented using the graphical simulation environment Simulink[1] that allows implementing both, continuous as well as discrete control mechanisms [13]. We run a fixed step, soft real-time simulation using the block from [14]. The SmartBody virtual character system serves as the output platform [15], while the open source m+m software [16] provides middleware transportation layer between Simulink and SmartBody.

3 Empirical Evaluation

3.1 Stimulus material and procedure

To evaluate the reflexive behavior system, we conducted and empirical study; we ask participants to compare outputs generate with the model to variants. The subsystems were tested individually, i.e. subsystems that are not tested were disable during the experiment. All variants were generated by modifying the original model. We aimed to compare valid alternatives, i.e. variants that constitute plausible variations of the model, rather than generating arbitrary behavior.

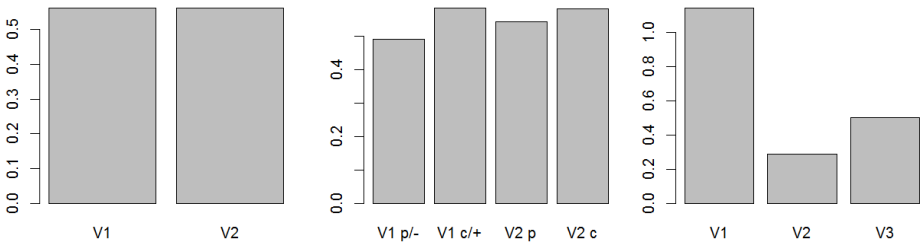
To test the appropriateness of the reflexive behavior to rapid vs. slow onset stimuli we used an auditory input that comprises a sequence of a slamming door and ringing phone. The behavior of the proposed system (V1 link) was a startle response after the door slam input and an orientation towards the location of the ringing phone. In the variation (V2 link), the agent we switched the sounds, and the agent was not startled by the phone and oriented towards the sound of the slammed door. We evaluated the influence of the affective response to external stimuli *(slow system affective response)* by comparing a strong negative response to a phone ringing (V1p/- link) with no display of affect (V2c link) to the same sound, and a positive affective response to an audio clip of a group chatting (V1c/+ link) to no affective response to the same sound (V2p link). All videos were 10s long. Lastly, we tested the s*low System habituation* by comparing the behavior towards the sound of chatting. The behavior of the proposed system (V1 link) was that the agent orients towards the sound then looks back at the camera (habituation), and as a third behavior orients again towards the sound source (re-visiting). In V2 (link) the agent shows no habituation, i.e. continues gazing in the direction of the sound, while in V3 (link) he shows habituation, but no re-visiting.

[1] www.mathworks.com/products/simulink

In the case of the exogenous behavior, the system was tested with internally generated signals (i.e. not recorded from the real-world). The test signals comprised of three components: 1) Amplitude time course, 2) Valence of the signal (constant over time), 3) Location in space. The amplitude time course is manually designed (as opposed to computed as an envelope) to mimic the key properties of the input signal such as speed of onset and duration. We assessed the realism of the videos using a scaled pairwise comparison. In this paradigm, participants are asked to indicate whether one video is much more, slightly more, or equally realistic.

3.2 Results

For the data analysis we used a tournament style scoring system: For each pairwise comparison between two videos we assign 1 or 2 points to the "winning video" (depending on whether participants chose "slightly more realistic" or "much more realistic", respectively). For the answer "Both videos are equivalent" both videos were given 0.5 points. The final score per video is the total score normalized by the number of matches played. A total of 343 pairs of videos were rated by 65 unique Amazon mechanical Turk workers, with age >18 years, and location U.S.A. Five pairs where the video did have a sound, but participant did not indicate which sound was played, were omitted.



(a) *(b)* *(c)*

Fig. 3. Evaluation of the exogenous reflexive system. a) Rapid and Slow system response. **V1**: Startle after door slam, orientation towards phone ring, **V2**: Startle response to phone sound, orientation towards door slam, b) Slow system-affective response. **V1p/-**: Sound: phone ring, strong negative response, **V1c/+**: Sound: chatting, positive response, **V2p**: Phone ring, no affective response, **V2c**: Chatting, no affective response. c) Slow System habituation (sound: chatting). **V2**: No habituation, **V3**: Normal habituation, no re-visiting.

Exogenous reflexive behavior: Somewhat surprisingly, swapping the door slam and the phone ringing sound did not yield a difference in perceived realism (Fig. 3a); startling when a door is slammed or when a phone rings, or, conversely, turning towards a phone ringing or towards a door that was slammed, are rated as equally realistic. All tested variations of the affective response in the slow system were rated very similarly (Fig. 3b). A positive response to chatting was virtually equivalent to no affective response to the same sound. Both, in turn, were rated as slightly more realistic than no affective response towards a phone ringing, or negative response respectively. Lastly,

the variations on the habituation response yielded the biggest differences in realism (Fig. 3c). Habituating to a stimulus but not re-visiting the location of the sound source was deemed more realistic than a sustained gazing in the direction of the sound. By far the most realistic behavior was generated by the proposed system, i.e. by habituation to the sound and subsequent re-visiting.

4 Discussion

In the exogenous subsystem, one of the most unexpected results was that swapping door slam and phone sound did not make a difference. In hindsight, it does indeed make sense that a person would, in addition to a startle response, also show an orientation response towards the rapid onset door slam sound. Conversely, startling, when a phone rings, is similarly something most people will have experienced personally. The mixed results regarding the affective response hint at the problem that the signal to noise ratio between affective response and distractors such as posture, eye movement etc. was not big enough. The open-ended question that asked participants about the reason for their assessment showed that some participants found the affective response was exaggerated, while others did not seem to have noticed it at all. Adding more channels for affective expressions besides facial animations would allow tuning down the amplitude of the latter (hence avoiding unrealistic exaggeration), while simultaneously making the affective response more detectable. The results regarding the habituation response seem to indicate that realism is best achieved for systems that neither involve an affective response nor require that participants pick up on the stochastic nature of a temporal distribution.

5 Conclusion

In this paper, we have presented ongoing work on the development of a reflexive behavior architecture. We follow a systems approach where we build models based on dynamic mechanisms underlying the actual behavior. One of the advantages of this approach is that a small set of canonical parameters can generate a wide range of different behaviors. In the initial phase of the development, the architecture parameters were set based mostly on the modeler's common sense. A next step will be to ground the parameter values empirically. Planned further developments of the architecture include the addition of more behaviors such as evasion, and, most importantly, the inclusion of the visual modality.

Acknowledgments

This work was partially supported by" Moving Stories" Canadian SSHRC grant.

References

1. Chollet, M., Wörtwein, T., Morency, L., Shapiro, A., Scherer, S.: Exploring feedback

strategies to improve public speaking. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp ’15. pp. 1143–1154. ACM Press, New York, New York, USA (2015).
2. Inderbitzin, M., Betella, A., Lanatá, A., Scilingo, E.P., Bernardet, U., Verschure, P.F.M.J.: The Social Perceptual Salience Effect. J. Exp. Psychol. Hum. Percept. Perform. 39, 62–74 (2013).
3. Schulman, D., Bickmore, T.: Changes in Verbal and Nonverbal Conversational Behavior in Long-Term Interaction. In: Proceedings of the 14th ACM international conference on Multimodal interaction (2012).
4. Krämer, N.C., Simons, N., Kopp, S.: The Effects of an Embodied Conversational Agent ´ s Nonverbal Behavior on User ´ s Evaluation and Behavioral Mimicry. Intell. Virtual Agents. 238–251 (2007).
5. Schulman, D., Bickmore, T.: Posture , Relationship , and Discourse Structure Models of Nonverbal Behavior for Long-Term Interaction. 106–112 (2011).
6. Ruhland, K., Andrist, S., Badler, J., Peters, C., Badler, N., Gleicher, M., Mutlu, B., Mcdonnell, R.: Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. Eurographics State-of-the-Art Rep. 69–91 (2014).
7. Pejsa, T., Andrist, S., Gleicher, M., Mutlu, B.: Gaze and Attention Management for Embodied Conversational Agents. ACM Trans. Interact. Intell. Syst. 5, 3:1–3:34 (2015).
8. Cafaro, A., Gaito, R., Vilhjálmsson, H.: Animating idle gaze in public places. Intell. Virtual Agents. 250–256 (2009).
9. Kokkinara, E., Oyekoya, O.: Modelling selective visual attention for autonomous virtual characters. Animat. Virtual. 361–369 (2011).
10. Kallmann, M., Monzani, J.-S., Caicedo, A., Thalmann, D.: ACE: A Platform for the Real Time Simulation of Virtual Human Agents. In: Computer Animation and …. pp. 73–84 (2000).
11. Maatman, R.M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 3661 LNAI, 25–36 (2005).
12. Lee, S.P., Badler, J.B., Badler, N.I.: Eyes alive. ACM Trans. Graph. 21, 637–644 (2002).
13. Saberi, M., Bernardet, U., DiPaola, S.: Model of Personality-Based, Nonverbal Behavior in Affective Virtual Humanoid Character. In: ICMI ’15: 2015 International Conference on Multimodal Interaction (2015).
14. Ivo Houtzager: Simulink Block for Real Time Execution, http://www.mathworks.com/matlabcentral/fileexchange/30953-simulink-block-for-real-time-execution.
15. Shapiro, A.: Building a character animation system. In: Motion in Games. pp. 98–109 (2011).
16. Bernardet, U., Schiphorst, T., Adhia, D., Jaffe, N., Wang, J., Nixon, M., Alemi, O., Phillips, J., DiPaola, S., Pasquier, P.: m+m: A Novel Middleware for Distributed, Movement Based Interactive Multimedia Systems. In: Proceedings of the 3rd International Symposium on Movement and Computing - MOCO ’16. pp. 21:1–21:9. ACM Press, New York, New York, USA (2016).