Influence of Directivity on the Perception of Embodied Conversational Agents' Speech

Jonathan Wendt*
Visual Computing Institute,
RWTH Aachen University, Germany

Benjamin Weyers
Human-Computer Interaction,
University of Trier, Germany

Jonas Stienen
Institute of Technical Acoustics,
RWTH Aachen University, Germany

Andrea Bönsch
Visual Computing Institute,
RWTH Aachen University, Germany

Michael Vorländer
Institute of Technical Acoustics,
RWTH Aachen University, Germany

Torsten W. Kuhlen
Visual Computing Institute,
RWTH Aachen University, Germany

(a) Speaker Directivity Visualized in an Exemplary Interaction (b) Virtual Stockroom

Figure 1: (a) A participant holding a picked up item (within the blue sphere), which he was asked for by the agent. The directivity of the agent's speech sound source is exemplarily visualized, for a more precise visualization of the used directivity see Fig. 2. (b) Top view of the stockroom, with an agent standing next to the scanner and shelves filled with 237 collectable packages.

ABSTRACT

Embodied conversational agents become more and more important in various virtual reality applications, e.g., as peers, trainers or therapists. Besides their appearance and behavior, appropriate speech is required for them to be perceived as human-like and realistic. Additionally to the used voice signal, also its auralization in the immersive virtual environment has to be believable. Therefore, we investigated the effect of adding directivity to the speech sound source. Directivity simulates the orientation dependent auralization with regard to the agent's head orientation. We performed a one-factorial user study with two levels (n=35) to investigate the effect directivity has on the perceived social presence and realism of the agent's voice. Our results do not indicate any significant effects regarding directivity on both variables covered. We account this partly to an overall too low realism of the virtual agent, a not overly social utilized scenario and generally high variance of the examined measures. These results are critically discussed and potential further research questions and study designs are identified.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; *User studies.*

KEYWORDS

virtual agents; virtual acoustics; directional 3D sound; social presence

ACM Reference Format:
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 *ACM International Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, PARIS, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308532.3329434

1 INTRODUCTION

Embedding human-like virtual agents into virtual reality applications becomes more frequent since they can be of avail for various tasks. Embodied conversational agents (ECAs) can act as trainers, interviewers, peers or simply enliven virtual environments (e.g., [3]). To this end, they should behave and appear naturally and plausibly. Therefore, not only their visuals and behavior are important but also the sounds created by them, e.g., life-like sounds like breathing [2], physical sounds like footsteps and rustling of the ECAs' clothes, or their vocalization. The latter is of particular importance when

*email: wendt@vr.rwth-aachen.de

IVA '19, July 2–5, 2019, PARIS, France

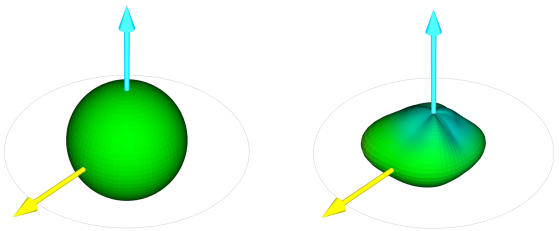
ACM ISBN 978-1-4503-6672-4/19/07.
https://doi.org/10.1145/3308532.3329434

interacting with ECAs. While many applications enable the user to perceive the location of a speech sound source, e.g., by using binaural sound synthesis [5], the speaker's orientation is often ignored. However, most natural sound sources do not radiate the sound uniformly into all directions but have some directional filtering effect, e.g., a human speaker sounds clearest when talking directly towards the listener while sounding muffled and less loud when talking away from the listener.

Rendering ECAs more realistically can improve the perceived co-presence of ECAs, as shown by Zibrek et al. [10]. Furthermore, Shin et al. [8] found that a more realistic 3D sound also has a positive influence on social presence. While they used recorded 3D sound, which naturally incorporates the sound source directivity of the recorded sources in the examined static setup, this has to be actively simulated for dynamically moving ECAs. As Mehra et al. [6] stated, sound source directivities improve the realism of the auralization and might thereby, if also applied to ECAs' speech, improve their acoustic realism and thus their perceived social presence. Therefore, we designed a within-subjects user study in which participants interact with an ECA in a virtual stockroom. The study's goal was to examine whether humans (subconsciously) prefer ECAs that are auralized using this directional effect and whether they are rated as being more realistic.

2 STUDY

We conducted a within-subjects study in a five-sided CAVE (5.25m× 5.25m), based on the findings of a pre-study by Wendt et al. [9], which examined the same effect but in a less plausible scenario. The study had one independent variable: *Auralization*, with the two levels *omnidirectional* and *directional*. So, we either used an omnidirectional sound radiation pattern or a measured directivity pattern of a human singer as provided by Kob [4], one per condition in a counterbalanced order. The participants were placed in a virtual stockroom, with the same dimensions as the CAVE, together with a male ECA (see Fig. 1(b)). The stockroom was filled with five rows of shelves with in total 237 packages of 16 different types and a scanner terminal at which the ECA stood during the entire interaction. Participants had to collect all packages for one of two predefined orders requiring to collect 17 items in total each. Therefore, the ECA read out what is needed next while looking at the terminal, e.g., *"For this order, we also have to find a teddy bear."*. In total, 10 requests per order were given, where some required multiple items of the same kind to be collected (cf. supplementary video [1]). Unnoticeable for the participants, they were observed by the experimenter by means of real-time video and audio streams. This enabled the experimenter to control the ECA via a Wizard-of-Oz paradigm by triggering reasonable utterances depending on the participants' actions or questions using natural language. One of these utterances per request included a turning of the ECA towards the shelf row where the requested item was placed and thereby highlighted the directivity pattern, since the direction-dependent changes, if present, are most noticeable during turning. Thereby the ECA gave the participants incremental hints for what and where to look. After finishing each order, the participants had to answer a questionnaire with subjective measures while staying in the CAVE. We measured the perceived social presence of the ECAs using the social presence score (SPS) questionnaire [1] and questions



(a) Omnidirectional (b) Human Singer Directivity

Figure 2: The used directivities, shown here at 125Hz with the yellow arrow pointing forward and the cyan one upward

regarding speech realism for each condition. SPS was used since it is, to our knowledge, the best available questionnaire measuring the concept of social presence, and the task was designed to elicit social interaction. Additionally to the subjective measures, we measured the minimal distance participants kept to the ECA, as these proxemics could potentially also be used to gain objective insights in the perceived social presence [1]. The time needed for each condition was also logged, as task-related measure.

After leaving the CAVE, participants were asked to fill out a post-study questionnaire asking them what they think was investigated and several questions to rate the experience and which of the two conditions they liked better concerning different aspects of the ECA.

To render and animate the ECA, we used *SmartBody* [7] and its human model *Brad*. The speech audio was generated by means of *Google Cloud Text-To-Speech*[2] and the *Sphinx-4* library[3] was used to generate the matching lip sync data for *SmartBody*'s lip-syncing. Since the ECA was, by design, often talking towards the wall, we also included the room response to the acoustical speech signals into the auralization. As directivity filter, a measured directivity of a human singer [5] was used (see Fig. 2(b)). This directivity filter changes the sound of the speech related to the orientation of the speaker, e.g., damping specific frequencies when the ECA is facing away from the participant. Furthermore, for the omnidirectional condition a directivity was used that is uniform in all directions and frequencies (see Fig. 2(a)) and is normalized to have the same amplitude in the frontal direction as the singer's directivity. Using directional filters for both conditions guaranteed that the acoustic signals were processed equally for both conditions and no difference was introduced by additional filtering. We confirmed with expert listening tests, that the difference between the auralization conditions was well noticeable.

3 RESULTS AND DISCUSSION

We conducted the study with 35 participants (9 female and 26 male, mean age = 24.61 years, SD = 4.18). Unfortunately, our results did not reveal any significant major effect. The SPS measures do not show significant differences between the two different auralization methods. Furthermore, when asked for a preference of one of the conditions, participants did not state a significant preference for either one. When analyzing the speech sound realism ratings that were posed right after each condition or the comparative ones that were posed after finishing both, no significant effect on the perceived realism could be found. The interference that led to these results could be manifold.

[1] https://youtu.be/noAF_ZB0_oQ

[2] https://cloud.google.com/text-to-speech/

[3] https://cmusphinx.github.io/

Although we designed our scenario to foster participant's engagement into a natural conversation with the ECA, only 34% engaged in a real conversation. By this, probably fewer subjects may have noticed the altered auralization. This might be caused due to the fact that participants deemed the task too simple and wanted to solve it on their own instead of asking for help, albeit being encouraged to ask in the task description. However, only considering cooperative participants does not yield any significant observations either.

Another challenge we faced is that evaluating the influence of the subtle auralization change using objective measures is complicated. We did not expect to find an influence on the completion time, since a higher realism of the voice should not change the difficulty of the task. Furthermore, the other objective measure, the kept distance to the ECA, did not yield any insightful information either. This is probably due to the fact that proxemics is a better measure to differentiate effects of eeriness [10] than subtle changes of the voice.

Generally, the studied auralization techniques are advanced with regard to sound realism. Therefore, some participants remarked that the overall realism of the interaction and the scene was too low, thus the added realism due to the directivity was unnoticeable with the applied measures. This becomes also apparent in the answers to the realism questionnaire items and the free field comments, which span from *"The voice of the ECA was also reducing the quality of the experience, it was very 'robotic' sounding."* to *"I did not focus on the speech/sound at all, as the speech/sound itself was the most natural thing in my opinion"*. Furthermore, this also hints to another deficiency, the usage of a synthetic TTS voice. We found that potentially more significant results could have been gathered if a recorded voice would have been used. On the other hand, recording a human in a natural way is also hard to achieve and would require trained native speakers. Thus, in order to improve our study design to gain more insight, we recommend using a recorded human voice for pre-defined utterances in follow-up studies.

While some participants stated that they found the ECA too silent and therefore hard to understand, in general, the participants found the ECA well understandable. However, picking the right loudness for the omnidirectional condition is hard, since omnidirectional speech sound sources are an artificial concept. We picked it such that the loudness is equal for both conditions when the ECA talks directly towards the participant. This, however, means that the accumulated sound energies the ECA radiates into the scene are different, because the directional filter damps the radiation in non-frontal directions. If then again these accumulated energies are matched the ECA using a directional radiation pattern would sound louder when directly talking towards the participant, which is even more noticeable and might distort the results since a louder ECA is better understandable.

The participants, who were left naïve to the investigated effect, were asked to speculate on the purpose of the study. Only 6 participants (17%) suspected that the study investigated the effect of ECAs' speech or movement. When told afterwards what the investigated effect was, one participant stated that he noticed the directivity effect during the study when the ECA was turning at least once. However, he also stated that he did not notice the absence of it. This potentially means that directivity can slightly increase the realism of speech, but normally users do not pay attention to it, especially if there are other aspects decreasing the realism of the virtual environment. This is also in line with the result that only 31% of the participant reported that they at all noticed a change in auralization and even for those no significant effects were apparent.

Another possible explanation for not finding any significant differences between the probed auralization conditions, is that there is no or only a small effect when adding directivity to the auralization of an ECA during the task at hand, at least on the perceived social presence and realism.

4 CONCLUSION

With this study, we intended to show that using directivity to auralize the speech of ECAs has an influence on the perceived realism of those and thereby on their social presence. However, the results did not reveal any significant effects. This can be partly accounted to the used scenario which did not force participants to engage in a natural and bi-directional conversation with the ECA and thereby had not focus them especially on the speech sound. Additionally, this could be caused by an overall too low realism in the context of which such an advanced technique only plays a marginal role.

We plan to examine further whether in a direct comparison between these conditions effects on perceived realism can be measured when participants are primed and know on what to focus. Therefore, we will design a study in which no artificially social task is involved, but only a monologue of an ECA is experienced, using recorded speech and motion. Beyond that, we want to examine the effect of dynamic directivities. That means that the directivity pattern is influenced by the currently uttered phoneme. It remains to examine whether that is at all distinguishable from the static directivity that was used in this paper.

ACKNOWLEDGMENTS

This work was funded by the project house ICT Foundations of a Digitized Industry, Economy, and Society at RWTH Aachen Univ.

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.

[2] Ulysses Bernardet, Sin-hwa Kang, Andrew Feng, Steve DiPaola, and Ari Shapiro. 2017. A Dynamic Speech Breathing System for Virtual Characters. In *Proceedings of 17th International Conference on Intelligent Virtual Agents*. Springer, Cham, 43–52.

[3] Andrea Bönsch, Jonathan Wendt, Heiko Overath, Ozgür Gürerk, Christine Harbring, Christian Grund, Thomas Kittsteiner, and Torsten W. Kuhlen. 2017. Peers at work: Economic real-effort experiments in the presence of virtual co-workers. In *IEEE Virtual Reality*. IEEE, 301–302.

[4] Malte Kob. 2002. *Physical Modeling of the Singing Voice.* Ph.D. Dissertation. RWTH Aachen University.

[5] Tobias Lentz. 2007. *Binaural technology for virtual reality.* Ph.D. Dissertation.

[6] Ravish Mehra, Lakulish Antani, Sujeong Kim, and Dinesh Manocha. 2014. Source and Listener Directivity for Interactive Wave-based Sound Propagation. *IEEE Transactions on Visualization and Computer Graphics* 20, 4 (2014), 495–503.

[7] Ari Shapiro. 2011. Building a character animation system. *Lecture Notes in Computer Science* 7060 LNCS (2011), 98–109.

[8] Mincheol Shin, Stephen W Song, Se Jung Kim, and Frank Biocca. 2019. Does sound make differences in an interpersonal relationship?: The Effects of 3D sound on Social Presence, Parasocial Relationship, Enjoyment, and Intent of Supportive Action. *International Journal of Human-Computer Studies* (2019).

[9] 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.*

[10] Katja Zibrek, Elena Kokkinara, and Rachel Mcdonnell. 2017. Don't Stand So Close To me: Investigating the effect of control on the appeal of virtual humans using immersion and a proximity-based behavioral task. In *Proceedings of the ACM Symposium on Applied Perception.*