Does the Directivity of a Virtual Agent's Speech Influence the Perceived Social Presence?

Jonathan Wendt* Benjamin Weyers* Andrea Bönsch* Jonas Stienen† Tom Vierjahn*
Michael Vorländer† Torsten W. Kuhlen*

* Visual Computing Institute, RWTH Aachen University, Germany
† Institute of Technical Acoustics, RWTH Aachen University, Germany
* JARA-HPC, Aachen, Germany

Abstract

When interacting and communicating with virtual agents in immersive environments, the agents' behavior should be believable and authentic. Thereby, one important aspect is a convincing auralization of their speech. In this work-in-progress paper a study design to evaluate the effect of adding directivity to speech sound source on the perceived social presence of a virtual agent is presented. Therefore, we describe the study design and discuss first results of a prestudy as well as consequential improvements of the design.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies

1 Introduction

In various current virtual reality applications, a need for believable conversational virtual agents (VAs) exists. They can for example represent interfaces to control and get feedback from an application [1] or present and teach [2]. As stated by Gratch et al., speech and lip-syncing are important aspects to create believable, embodied conversational agents [3]. Among other things this means that an appropriate auralization [4] for their voices is necessary. Auralization describes how the sound of different virtual sound sources is presented to a human listener who is immersed into a virtual audio scene. Therefore, often binaural auralization is used that renders two different sound signals, one for each ear considering the characteristics of a listener's head, the so called head-related transfer function (HRTF) [5]. It simulates the natural hearing experience, thus humans are able to localize sound sources in 3D space. Moreover the directivity of every sound source with regard to its orientation, i.e., the sound radiation characteristics considering direction and frequency, is important. For example, directivity has the effect of a human speaking towards oneself sounding louder and less muffled than one facing away. This directivity for humans can be measured as described by Kob [6] and can then be simulated for virtual sound sources during binaural rendering. Our research focus lies on the investigation of how changing the directivity of the speech sound source affects the subconsciously social presence of the VA perceived by an immersed human user.

In this work-in-progress paper we present the design of a study that aims at investigating this research question. Therefore, we alternate the way the speech of a VA is auralized between *omnidirectional*, where the orientation towards the listener is indistinguishable, and *directed*, where the effect of this orientation is audible.

*e-mail: {wendt|weyers|boensch|vierjahn|kuhlen}@vr.rwth-aachen.de
†e-mail: {jonas.stienen|michael.vorlaender}@akustik.rwth-aachen.de

Figure 1: Top view of the stockroom, with a virtual agent in the middle and 18 items placed on shelves. The items that are not on the top shelves are not visible.

We will first present the study design in Section 2 and then give and discuss the results of the conducted prestudy in Section 3. Conclusively we will discuss the next steps that we want to undertake to improve the study design in Section 4.

2 Study Design

We plan to conduct a within-subject study in a five-sided CAVE. It has two independent variables: *Auralization* and *Gender*. We use two different VAs (*male* and *female*) that are auralized using either no directivity (*omnidirectional*) or using the directivity of a human speaker (*directivity*) as provided by Kob [6]. This leads to four different conditions. The participants are placed in a virtual stockroom with one VA (see Figure 1). Participants have to fulfill a search task. Therefore, the VA utters demands for specific items in various directions facing towards and away from the participant to exhibit the different auralization techniques. We measure the perceived social presence of the VAs using the social presence score (SPS) questionnaire [7] after each condition which the participant can directly answer within the CAVE.

2.1 Experimental Design and Task

After reading a brief introduction and answering a demographic questionnaire, participants enter the CAVE and have the opportunity to familiarize with both VAs side-by-side in an otherwise empty

Figure 2: A participant returning an item, which is attached to the pointing device, to the virtual agent.

scene. The VAs say a brief welcoming sentence for the participants to get accustomed to their speech. Subsequently, a virtual stockroom is displayed (see Figure 1), which has exactly the size of the CAVE, so the participants can navigate by means of physical walking. The VA to be used in the respective condition is placed close to the middle of the room, so he/she does not stand directly in between two shelves and therefore gives the participant enough space to pass, avoiding collisions. The shelves are filled with several boxes and 18 items. This number was chosen so that the algorithm, explained below, has a sufficient number of items to pick from. The items are placed at well reachable places on the shelves, evenly spread around the VA. The items are randomly swapped between conditions to avoid learning effects.

The VA utters a request for one particular item at a time. Therefore, the VA first turns and looks towards the item he/she will ask for, before speaking a sentence like *"Please bring me the green basket"*. These sentences are predefined and differ slightly for each item, as recommended in [8]. By turning towards the item, differences in auralization become noticeable, since the sound does not change if the VA and thereby the directed sound source is facing the participant for every utterance. Following this idea, the item that has to be picked up next is determined based on the angle

$$\theta = \angle(\mathbf{P}_{item} - \mathbf{P}_{agent}, \mathbf{P}_{user} - \mathbf{P}_{agent})$$

which is based of the positions of the item ($\mathbf{P}_{item}$), the participant ($\mathbf{P}_{user}$) and the VA ($\mathbf{P}_{agent}$) projected onto the floor plane. Participants are asked to find 12 different items, that are chosen such that we have an equal number of cases with $\theta < 45^\circ$ (front), $45^\circ \leq \theta < 135^\circ$ (side) and $135^\circ \leq \theta$ (back), namely four items each. This way facing directions in all four quadrants around the VA based on the respective participant position are used. Left and right are not distinguished, since they do not exhibit different sounds related to directivity. To find the item to ask for next, we first randomly chose a facing direction and then pick one of the remaining items in that quadrant. Therefore, more items need to be present in the scene than should be picked.

The participant has to pick up the demanded item. Therefore, he/she has to walk towards it and use a grabbing metaphor with a pointing device (6DOF + Buttons). With the item attached to the pointing device (see Figure 2), the participant has to walk back to the VA and bring the item close enough to the VA. Once the items distance to the VA in the floor plane is below 40 cm it disappears, the VA faces the participant and randomly says one of the three predefined thank-you sentences. If not all 12 items have been picked up yet, the VA waits 1.5 s and then turns towards another item and asks for it. This time is chosen such that the VA has enough time to finish the thank-you sentence before he/she starts turning.

After finishing all 12 pick-ups the scene fades out and an empty scene (only a blue floor plane) with a questionnaire is displayed which the participant is asked to answer using the same pointing device. After each of the four conditions, participants are asked to answer the 5-item SPS questionnaire as proposed by Bailenson et al. [7]. As soon as the fourth questionnaire is answered, the familiarization scene is displayed again and the participants are asked to answer a post-study questionnaire outside of the CAVE asking for preferences of specific conditions.

The participants are equally distributed on the randomized sequences of conditions, to counter any order effects. Thereby both conditions of one *Gender* are always done right after each other, so it is potentially easier for the participants to specify their preference for *Auralization* in the post-study questionnaire. This leads in total to 8 different possible sequences. Additionally to the SPS questionnaires, the distance that the participants keep to the VA are measured and the minimal distance per condition is stored as an objective measure. Bailenson et al. stated that the distance kept, related to the personal space, can be used to measure perception of social presence [7]. Therefore, the virtual stockroom is deliberately designed in a way, that the participants, who are asked to avoid collision with virtual objects, have to pass close by the VAs (see Figure 1). This way we hope to measure a correlation between the SPS and the minimal distance.

2.2 Equipment

The study is conducted in a five-sided CAVE with a size of $5.25\,\mathrm{m} \times 5.25\,\mathrm{m} \times 3.30\,\mathrm{m}$ ($w \times d \times h$) in which the participant can walk freely since the virtual scene matches these dimensions. The participant wears tracked active stereo glasses and interacts with an ART Flystick 2. The ceiling is equipped with an acoustic system consisting of 12 studio loudspeakers and 9 sub-woofers. It can be used to generate two separate virtual sound sources next to the ears of the participant using crosstalk cancellation [9] to generate binaural audio. Furthermore two surveillance cameras are mounted at the ceiling with which the examiner can monitor the participant. To render and animate the VAs, *SmartBody* [10] is used, from which the human models *Brad* and *Rachel* are utilized. *SmartBody* can also perform lip-syncing. The speech audio is produced using the text-to-speech engine *CereVoice*[1].

3 Prestudy Results and Discussion

To evaluate the study design, we conducted a preliminary study with 8 participants executing this design. According to the post-study questionnaire, all but one did not realize which parameter we had tested, when asked afterwards what was changed between the conditions apart from the VAs' gender. However, when asked whether they noticed a difference in the auralization of the speech, five participants affirmed that they noticed some change. This indicates that in general this setup could be used to examine subconscious effects of auralization.

Furthermore, we noticed in the data logs, that the prerequisite of equally distributed utterance directions was violated in 9 of 32 conditions. In these cases, one of the directions was only used three times and thereby another direction five times. This should be adapted for the exhaustive study by adding more items and thereby more potential item directions to pick from. During the prestudy no one had problems finding the requested items. We also added the possibility to repeat the lastly uttered sentence, which however

[1] https://www.cereproc.com/

nobody utilized. Additionally, we noticed that many participants moved while the VAs were speaking, although asked not to do so in the task description. This might have influenced the hearing experience and should therefore be further prevented during the full study. However, having the VA turn while speaking could enhance the audibility of the directivity. For the agents' speech we used a synthetic voice. When asked afterwards about this synthetic voice, seven of the participants stated that they would have preferred recorded speech, since they state that the synthetic voice had both a negative influence on their feeling of being there and of interacting with a real person.

In the prestudy we tried to measure improvements of social presence. Additionally to the SPS questionnaires and minimal distances kept, we asked for the preference of individual conditions evaluated in the post-study questionnaire. However, when being asked for the preferred condition per gender after the study, only half of the participants had a specific preference, the others answered with *no preference* or *cannot remember*. Therefore, this question should potentially be embedded in the questionnaire directly after the second condition with the same gender, so they can potentially better remember any differences.

Looking at the recorded answers of the SPS questionnaires, no trends for improvements of social presence between *omnidirectional* and *directivity* conditions is noticeable. The difference in SPS ($SPS_{dir} - SPS_{oni}$), between all pairs of conditions has a mean of -0.2 (SD: 4.01), while SPS can take on values from -15 to 15. This does not seem very promising for a large scale study, so better choices for questionnaires or scores should be evaluated to measure an effect if it exists at all. Furthermore, the considered minimal distances to the VAs seem to rather exhibit order effects (M: -0.076 SD: 0.069) than being influenced by the *Directivity* (M: 0.025 SD: 0.102). So probably an additional training condition should be added for each *Gender* before the two conditions varying the *Directivity* begin. Then again, this objective distance measure might not prove insightful after all. We used VAs with different genders to counter gender effects, which at least for the distance kept to a VA cannot be ruled out (cf. [11]). Furthermore this allowed us to gather more data using different synthetic voices. Besides, we will evaluate the possibility to use recorded speech as most of the participants rated the synthetic voice to decrease their perceived social presence. Therefore, it is not entirely clear how this rating influences the effect we want to investigate.

4 Next Steps

As mentioned before, further steps before conducting a complete study are to evaluate other options for social presence questionnaires to potentially get more reliable results. This could be beneficial since some of the participants complained about not being able to consistently answer the used SPS questionnaire. Moreover, some more items will be placed in the scene to achieve a consistent sampling of speaking directions. Furthermore, the preference question after two conditions with the same gender will be posed right after those, to hopefully reduce indifferent answers. A longer familiarization phase within the virtual stockroom before the two conditions per VA should be introduced to further reduce habituation effects. Additionally, it will be evaluated whether using real, prerecorded speech improves the perceived social presence of the VA. Lastly we will evaluate whether having the VA turn towards the requested item while speaking will emphasize audible directivity effects further.

Acknowledgments

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

References

[1] E. André and C. Pelachaud, "Interacting with Embodied Conversational Agents," in *Speech Technology*. Springer US, 2010, pp. 123–149.

[2] S. Kopp, L. Gesellensetter, N. C. Krämer, and I. Wachsmuth, "A Conversational Agent as Museum Guide - Design and Evaluation of a Real-World Application," in *Intern. Workshop on Intell. Virtual Agents*, 2005, pp. 329–343.

[3] J. Gratch, J. Rickel, E. André, J. Cassell, E. Petajan, and N. Badler, "Creating Interactive Virtual Humans: Some Assembly Required," *IEEE Intell. Sys.*, 2002.

[4] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization An Overview," *J. Audio Engin Soc*, vol. 41, no. 11, pp. 861–875, 1993.

[5] D. R. Begault, "3-D Sound for Virtual Reality and Multimedia," 2000.

[6] M. Kob, "Physical Modeling of the Singing Voice," Ph.D. dissertation, RWTH Aachen University, 2002.

[7] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis, "Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments," *Presence: Teleoperators and Virtual Environments*, vol. 10, no. 6, pp. 583–598, 2001.

[8] A. Bönsch, T. Vierjahn, and T. W. Kuhlen, "Evaluation of Approaching-Strategies of Temporarily Required Virtual Assistants in Immersive Environments," in *IEEE Symp. 3D User Interfaces*, 2017, pp. 69–72.

[9] B. Masiero and M. Vorländer, "A Framework for the Calculation of Dynamic Crosstalk Cancellation Filters," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 1345–1354, 2014.

[10] A. Shapiro, "Building a Character Animation System." Springer, 2011, pp. 98–109.

[11] A. Bönsch, B. Weyers, J. Wendt, S. Freitag, and T. W. Kuhlen, "Collision Avoidance in the Presence of a Virtual Agent in Small-Scale Virtual Environments," *IEEE Symp. 3D User Interfaces*, pp. 145–148, 2016.