Toward a human-like sound perception for reactive virtual agents

Audrey Pichard, Gauthier Couzon, Eliott Zimmermann, Pierre Raimbaud

Toward a Human-like Sound Perception for Reactive Virtual Agents

Gauthier Couzon
gauthier.couzon@enise.fr
ENISE, Centrale Lyon
Saint-Etienne, France

Audrey Pichard
audrey.pichard@enise.fr
ENISE, Centrale Lyon
Saint-Etienne, France

Eliott Zimmermann
eliott.zimmermann@enise.ec-lyon.fr
Univ Lyon, Centrale Lyon, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, ENISE
Saint-Etienne, France

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

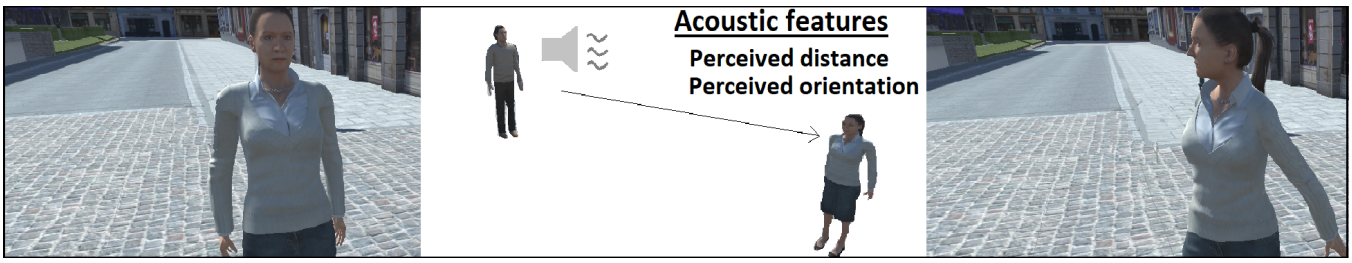

Figure 1: An Intelligent Virtual Agent having a reactive behaviour in the direction of a sound coming from the right

ABSTRACT

Human social interactions rely on multisensory cues. In this regard, visual and auditory cues are paramount during the initiation of an interaction. In this preliminary work, we propose an approach to let Intelligent Virtual Agents (IVAs) simulating sound perception capabilities. Our model targets to control IVAs' reactive behaviour through their analysis of perceived other agents' emitted sounds. For that, we explored auditory features close to the human system.

CCS CONCEPTS

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

KEYWORDS

virtual agents, reactions, animation, sound perception, social interactions

ACM Reference Format:
Gauthier Couzon, Audrey Pichard, Eliott Zimmermann, and Pierre Raimbaud. 2023. Toward a Human-like Sound Perception for Reactive Virtual Agents. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Corresponding author

Conference'17, July 2017, Washington, DC, USA
2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As humans are social beings [16], social behaviours are paramount to any daily interaction. Moreover, humans expect other beings to have social behaviours with them, as shown in the literature for animals [1, 32], robots [17, 26], and virtual agents [5, 10].

In the field of virtual reality (VR), virtual agents are key to convey emotions and to allow for social interactions. They contribute to the social presence felt by users [29], along with their immersion [27] and engagement in VR [28]. However, to elicit such effects and to observe complex social behavioural responses [20], the design of virtual agents must overcome several issues. Yet, solutions are highly context-depend, for example for the visual appearance of agents, the appropriate quality level varies according to the quality of the environment and its purposes [31], e.g., when seeking naturalness in interactions, a photorealistic appearance elicits more realistic social responses [40]. Similarly, realistic motions, as well as speech quality, are key for VR users' perception of virtual agents [35].

Challenges remain to improve agents' believability [13], especially for social interactions, toward reactive behaviour modelling, contributing to making them Intelligent Virtual Agents (IVA). For this, Raimbaud et al. [30] proposed a visual perception approach, based on the analysis of how the motions of an agent are perceived from an other agent viewpoint. From the context and motion-depend features computed on the observed motions, reactions are triggered on the observer. However, no other kind of perception than the *visual* one was used in this study, despite humans daily use other types of cues to perceive an action before reacting to it.

Humans rely on crossmodal perceptions [15] that they process and balance to act and react, in real life and in VR [37], with real

humans and other agents (robots [36], virtual agents [24]). In many contexts, *sound* perception is paramount – usually combined with other sensory perceptions [14, 34], particularly to draw and maintain attention. In driving context, Wang et al. [38] found that drivers' performance improved through new usages of sound for advisory information (replacing visual cues), whereas Bellotti et al. [3] showed that its spatialisation improved driving, especially in case of low visibility. Similarly, it has been shown that spatialised sounds helped listeners to understand recorded music concerts, either regarding the music itself [2] or its social context [33]. Thus, in these contexts and based on human abilities to localise sources [8], spatialised sounds have succeeded in triggering oriented actions or reactions. This has also been observed on VR users [19]. Moreover, Huang et al. [22, 39] implemented a sound localisation approach to drive virtual agents' behaviours (e.g., chasing agents), relying on a sound field propagation model, used then to determine the source origin. In line with this, Chemistruck et al. [6] built another energy-wave propagation model, improved with masking and reverberation components. Finally, Cowan et al. [9] proposed to use a spatial graph and to estimate the sound propagation by computing its shortest path between the source and the receiver, from occlusions. However, these approaches did not intend to mimick the human receiving auditory system. Therefore, we propose in this paper a new model for sound reactive behaviours of virtual agents, from an egocentric viewpoint and inspired by human perception. Section 2 presents it, Section 3 a case study, and Section 4 its limitations and perspectives.

2 EGOCENTRIC SOUND PERCEPTION MODEL

We present here a new approach that contributes to the modelling of reactive behaviours – either between IVAs or with a VR user. It relies on a human-like egocentric perception of sounds, proposing thus a different perspective compared to previous work on agents in virtual environments [6, 9, 22, 39]. Fig. 2 displays our approach, which encompasses two steps: the computation of acoustic features on sounds as perceived from the receiver viewpoint, and a result synthesis step to induce reactive behaviours on the receiving agent.

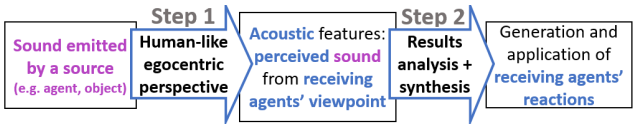

Figure 2: Sound perception model for IVA reactive behaviours

2.1 Acoustic features computation

Our first step consists in computing acoustic features from the receiver viewpoint, regardless of source sound characteristics and in a human-like perception way. We present three main features i) the perceived sound intensity, ii) the perceived distance, and iii) the perceived orientation; yet, our model is not limited to these ones.

In physical terms, the sound power P describes the behaviour of a sound source, independently from the distance, position and environment, whereas the sound intensity I refers to its power in a given area, thus measured at a specific position and representing the sound information perceived there. Therefore, capturing through a listener the sound intensity at the receiver position is a paramount feature to understand the local perception of sound. In addition, the *perceived intensity* feature I can be used to compute a *perceived distance* r to the source, through the following formula [23], valid for source emission in spherical volumes and knowing its power P:

$$I = \frac{P}{4\pi r^2} \quad \Longleftrightarrow \quad r = \sqrt{\frac{P}{4\pi I}}$$

It should be noted that, in a virtual environment and thus for numerical sounds, the computation of power and intensity depends on the induced volume of the different software layers (operating system, applications). A common approach is thus to use a reference power level value of the sound on a given volume, and to measure the intensity level at the virtual agent's position with the same volume, allowing then for the computation of the distance r.

The Interaural Time Difference (ITD) [25] refers to the time between the perception of a sound in the left ear and in the right ear, which is perceptible and used by humans for source localisation. The ITD can be expressed from the interaural distance (r_{int}), the sound celerity c in the environment, and the θ azimuth between the sound emitter direction and the axis passing through the ears, and therefore the *perceived orientation feature* can be computed from:

$$\Delta t = \frac{r_{int} \cdot (\theta + \sin(\theta))}{c}$$

2.2 Results synthesis and reactions generation

Mimicking the perception-action loop used by humans [21], we propose to generate and trigger reactions on the receiving agent [12] from the perceived acoustic features. This second step consists in synthesising the acoustic features results to drive realistic reaction behaviours on the IVA. We present here some possible analyses and syntheses from the acoustic features previously presented.

An analysis can be done on the perceived intensity by comparing it to a threshold value. For example, this one can be determined as the intensity of the ambient sound of the environment (background noise), or as a fixed minimum value for sounds to be considered as "triggering sources". In both cases, a reaction would be triggered on the IVA when the perceived intensity is higher to the threshold value. Thresholds on the perceived distance can be used similarly.

Another analysis can be performed on the perceived distance, along with the perceived orientation. The combined result can be used to determine a perceived position for the sound source from the receiver agent's viewpoint, and therefore to generate a reactive orientation motion in the direction of this computed position.

3 CASE STUDY: A CALL IN THE STREET

We present an illustrative case study to exhibit our approach: an IVA calls another IVA in the street – at a distance of $5m$ and at $90°$ on its right, to draw its attention by shouting. In this case, the perceived distance and perceived orientation acoustic features are computed in the first step of our model (results: $4.89m$ and $90°$). From this, a perceived position is deduced, and a combination between torso, head and eye orientations is generated for a reactive behaviour of the receiving IVA toward the estimated sound source, realistically simulating a response to the calling of the other IVA. Fig. 1 shows this case study in a virtual environment where we implemented our model, here in Unity with sound spatialisation by SteamAudio.

4 LIMITATIONS AND FUTURE WORK

A first limitation, in the current development state of our model, is that we use sound power to compute the perceived distance feature. In real-life, humans use their ability to recognise the type of a sound, which they associate to a known power when they are familiar to it, to estimate then its distance [4, 11]. It would be interesting to use a similar approach in our model, e.g., through deep-learning approaches to recognise the type of the perceived sounds. Then, with our current implementation, results accuracy for the perceived azimuth orientation can vary depending on the IVAs' relative positions, e.g., with ±0-1° accuracy for a 90° angle, and ±5-10° for 135°. Even though human accuracy to localise sound can be up to 5 or 10° compared the real position in most conditions [18], our model could be improved by using more "human-like physiology" measures than only the ITD, e.g., interaural phase and level differences etc. This would also contribute to expand our human-like model with more features such as the perceived elevation angle.

As future work, our approach could be integrated in a more complex framework for IVAs' reactive behaviours, where multimodal egocentric perceptions would be used as humans do (e.g. combining our sound-based model with vision-based approaches [30]), as well as other types of approaches (e.g. deep-learning approaches where agents are trained to navigate to a sound source, as recently developed for robots [7]). Finally, we also aim to extend the use of our model to other case studies and contexts, notably with multiple "receiver agents", and also in VR with users that would interact with the virtual agents, and trigger reactive behaviours on them.

REFERENCES

[1] Catherine E Amiot and Brock Bastian. 2015. Toward a psychology of human–animal relations. *Psychological bulletin* 141, 1 (2015), 6. https://doi.org/10.1037/a0038147

[2] Natasha Barrett and Marta Crispino. 2018. The impact of 3-D sound spatialisation on listeners' understanding of human agency in acousmatic music. *Journal of New Music Research* 47, 5 (2018), 399–415. https://doi.org/10.1080/09298215.2018.1437187

[3] Francesco Bellotti, Riccardo Berta, Alessandro De Gloria, and Massimiliano Margarone. 2002. Using 3d sound to improve the effectiveness of the advanced driver assistance systems. *Personal and ubiquitous computing* 6 (2002), 155–163. https://doi.org/10.1007/s007790200016

[4] Erik Berglund and Joaquin Sitte. 2005. Sound source localisation through active audition. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, IEEExplore, 3 Park Avenue, 17th Floor, NY, USA, 653–658. https://doi.org/10.1109/IROS.2005.1545032

[5] Donato M. Cereghetti, Styliani Kleanthous, Christophoros Christophorou, Christiana Tsiourti, Cindy Wings, and Eleni Christodoulou. 2015. Virtual partners for seniors: Analysis of the users' preferences and expectations on personality and appearance. In *AmI (Workshops/Posters)*. Vol. 1528. CEUR-WS, Aachen, Germany.

[6] Mike Chemistruck, Andrew Allen, John Snyder, and Nikunj Raghuvanshi. 2021. Efficient acoustic perception for virtual AI agents. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 4, 3 (2021), 1–13. https://doi.org/10.1145/3480139

[7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, Springer International Publishing, Cham, 17–36. https://doi.org/10.1007/978-3-030-58539-6_2

[8] Fang Chen. 2002. The reaction time for subjects to localize 3D sound via headphones. *Journal of the audio engineering society* Journal of the audio engineering society, 1 (2002), 1–4.

[9] Brent Cowan, Bill Kapralos, and KC Collins. 2020. Realistic Auditory Artificial Intelligence: Spatial Sound Modelling to Provide NPCs with Sound Perception.

[10] Stephen Cranefield and Guannan Li. 2010. Monitoring Social Expectations in Second Life. In *Coordination, Organizations, Institutions and Norms in Agent Systems V*, Julian Padget, Alexander Artikis, Wamberto Vasconcelos, Kostas Stathis, Viviane Torres da Silva, Eric Matson, and Axel Polleres (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 133–146. https://doi.org/10.1007/978-3-642-14962-7_9

[11] Edvinas Danevičius, Frederik Stief, Konrad Matynia, Morten Læburgh Larsen, and Martin Kraus. 2021. 3D Localisation of Sound Sources in Virtual Reality. In *Interactivity and Game Creation: 9th EAI International Conference, ArtsIT 2020, Aalborg, Denmark, December 10–11, 2020, Proceedings 9*. Springer, Springer International Publishing, NY, USA, 307–319.

[12] E. Datteri, G. Teti, C. Laschi, G. Tamburrini, G. Dario, and E. Guglielmelli. 2003. Expected perception: an anticipation-based perception-action scheme in robots. *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)* 1 (2003), 934–939 vol.1. https://doi.org/10.1109/IROS.2003.1250748

[13] Virginie Demeure, Radosław Niewiadomski, and Catherine Pelachaud. 2011. How is believability of a virtual agent related to warmth, competence, personification, and embodiment? *Presence* 20, 5 (2011), 431–448. https://doi.org/10.1162/PRES_a_00065

[14] Nicolò Dozio, Emanuela Maggioni, Dario Pittera, Alberto Gallace, and Marianna Obrist. 2021. May I smell your attention: Exploration of smell and sound for visuospatial attention in virtual reality. *Frontiers in psychology* 12 (2021), 671470. https://doi.org/10.3389/fpsyg.2021.749419

[15] Jon Driver and Charles Spence. 1998. Crossmodal attention. *Current opinion in neurobiology* 8, 2 (1998), 245–253. https://doi.org/10.1016/S0959-4388(98)80147-5

[16] Richard P Ebstein, Salomon Israel, Soo Hong Chew, Songfa Zhong, and Ariel Knafo. 2010. Genetics of human social behavior. *Neuron* 65, 6 (2010), 831–844. https://doi.org/10.1016/j.neuron.2010.02.020

[17] Autumn Edwards, Chad Edwards, David Westerman, and Patric R Spence. 2019. Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior* 90 (2019), 308–314. https://doi.org/10.1016/j.chb.2018.08.042

[18] Rachel Ege, A Opstal, and Marc M Van Wanrooij. 2018. Accuracy-precision trade-off in human sound localisation. *Scientific reports* 8, 1 (2018), 1–12.

[19] Patrick Flanagan, Ken I McAnally, Russell L Martin, James W Meehan, and Simon R Oldfield. 1998. Aurally and visually guided visual search in a virtual environment. *Human factors* 40, 3 (1998), 461–468. https://doi.org/10.1518/001872098779591331

[20] Maia Garau, Mel Slater, David-Paul Pertaub, and Sharif Razzaque. 2005. The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators & Virtual Environments* 14, 1 (2005), 104–116. https://doi.org/10.1162/1054746053890242

[21] James Jerome Gibson and James J Gibson. 1986. *The ecological approach to visual perception*. Vol. 1. Psychology Press, New York, USA. https://doi.org/10.4324/9780203767764

[22] Pengfei Huang, Mubbasir Kapadia, and Norman I. Badler. 2013. SPREAD: Sound Propagation and Perception for Autonomous Agents in Dynamic Environments. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Anaheim, California) *(SCA '13)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/2485895.2485911

[23] Finn Jacobsen. 1997. An overview of the sources of error in sound power determination using the intensity technique. *Applied Acoustics* 50, 2 (1997), 155–166. https://doi.org/10.1016/S0003-682X(96)00047-3 Sound Intensity Measurement.

[24] Alberto Jovane, Pierre Raimbaud, Katja Zibrek, Claudio Pacchierotti, Marc Christie, Ludovic Hoyet, Anne-Hélène Olivier, and Julien Pettré. 2023. Warping character animations using visual motion features. *Computers & Graphics* 110 (2023), 38–48. https://doi.org/10.1016/j.cag.2022.11.008

[25] George F Kuhn. 1977. Model for the interaural time differences in the azimuthal plane. *the Journal of the Acoustical Society of America* 62, 1 (1977), 157–167. https://doi.org/10.1121/1.381498

[26] Minae Kwon, Malte F Jung, and Ross A Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, IEEExplore, 3 Park Avenue, 17th Floor, NY, USA, 463–464. https://doi.org/10.1109/HRI.2016.7451807

[27] Stylianos Mystakidis. 2020. Distance education gamification in social virtual reality: A case study on student engagement. In *2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA*. IEEE, IEEExplore, NY, USA, 1–6. https://doi.org/10.1109/IISA50023.2020.9284417

[28] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. 2020. Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI* 7 (2020), 92. https://doi.org/10.3389/frobt.2020.00092

[29] Sandra Poeschl and Nicola Doering. 2015. Measuring co-presence and social presence in virtual environments–psychometric construction of a german scale for a fear of public speaking scenario. *Annual Review of Cybertherapy and Telemedicine 2015* 219, 1 (2015), 58–63.

[30] Pierre Raimbaud, Alberto Jovane, Katja Zibrek, Claudio Pacchierotti, Marc Christie, Ludovic Hoyet, Julien Pettré, and Anne-Hélène Olivier. 2021. Reactive Virtual Agents: A Viewpoint-Driven Approach for Bodily Nonverbal Communication. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Japan) *(IVA '21)*. Association for Computing Machinery,

New York, NY, USA, 164–166. https://doi.org/10.1145/3472306.3478351

[31] Lazlo Ring, Dina Utami, and Timothy Bickmore. 2014. The Right Agent for the Job?. In *Intelligent Virtual Agents*, Timothy Bickmore, Stacy Marsella, and Candace Sidner (Eds.). Springer International Publishing, Cham, 374–384. https://doi.org/10.1007/978-3-319-09767-1_49

[32] Verónica Sevillano and Susan T. Fiske. 2016. Animals as Social Objects. *European Psychologist* 21, 3 (2016), 206–217. https://doi.org/10.1027/1016-9040/a000268

[33] Mincheol Shin, Stephen W Song, Se Jung Kim, and Frank Biocca. 2019. The effects of 3D sound in a 360-degree live concert video on social presence, parasocial interaction, enjoyment, and intent of financial supportive action. *International Journal of Human-Computer Studies* 126 (2019), 81–93. https://doi.org/10.1016/j.ijhcs.2019.02.001

[34] Barbara G Shinn-Cunningham. 2008. Object-based auditory and visual attention. *Trends in cognitive sciences* 12, 5 (2008), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

[35] Sean Thomas, Ylva Ferstl, Rachel McDonnell, and Cathy Ennis. 2022. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEExplore, 3 Park Avenue, 17th Floor, NY, USA, 11–20. https://doi.org/10.1109/VR51125.2022.00018

[36] Elena Torta, Jim van Heumen, Raymond H Cuijpers, and James F Juola. 2012. How can a robot attract the attention of its human partner? a comparative study over different modalities for attracting attention. In *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012. Proceedings 4*. Springer Berlin Heidelberg, Springer, Berlin, Germany, 288–297. https://doi.org/10.1007/978-3-642-34103-8_29

[37] Alexandra Voinescu, Liviu Andrei Fodor, Danaë Stanton Fraser, and Daniel David. 2020. Exploring attention in vr: effects of visual and auditory modalities. In *Advances in Usability, User Experience, Wearable and Assistive Technology: Proceedings of the AHFE 2020 Virtual Conferences on Usability and User Experience, Human Factors and Assistive Technology, Human Factors and Wearable Technologies, and Virtual Environments and Game Design, July 16-20, 2020, USA*. Springer, Springer International Publishing, NY, USA, 677–683. https://doi.org/10.1007/978-3-030-51828-8_89

[38] MinJuan Wang, Sus Lundgren Lyckvi, Chenhui Chen, Palle Dahlstedt, and Fang Chen. 2017. Using advisory 3D sound cues to improve drivers' performance and situation awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, NY, USA, 2814–2825. https://doi.org/10.1145/3025453.3025634

[39] Yu Wang, Mubbasir Kapadia, Pengfei Huang, Ladislav Kavan, and Norman I Badler. 2014. Sound localization and multi-modal steering for autonomous virtual agents. In *Proceedings of the 18th meeting of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. Association for Computing Machinery, New York, NY, USA, 23–30. https://doi.org/10.1145/2556700.2556718

[40] Katja Zibrek, Sean Martin, and Rachel McDonnell. 2019. Is photorealism important for perception of expressive virtual humans in virtual reality? *ACM Transactions on Applied Perception (TAP)* 16, 3 (2019), 1–19. https://doi.org/10.1145/3349609