A Perceptual Sensing System for Interactive Virtual Agents: towards Human-like Expressiveness and Reactiveness

Alberto Jovane, Pierre Raimbaud

To cite this version:

Alberto Jovane, Pierre Raimbaud. A Perceptual Sensing System for Interactive Virtual Agents: towards Human-like Expressiveness and Reactiveness. Julien Pettré; Barbara Solenthaler; Rachel McDonnell; Christopher Peters. 16th Annual ACM SIGGRAPH Conference on Motion, Interaction and Games (ACM MIG 2023), Nov 2023, Rennes, France. ACM, Proceedings of Conference on Motion, Interaction and Games (MIG ’23), 2023. hal-04418557

HAL Id: hal-04418557
https://hal.science/hal-04418557

Submitted on 26 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License
A Perceptual Sensing System for Interactive Virtual Agents: towards Human-like Expressiveness and Reactiveness

Alberto Jovane
jovanea@tcd.ie
Trinity College Dublin
Dublin, Ireland

Pierre Raimbaud
pierre.raimbaud@ec-lyon.fr
Univ Lyon, Centrale Lyon, CNRS, INSA Lyon, UCBL,
LIRIS, UMR5205, ENISE
Saint-Etienne, France

Figure 1: Representation of communication losses case study. In (a) the two agents, the red being the one controlled by our system, talk on a train platform. Then, we show the effect that the environment noises have on the communication: in (b) the noise of the moving train impacts the conversational sound perception; in (c) the presence of train passengers produce visual and sound interferences on the two agents’ communication. (d) shows the result of our detection and adaption model via a visual and sound sensing capabilities system given to the controlled virtual agent: higher speech level and gesture amplitude.

ABSTRACT
Social communication interactions are paramount in human daily-life. In virtual environments, virtual agents allow for more social presence to the users, and can trigger social behaviours on them. However, the complexity of real-life situations, notably with external stimuli perturbations (noise, visual occlusions etc.) is not handled by classical interaction models. We propose here a model that includes an acoustic and visual sensing system, to compute and evaluate agent-dependent perceptions from the agents and their surroundings, before triggering expressive actions and reactions.

CCS CONCEPTS
• Computing methodologies → Simulation environments; Motion processing; Procedural animation.

KEYWORDS
virtual agents, animation, perception, interaction, expressiveness

1 INTRODUCTION
Since several decades, computer science technologies have allowed for the creation of Virtual Environments (VE), which can be used either passively such as in 3D movies or actively such as in videogames or in Virtual Reality (VR) [15]. In both cases, VE purposes can be very diverse, such as education, entertainment, professional training [26]. Some VEs try to replicate real world spaces and conditions, whereas others explore fantastic possibilities; either way, in most cases humans expect VEs to be populated by social beings [17, 23]. Populated VR environments usually produce a feeling of social presence to its users [11], which contributes to improve their global experience [27]. In multi-users environments, users’ representations through avatars permit to achieve such “populating objective” [19]. Nonetheless, despite recent improvements of Internet and extended reality (XR) technologies [5, 28], most XR applications still focus on single-user context. The design of populated VEs remains thus highly dependent on the development of virtual human agents [21].

The design of virtual agents has significantly improved in the last years [10], which can notably be explained by the continuous enhancement of 3D models production and rendering capabilities [20]. This is easily noticeable in movies and videogames, and through the emergence of multiple artificial intelligence techniques, observable on autonomous agents, either in commercial products – e.g., voice assistants [16, 18], or state-of-the-art 3D virtual humans [3]. Virtual agent improvements are usually tagged as realism or believability enhancements: this encompasses not only the agent external appearance, but many different aspects, which must synergise to avoid context and characteristic-dependent issues like the uncanny valley effect [12]. Therefore, together with physical appearance [29], the literature showed the importance of facial expressions [9], body motions [25] or non-verbal communication cues in general [1]. Similarly, research has been conducted about the virtual agents’ abilities to interact with the environment and with VR users [13, 22] through actions like posture/facial expressions changes, verbal communication, navigation, objects manipulation, physical and social behaviours, improving the virtual agents’ believability [7].
Here, we focus on virtual agents designed with the aim of socially interacting between them or with VR users. In such a context, agents usually benefit from various interactions capabilities such as conversation [6] with lip synchronisation and facial expression [4, 24], gesture synthesis [2, 8], or emotion mimicry [14]. All these capability examples typically assume a two-way interaction where interactants face each other in a social context free of any disturbance for their exchange. To ensure this, videogames commonly focus the user attention on the interacting agent by, for example, blurring out all the surroundings, but the virtual representation of realistic situations with "perturbed-context interactions" cannot be restricted to this. The literature currently lacks from a model that would allow for the design of such situations with interactive virtual agents able to behave accordingly. In such a model, agents should be able to convey human-like expressions and reactions in accordance with their awareness of the surroundings, making it more convincing for VR users. Our paper proposes a model to enable virtual agents to behave like this by providing them with a sensing system to visually and audibly perceive their environment and thus to adapt their user-interaction behaviours in real-time.

2 A MODEL BASED ON PERCEPTUAL SENSING SYSTEM FOR INTERACTIVE AGENTS

We propose here a model for virtual agents to provide them with a human-like sensing system – visual and acoustic (so far), for more realistic expressions and reactions when interacting with other agents or VR users. In dyadic contexts, composed of one virtual agent employing our model and another interactant (agent or VR user), the model functionalities rely on a new two-way sensing system for the agent, i.e. not only with a sensory user-feedback way, but also with a reverse-way of interpretation of signals on the controlled virtual agent involved in the interaction with the user.

In details, the system relies on the evaluation of emitted/perceived signals (sound and visual) with all the other potential interactants. The dimensionality of these signals is simplified by the computation of salient features, which are related to non-verbal communication elements. Examples of visual features are the apparent occupation of the other’s field of view, the apparent amplitude, velocity and direction of gestures and posture. Similarly, example of acoustic features are: sound intensity, interaural time difference, which allows for the computation of source perceived distance and orientation.

Relying on our sensing system, we propose a three-step interaction model (see Figure 2): 1) constant evaluation of two-ways visual and acoustic signals (from the controlled agent to one communication interactant), 2) synthesis, evaluation of the perceived signals to interpret and decide the next behaviour, and 3) triggering of the appropriate (re)actions on the agent. This requires a preliminary phase of features identifications and thresholds tuning for each reaction. These elements are case-specific, and can be extracted from data or user perception evaluations through the recreation of sample scenarios notably with simulated external interferences.

3 BENEFITS AND RELEVANT CASE STUDIES

Our system provides insights about the visual and sound cues from the own perspective of an agent but also from the other interactants. This enables more realistic behaviours such as more adapted reactions, or actions that would be more easily perceived. With our system, the communication state between the agents is available to them, enabling them to behave according to this knowledge.

Then, social situations with communication losses are particularly interesting for the use of our model. Indeed, classical interaction systems carefully generate voice and gestuality of the agent, but without considering if the intended messages are properly received, which is problematic in complex VE situations e.g., when VR users and agents can freely interact, and thus produce not-prepared and not-precontextualised events. In such scenarios, the communication flow might be disturbed by external noises and visual occluders, jeopardizing the communication meaning. Agents using our system would be able to communicate accordingly to the situation since existing disturbances could be evaluated by comparing agents-perceived signals features with emitted ones, and faced by triggering appropriate actions/reactions (e.g., clarification, repetition, indications about the original communication failure).

An example of communication losses situation is: two agents are talking in a platform of a virtual train station; a train comes and starts stopping close to the red agent (see Figure 1), making a lot of noise; with our system, the black agent can perceive red agent’s difficulties to ear and thus increase its talking-volume; moreover, a crowd of agents gets out of the train, making noise and occluding agents’ mutual gaze line of sight, but our perceptually-driven agents can adapt their face position to improve their communication.

4 LIMITATIONS AND FUTURE WORK

Our perceptual sensing system for interactive virtual agents might nonetheless suffers from some outcomes. First, it would require optimisation to avoid performance drop in case of crowded virtual environments with multiple interactions (social groups, in and out-group interactions, multiple VR users etc.). In any cases, it would be recommended to optimise the time dedicated to fully new computations of perceived acoustic and visual features, depending on the real-time likeliness for an agent to be interacting with others. Another limitation is the implementation of just two senses, whereas real humans have a more extended perception (haptics, proximity etc.). It could also be interesting to couple our system with a cognitive model, to improve our synthesis-evaluation-interpretation step, since real humans perform cognitive operations after perceiving (decision-making) and it would formalise rules/cases of.

Figure 2: Top: the general communication case between the red agent (controlled by our system) and other users/agents, with exchanged signals in the VE. Bottom: our 3-step model.
action/reaction decisions. Finally, our system must be tuned and stress-tested on multiple case studies, particularly the ones that explained in Section 3. Our future work will focus on completing our implementation and testing it on case studies, as well as evaluations with users, either through videos or involving them in VR.

ACKNOWLEDGMENTS
To Robert, for the bagels and explaining CMYK and color spaces and, to Eliott Zimmermann, for all the conversations regarding human audio perception.

REFERENCES