# Doctoral Consortium: Verbal Interactions with Embodied Conversational Agents

Jonathan Ehret
ehret@vr.rwth-aachen.de
Visual Computing Institute, RWTH Aachen University
Aachen, Germany

## KEYWORDS

verbal communication, non-verbal communication, gestures, turn-taking, embodied conversational agents (ECAs), virtual acoustics, prosody, directivity, social presence, face-to-face communication, virtual reality

**ACM Reference Format:**
Jonathan Ehret. 2022. Doctoral Consortium: Verbal Interactions with Embodied Conversational Agents. In *Proceedings of IVA '22: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents (Doctoral Consortium at IVA '22).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Embedding virtual humans into virtual reality (VR) applications can fulfill diverse needs. These, so-called, embodied conversational agents (ECAs) can simply enliven the virtual environments, act for example as training partners [11], tutors [4] or therapists [5], or serve as advanced (emotional) user interfaces to control immersive systems. The latter case is of special interest since we as human users are specifically good at interpreting other humans. ECAs can enhance their verbal communication with non-verbal behavior and thereby make communication more efficient. For example, backchannels [2, 3, 12], like nodding or signaling not understanding, can be used to give feedback while a user is speaking. Furthermore, gestures, gaze, posture, proxemics and many more [18] no-verbal behaviors can be applied. Additionally, turn-taking can be streamlined when the ECA understands when to take over the turn and signals willingness to yield it once done [13]. While many of these aspects are already under investigation from very different disciplines, operationalizing those into versatile, virtually embodied human-computer interfaces remains an open challenge. To this end, I conducted several studies investigating acoustical effects of ECAs' speech, both with regard to the auralization in the virtual environment and the speech signals used. Furthermore, I want to find guidelines for expressing both turn-taking and various backchannels that make interactions with such advanced embodied interfaces more efficient and pleasant, both when the ECA is speaking and during listening. Additionally, measuring social presence (i.e., the feeling of being there and interacting with a "real" person [14]) is an important instrument for this kind of research, since I want to facilitate exactly those subconscious processes of understanding other humans, which we as humans are particularly good at. Therefore, I want to investigate objective measures for social presence (see [14]).

In the following, I will give a brief overview of my accomplished and planned research and some methodology I develop and use for that.

## 2 RESEARCH AGENDA

In this section I will outline my research agenda. Therefore, I will first summarize research I performed in the first years of my doctoral studies, namely investigating different qualitative aspects of auralizing an ECA's speech. Afterwards, I will give an outlook on how I plan to proceed.

### 2.1 Virtual Agents' Speech

As verbal communication requires acoustic signals to be presented to the user, the question arises how those should be presented. Modern VR applications often include binaural audio rendering. Here, the position of the sound source relative to the user's ears is considered, so when the user looks around the sound presented to each ear changes [16]. While this already increases realism, the directionality of the sound sources is often not taken into account. In reality, for example, human speakers sound muffled and less loud when they speak away from the listener as compared to when they were directly facing the listener. I investigated whether modeling this directivity for virtual human speakers influences their perceived social presence. To that end, I teamed up with colleagues from the Institute of Hearing Technology and Acoustics to conduct a study [19, 20] consisting of an interaction between one ECA and the participant in an object search task which involved a lot of turning of the ECA and listening to the agent from very different angles. However, I was unable to find a significant difference in the perceived social presence when comparing omnidirectional speech sound sources with those using directivity. Since I mainly attributed this to an overall too low realism and a high variance in the examined measures (e.g., Social Presence Survey [1]), I conducted a second study. This time also additionally considering dynamic directivity [8], i.e., dynamically changing the directivity pattern in accordance with the currently uttered phoneme. In this study, we found indications that participants were unable to distinguish dynamic from static (non-changing) directivity, but were well able to distinguish it from omnidirectional speech sources. However, there was no clear preference for the more realistic directional case.

The aforementioned studies were conducted using both synthetic and recorded speech (in [20] and [8] respectively). Therefore,

*Doctoral Consortium at IVA '22, Sept. 6–9, 2022, Faro, Portugal*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

I also wanted to investigate how large the influence of this synthetic voice, and especially the often slightly unnatural prosody produced by Text-to-Speech (TTS) engines, is on the perceived naturalness of the ECAs (cf. [15]). Furthermore, I investigated whether seeing a virtual ECA presenting the speech moderates this effect, as it might change expectations [6]. I conducted an online survey in cooperation with colleagues from the linguistics and acoustics department, presenting videos of two conversing ECAs or audio-only to participants, with natural speech, fully synthetic speech, or human speakers imitating the unnatural prosody of the TTS engine used [6]. The ECAs were however in contrast to the aforementioned studies not aware of the participants listening. The results indicated that inadequate prosody (as often produced by TTS engines) has a strong influence on perceived naturalness. Thus, I will focus on recorded speech for further research. Furthermore, although the results were not perfectly conclusive, I will use static directivity for speaking ECAs.

## 2.2 Virtual Agents' Non-verbal Behavior

Additionally to the verbal behavior already looked at, I want to investigate non-verbal behavior of conversing ECAs. I especially want to look into turn-taking signaled by ECAs and their co-verbal gestures. I teamed up, additionally to colleagues from the acoustics institute, with colleagues from the psychology department. The psychologists developed a task (i.e., heard text recall (HTR) task [10]), during which a participant has to listen to a family story being presented by two speakers (see Figure 1) and then has to answer several questions regarding family constellations etc., which sometimes can only be answered by combining information from several utterances. Additionally to these questions, a dual-task is performed (cf. [10]) to evaluate listening effort. The plan of this project is to investigate whether this task is valuable to research listening effort in a closer-to-real-life situation compared to classical psychological paradigms. Furthermore, the influence of several aspects regarding acoustic and visual fidelity (e.g., display technology but also ECA behavior) should be investigated [9]. I plan to furthermore explore whether performance in this task (both in the primary task, i.e., answering the question, and secondary task, e.g., reacting to vibrotactile stimuli while listening), can be used as a more objective measure of social presence. I hypothesize that observing ECAs who behave differently from what we would expect from real humans might introduce additional cognitive load and thereby deduce performance in the task. In the planned experiments, I want to use both fitting and non-fitting co-verbal gestures presented by the ECAs while speaking. Since the ECAs take turns while presenting, my goal is to develop a set of rules of understandable non-verbal turn-taking cues given by the ECAs, including gestures, gazing, and breathing (cf. [17]). I plan to evaluate whether participants are able to correctly interpret those in a user study [7], where the ECA will be controlled in a Wizard-of-Oz paradigm. After evaluating and potentially improving those turn-taking cues, I will investigate whether signaling turn-taking at inappropriate points in time, e.g., signal yielding the turn when the ECA actually continues with the next utterance, increases the aforementioned listening effort. If such correlations can be found, this can be a valuable tool to objectify perceived social presence beyond the currently used questionnaires (cf. [14]) for several dimensions of non-verbal behavior (namely planned for turn-taking, co-verbal gestures, and potentially backchanneling).

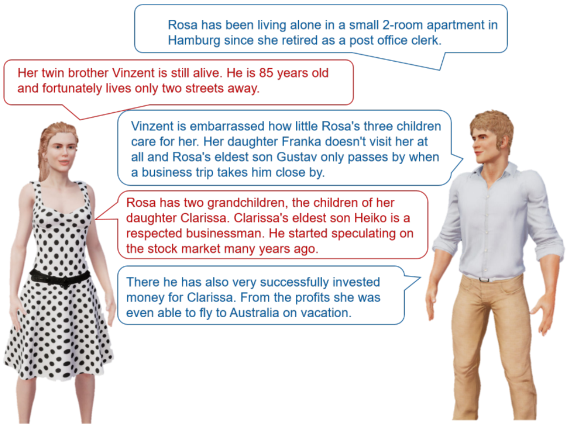
Rosa has been living alone in a small 2-room apartment in Hamburg since she retired as a post office clerk.
Her twin brother Vinzent is still alive. He is 85 years old and fortunately lives only two streets away.
Vinzent is embarrassed how little Rosa's three children care for her. Her daughter Franka doesn't visit her at all and Rosa's eldest son Gustav only passes by when a business trip takes him close by.
Rosa has two grandchildren, the children of her daughter Clarissa. Clarissa's eldest son Heiko is a respected businessman. He started speculating on the stock market many years ago.
There he has also very successfully invested money for Clarissa. From the profits she was even able to fly to Australia on vacation.

**Figure 1: Excerpt of a family story told by two ECAs as part of the HTR paradigm [9].**

# 3 INFRASTRUCTURE

A lot of animation data is needed for the ECAs. I already recorded facial animations when recording two actors giving the HTR task's speech by means of the *ARKit* using an *iPhone's TrueDepth Sensor*. Since I do not have direct access to an optical motion capture system, body animations are recorded using an *HTC Vive* headset in combination with 6 *Vive Trackers* and two *Vale Index Controllers*, which are capable of rudimentary tracking the fingers.

To allow an efficient implementation of the planned user studies, I develop a study framework for Unreal Engine. This framework facilitates to quickly set up multi-factorial studies, supporting the developer in randomization, logging, and controlling the study in general by providing experimenter interfaces. It is already made available to my collaborators and several students have used it to conduct experiments for their theses. I already systematically gathered overall positive feedback from the aforementioned users and a formal validation is planned in the future.

# 4 CONCLUSION

In this extended abstract, I gave a brief overview of my research agenda towards creating efficient and pleasant to use virtually embodied human-computer interfaces by means of ECAs. I introduced research I conducted with regard to speech source directivity and prosody. Furthermore, I outlined a future research path using a two talker scenario to investigate turn-taking, co-verbal gestures and the HTR dual-task paradigm for potentially objectively measuring social presence.

# ACKNOWLEDGMENTS

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SPP2236 AUDICTIVE – 444724862, Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments.

# REFERENCES

[1] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. 2001. Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments. *Presence: Teleoperators and Virtual Environments* 10, 6 (2001), 583–598. https://doi.org/10.1162/105474601753272844

[2] Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, Marc Schröder, and Catherine Pelachaud. 2010. Multimodal Backchannels for Embodied Conversational Agents. In *International Conference on Intelligent Virtual Agents*. 194–200. https://doi.org/10.1007/978-3-642-15892-6_21

[3] Hendrik Buschmeier and Stefan Kopp. 2018. Communicative Listener Feedback in Human-Agent Interaction: Artificial Speakers Need to Be Attentive and Adaptive Socially Interactive Agents Track. In *Proc. ofthe 17th International Conference on Autonom- ous Agents and Multiagent Systems (AAMAS 2018)*. 1213–1221. www.ifaamas.org

[4] Susan Wagner Cook, Howard S. Friedman, Katherine A. Duggan, Jian Cui, and Voicu Popescu. 2017. Hand Gesture and Mathematics Learning: Lessons From an Avatar. *Cognitive Science* 41, 2 (2017), 518–535. https://doi.org/10.1111/cogs.12344

[5] David Devault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 1061–1068. http://simsensei.ict.usc.edu/

[6] Jonathan Ehret, Andrea Bönsch, Lukas Aspöck, Christine T. Röhr, Stefan Baumann, Martine Grice, Janina Fels, and Torsten W. Kuhlen. 2021. Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech? *ACM Transactions on Applied Perception* 18, 4 (oct 2021), 21:1–15. https://doi.org/10.1145/3486580

[7] Jonathan Ehret, Andrea Bönsch, and Torsten W. Kuhlen. 2022. Natural Turn-Taking with Embodied Conversational Agents. In *IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)*. https://doi.org/10.1080/14992027.2021.1922765

[8] Jonathan Ehret, Jonas Stienen, Chris Brozdowski, Andrea Bönsch, Irene Mittelberg, Michael Vorländer, and Torsten W. Kuhlen. 2020. Evaluating the Influence of Phoneme-Dependent Dynamic Speaker Directivity of Embodied Conversational Agents' Speech. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, IVA 2020*. Association for Computing Machinery, Inc. https://doi.org/10.1145/3383652.3423863

[9] Janina Fels, Cosima A Ermert, Jonathan Ehret, Chinthusa Mohanathasan, Andrea Bönsch, Torsten W Kuhlen, and Sabine J Schlittmeier. 2021. Listening to, and Remembering Conversations between Two Talkers: Cognitive Research using Embodied Conversational Agents in Audiovisual Virtual Environments. In *47. Jahrestagung für Akustik, Wien (Austria), 15 Aug 2021 - 18 Aug 2021*. 1328–1331.

[10] Edina Fintor, Lukas Aspöck, Janina Fels, and Sabine J. Schlittmeier. 2021. The role of spatial separation of two talkers' auditory stimuli in the listener's memory of running speech: listening effort in a non-noisy conversational setting. *International Journal of Audiology* (2021), 1–9. https://doi.org/10.1080/14992027.2021.1922765

[11] Jonathan Gratch, David DeVault, and Gale Lucas. 2016. The Benefits of Virtual Humans for Teaching Negotiation. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents*. 283–294. https://doi.org/10.1007/978-3-319-47665-0_25

[12] Lixing Huang, Louis Philippe Morency, and Jonathan Gratch. 2011. Virtual rapport 2.0. In *Intern. Workshop on Intell. Virtual Agents*. Springer, Berlin, Heidelberg, 68–79. https://doi.org/10.1007/978-3-642-23974-8_8

[13] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling?. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*. ACM, New York, NY, USA, 131–138. https://doi.org/10.1145/3472306.3478360

[14] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Front. Robot. AI* 5, 114 (2018). https://doi.org/10.3389/frobt.2018.00114

[15] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction. *Comput. Surveys* 54, 4 (2021), 1–43. https://doi.org/10.1145/3386867

[16] Stefania Serafin, Michele Geronazzo, Cumhur Erkut, Niels C. Nilsson, and Rolf Nordahl. 2018. Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions. *IEEE Comput. Graph.* 38, 2 (mar 2018), 31–43. https://doi.org/10.1109/MCG.2018.193142628

[17] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech and Language* 67, 101178 (2021). https://doi.org/10.1016/J.CSL.2020.101178

[18] Isaac Wang and Jaime Ruiz. 2021. Examining the Use of Nonverbal Communication in Virtual Agents. *International Journal of Human–Computer Interaction* (2021), 1–26. https://doi.org/10.1080/10447318.2021.1898851

[19] Jonathan Wendt, Benjamin Weyers, Andrea Bönsch, Jonas Stienen, Tom Vierjahn, Michael Vorländer, and Torsten W. Kuhlen. 2018. Does the Directivity of a Virtual Agent's Speech Influence the Perceived Social Presence?. In *Virtual Humans and Crowds for Immersive Environments (VHCIE), IEEE*.

[20] Jonathan Wendt, Benjamin Weyers, Jonas Stienen, Andrea Bönsch, Michael Vorländer, and Torsten W. Kuhlen. 2019. Influence of Directivity on the Perception of Embodied Conversational Agents' Speech. In *Proc. Int. Conf. Intell. Virtual Agents*. ACM, 130–132. https://doi.org/10.1145/3308532.3329434