Audiovisual Coherence:
Is Embodiment of Background Noise Sources a Necessity?

Jonathan Ehret[1]* Andrea Bönsch[1]* Isabel S. Schiller[2] Carolin Breuer[3]
Lukas Aspöck[3] Janina Fels[3] Sabine J. Schlittmeier[2] Torsten W. Kuhlen[1]

[1] Visual Computing Institute, RWTH Aachen University, Germany
[2] Work and Engineering Psychology, RWTH Aachen University, Germany
[3] Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany

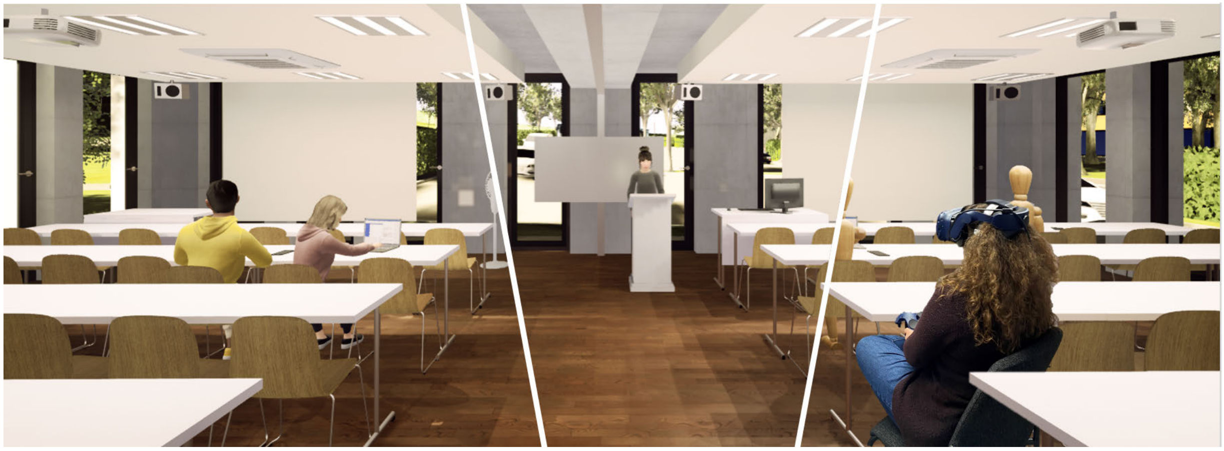

Figure 1: Three levels of background sound source visualization fidelity. Left: Animated virtual characters and, e.g., a moving fan (condition `Animated`); Center: No Visualization of background sound sources (condition `None`); Right: Peers visualized as non-moving wooden mannequins and other sources as static objects (condition `Static`). The female speaker in the center is identically visualized at all levels. A participant wearing a head-mounted display is embedded to show the seating position of participants during the study.

Abstract

Exploring the synergy between visual and acoustic cues in virtual reality (VR) is crucial for elevating user engagement and perceived (social) presence. We present a study exploring the necessity and design impact of background sound source visualizations to guide the design of future soundscapes. To this end, we immersed $n = 27$ participants using a head-mounted display (HMD) within a virtual seminar room with six virtual peers and a virtual female professor. Participants engaged in a dual-task paradigm involving simultaneously listening to the professor and performing a secondary vibrotactile task, followed by recalling the heard speech content. We compared three types of background sound source visualizations in a within-subject design: no visualization, static visualization, and animated visualization. Participants' subjective ratings indicate the importance of animated background sound source visualization for an optimal coherent audiovisual representation, particularly when embedding peer-emitted sounds. However, despite this subjective preference, audiovisual coherence did not affect participants' performance in the dual-task paradigm measuring their listening effort.

Index Terms: General and reference—Empirical studies; Human-centered computing—Virtual reality; Human-centered computing—User studies; Computing methodologies—Perception;

1 Introduction

In virtual reality (VR) applications, strategically designing visual and acoustic features plays a crucial role in enhancing (social) presence and perceived realism [12]. Consequently, such design elements also contribute to improved user engagement [17], encompassing factors like the listening experience and cognitive performance. This strategic design can be implemented through various means. For instance, optimizing visual signals such as using higher-quality renderings [11] or allowing user interactions within the immersive virtual environment (IVE) [29] have demonstrated efficiency. Moreover, Kim et al. [13] found that the visual embodiment of a virtual agent (VA) as the user's interaction partner significantly enhances the perceived social presence compared to audio-only interactions. Additionally, integrating animated behavior indicating social cues like gestures and facial expressions during user-commanded actions enhances user engagement more effectively than less interpretable VA behavior where users may not readily discern the VA's actions or intentions [13]. In the acoustic domain, integrating stimuli coherent with the virtual scene and actions taking place, e.g., tailored soundscapes or footstep sounds [12], contribute to a more immersive experience [10]. Furthermore, spatially rendered VA speech using binaural audio significantly enhances social presence when compared to non-spatial audio formats (mono and stereo) [4].

When visual and acoustic signals closely align semantically, despite minor temporal or spatial differences, they synergize into an integrated audiovisual signal [14, 28]. This phenomenon raises the question of how different visual representations for the same sound influence audiovisual integration and, more importantly, affect the perceived (social) presence and, thus, user engagement. We address

*These authors contributed equally to this work.
e-mail: {ehret|boensch}@vr.rwth-aachen.de

this question specifically concerning background sounds, an integral part of tailored soundscapes, intended to enhance IVE vibrancy without disturbing users or depleting their cognitive resources.

We examined whether there is a requirement to visually depict background sound sources in VR, emitted either from VAs populating the IVE or non-human scene elements, and aimed to determine the required **audiovisual coherence**, particularly in synchronizing the audio and visual elements, to strike a delicate balance: enhancing (social) presence while mitigating disturbances arising from the representation style of the background noise sources, thereby ensuring the user's optimal performance in the cognitive task at hand, such as attentive listening and efficient processing of speech content. Furthermore, we aimed to explore whether there are differences in the subjective perception of background sounds emanating from VAs compared to non-human sources. To this end, we compared three distinct **visualization fidelities** in terms of the accuracy of the visual elements (see Fig. 1), based on Kim et al.' approach [13], in a within-subject study: (i) without visualizing background sound sources (`None`), (ii) non-animated placeholders without illustrating what is causing the sound (`Static`), and (iii) animated visuals showing the precise origin of the sound (`Animated`).

To prevent participants from focusing directly on the specific audiovisual signals, we utilized a dual-task paradigm [9, 22, 23]. This paradigm was carefully designed to evaluate participants' ability to simultaneously maintain their performance on a primary (listening) task — attentive listening to a VA's speech content — while engaging in a secondary (vibrotactile) task within the IVE [23], before recalling the memorized information to answer questions about its content. Besides cognitively challenging the user, this dual-task paradigm was instrumental in objectively assessing (i) participants‘ memory performance in the listening task and (ii) participants' accuracy and response times in the secondary task as behavioral indicators for listening effort (LE), in the following referred to as behavioral LE. In an explorative fashion, we investigated whether there is a potential correlation between visual fidelity and participants' behavioral LE. We aimed to determine if behavioral LE could serve as a viable objective metric for assessing the optimal audiovisual coherence of an IVE. Complementing this, we collected subjective measures, including user ratings on perceived LE and (social) presence, to gain a nuanced understanding of participants' experiences and the impact of audiovisual coherence on their (social) presence and engagement in an IVE.

The remainder of this paper comprises details of our user study (Sect. 2), results (Sect. 3), discussion of findings (Sect. 4), and concludes with a summary (Sect. 5).

2 METHOD

In this chapter, we provide a brief overview of our IVE comprising different background sounds, and the speech material used before delving into the specifics of the study's design and procedure.

2.1 Virtual Environment

As the visual setting, a seminar room [16] was chosen and participants consistently occupied a specific desk in the third row (see Fig. 1). Six *MetaHuman*[1] models were seated in the room to simulate fellow students, creating a more realistic scenario. The peers were strategically positioned, with some directly in the participants' field of view and two peers in the back. An embodied conversational agent (ECA) [3], representing a female university professor, stood in front of the class at a lectern and was also visualized using a MetaHuman, animated with an idle animation and the recorded facial movement when speaking. To fit the speech sound featuring a read-out style, we implemented a gazing schedule such that the ECA looked down towards the lectern at the beginning of each sentence and then alternated her gaze between the virtual peers and the participant, following the gaze dynamics described in [20].

For the IVE's soundscape, we incorporated three classes of **background noise sources**, employed via binaural acoustics as suggested in [4]: (i) Human sounds produced by the peers (i.e., coughing, whispered conversation, laptop typing, or yawning) and non-human sounds originating from sources (ii) within the seminar room (i.e., a fan, window blinds closing, or phones ringing) and (iii) outside (i.e., a car passing by the window or a dog barking). Examples of some background noise source representations can be seen in Fig. 2. While animations for objects (e.g., the fan rotating and turning left to right) were simple to implement, animations for the sounds produced by the virtual peers were not readily available. They had to be coordinated with the sounds, which were acoustically recorded under controlled conditions. To this end, we used a simple motion capture setup comparable to the one described in [7], where off-the-shelf VR hardware with additional foot trackers is used in combination with inverse kinematics to record full-body animations. These animations were then manually post-processed to eliminate tracking errors. Additionally, we recorded several seated idle animations to introduce diversity in the movements among the virtual peers. While the fan in front of the seminar room was continuously operating throughout the entire study, we ensured a balanced distribution of both the quantity and class of the remaining background sounds between different runs in the study. To this end, we manually created 22 schedules for the execution of the sounds while the ECA was speaking (Sect. 2.2), a number derived from the study procedure outlined in Sect. 2.5.

2.2 Study Task

For the primary task, we asked participants to listen to family stories narrated by the ECA standing in front of the class (see Fig. 1). We utilized texts from the established Heard Text Recall (HTR) paradigm [24], consisting of 34 German texts providing information on three generations of family members, with nine questions per text[2]. We recorded 20 (derived from the study procedure detailed in Sect. 2.5) of these stories in a hemi-anechoic chamber at the Institute for Hearing Technology and Acoustics. A female voice expert (a speech-language pathologist and voice researcher) read the texts, each lasting between 53 and 62 s. In addition to the voice recordings, we captured facial movements, using an iPhone XR and ARKit, to later animate the virtual speaker's face (see, e.g., [5]). Following the text presentation, participants sequentially answered nine questions displayed on projection screens located to the left and right of the ECA (see Fig. 1). These questions pertained to family relationships, hobbies, etc., either directly provided in the text or inferred from various pieces of information. Participants answered these questions verbally and the correctness of their responses was logged by the experimenter.

In order to quantify participants' LE, we employed a dual-task paradigm that comprised the HTR as the primary listening task and a vibrotactile secondary task [18, 23]. Both were conducted alone (`Single-Task` baseline) and in parallel (`Dual-Task` condition). Specifically, while listening to the ECA's speech, participants reacted to vibration patterns presented via two handheld controllers by clicking a button on either the right or left controller. Based on the cognitive load theory [19], a decrease in task performance (more errors or increased response time) was taken as an indicator of higher listening effort in the respective listening condition.

2.3 Study Design

We conducted a within-subject study evaluating the influence of audiovisual coherence of background sound sources on perceived (social) presence and user engagement. While the sound itself was kept identical, we varied the visual fidelity across three levels: `None`,

[1] https://www.unrealengine.com/metahuman

[2] Recordings of the texts and respective face trackings are available in [8].

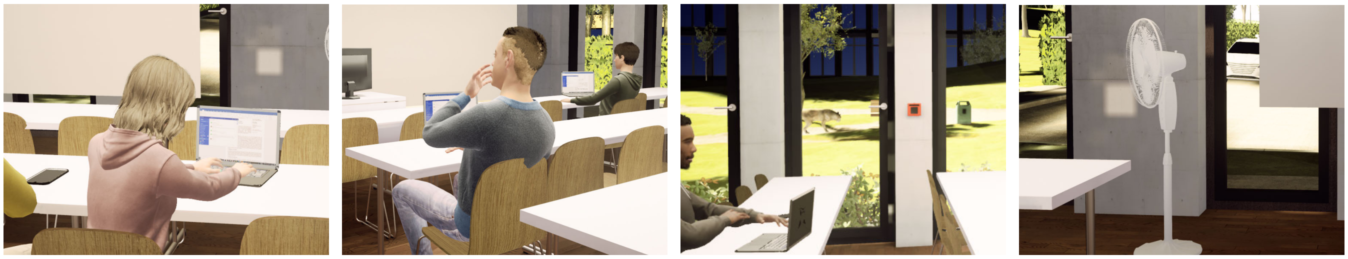

Figure 2: Examples of high fidelity background sound representations. From left to right: A virtual peer typing and a vibrating mobile phone on the table; a peer coughing; a barking dog crossing by outside the window; a fan in the front turning left to right with a spinning rotor.

`Static`, and `Animated`. In the first level, no representation for the origin of the sounds was shown (see fan and peer missing in the center of Fig. 1). In the `Static` condition, objects were placed as placeholders at every source of a sound but they, for example, did not move in the case of the fan or the car outside. Furthermore, we replaced the virtual peers with static, non-animated wooden mannequins to avoid eeriness effects of static highly-detailed VAs. The `Animated` condition featured representations of background sounds in high fidelity, as described in the previous section.

We expected the following hypotheses to be confirmed:

H1 Animated background sound sources are preferred over static visualizations which are preferred over no visualizations.

H2 (Social) Presence positively correlates with higher fidelity.

These two hypotheses are motivated, e.g., by the results of Kim et al. [13]. They compared users' perceptions of three types of virtual interaction partners. These partners were a disembodied voice, a VA with embodied gestures, and a VA with both embodied gestures and locomotion. They found that visual embodiment and plausible social behavior, encompassing gestures and locomotion and thus a fully animated VA, can significantly enhance users' perception of VAs in terms of social presence, comfort, and engagement, creating a more natural and intuitive interaction experience. Furthermore, also the difference between background noise being produced by other (virtual) humans in comparison to other, non-human sources should be explored. Additionally, we aimed to exploratorily assess, whether behavioral LE is correlated with the fidelity level of background sound source visualizations and thus the audiovisual coherence. We carefully suggested a potential negative correlation, implying that high visual fidelity (`Animated`) with dynamic motions might induce attention capture, potentially disturbing users, and diverting them from their primary cognitive task at hand.

2.4 Apparatus

The experiment was implemented using the *Unreal Engine* (version 4.27) and the *StudyFramework* plugin [6], the latter facilitates setting up and conducting factorial-design studies as ours. Participants wore an HTC Vive Pro Eye while being seated in a sound-proof hearing test booth (A:BOX, Desone Modulare Akustik, Berlin, Germany) with the dimensions 2.3 m × 2.3 m × 1.98 m ($w \times d \times h$) and a room volume of approximately 10.5 m^3. The audio was played over Sennheiser HD 650 headphones and the binaural dynamic live-rendering using an artificial head HRTF in a 1x1° resolution [25] was done using *Virtual Acoustics*[3] including RAVEN [26] for room simulation. All background sounds being made by humans used a human singer directivity filter and the sound of outside sound sources was combined with a transmission filter for the windows and played at an appropriate window.

2.5 Study Procedure

Upon written informed consent and eligibility check, participants were allowed to take part in the study. They were seated in the sound-proof booth at a table, position-wise exactly matching the virtual desk in the seminar room. They were equipped with headphones (Sennheiser HD 650) and a head-mounted display (HTC Vive Pro Eye) with two controllers. First, participants completed a practice block of the vibrotactile task (no HTR text, 1 sound schedule), followed by a single vibrotactile baseline block (no HTR text, 1 sound schedule). Next, we presented two HTR texts [24] to practice the primary (listening) task (2 HTR texts, 2 sound schedules). All of the above were conducted in the `None` condition. This was followed by the baseline block of the listening task, containing three texts, one for each visual fidelity level in counterbalanced order (3 HTR texts, 3 sound schedules). Afterward, there were also three texts for practicing the dual-task paradigm, counterbalanced in all three conditions (3 HTR texts, 3 sound schedules), followed by a short break. After that, three experimental blocks followed. Each block contained four repetitions of a text being presented with parallel vibrotactile tasks and questions being asked, all using the same visualization fidelity (3×4 HTR texts, 3×4 sound schedules). The order of these blocks was counterbalanced, and the assignment of texts and background sound schedules were randomized between participants. This procedure resulted in 22 trials in total (6 for practice, 4 for single-task baseline, 12 for dual task), requiring 20 HTR texts and 22 sound schedules.

After each dual-task experimental block, participants were asked to fill out an intermediate questionnaire, rating their perceived presence using the igroup presence questionnaire [27]. Social presence of the ECA was rated using the anthropomorphism construct of the Godspeed questionnaire [2] accompanied by the question "The speaker appeared to be sentient (conscious and alive) to me" (German: "Die Sprecherin wirkte auf mich wie ein fühlendes Wesen (mit Bewusstsein, lebendig)"). This last question is one of five items of the Social Presence Survey (SPS) [1] and was used in isolation as in [5] to enhance measuring the perceived anthropomorphism with a further social dimension. Social presence was only evaluated for the speaker in front, as the virtual peers were not visually present at all visual fidelity levels. Additionally, six questions assessing participant's subjective listening impression were asked, ranging from "How strong was your listening effort?" (the only item referring directly to perceived LE) over 'To what extent did you feel disturbed or bothered by background noise?" to "How in need of recovery do you feel right now?", based on [21]. These were accompanied by four questions asking whether participants felt in company apart from the speaker, and how plausible and real the background sound and the speech of the lecturer were perceived, e.g., "To what extent did the background noises resemble a real environment?". All of these were rated on a 5-point Likert scale from "not at all" (German: "gar nicht") to "extremely" (German: "außerordentlich"). After finishing all three blocks, a final questionnaire was posed, asking for demographics and a ranking of the visualization fidelity levels. In

[3] https://www.virtualacoustics.org/

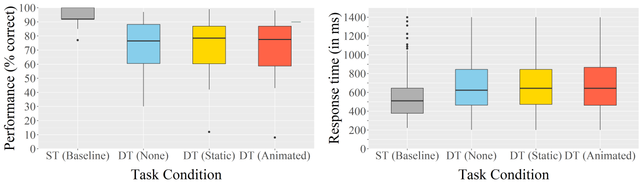

Figure 3: Secondary (vibrotactile) task results for the performance outcomes in % correct (left) and response time in ms (right) as a function of visual fidelity and task condition (`Single-Tasking` (ST) vs. `Dual-Tasking` (DT)).

this questionnaire, participants were also asked to recall all background sounds they remembered and on the next page to rank the actual sounds (given) by their disturbance. Furthermore, they had to rate the disturbance of three aspects (i.e., missing or static visual representations, and non-continuous background sounds during answering questions). In total, the experiment lasted for around 90 minutes, of which 50 to 60 minutes were spent immersed. The study was approved by the ethics committee of the Faculty of Arts and Humanities (ref. 2022_016_FB7_RWTH Aachen) and the experimental protocol was carried out in accordance with the Declaration of Helsinki.

3 Results

The analysis was performed using R (version 4.3.2).

3.1 Participants

Thirty-six persons participated in our study. We excluded three due to technical reasons (e.g., tracking problems), three due to self-reported restricted (and non-corrected) hearing or vision, and two due to failing the audiometry screening ($\leq$ 25 dB HL according to pure-tone audiometry between 125 and 8000 Hz using an Auritec ear3.0 audiometer). Furthermore, no subjective data was stored for one participant, so he/she was excluded as well. The remaining 27 persons (14 male, 12 female, 1 diverse) reported a mean age of 23.4 years ($SD = 3.8$). Five of the participants (18.5%) reported having never used virtual reality (VR) before, eight (29.6%) only once before, 12 (44.5%) less than 10 times, and the rest (7.4%) more frequently. One participant had to be further excluded from the objective evaluation (behavioral LE) due to errors by the experimenter when logging data.

3.2 Behavioral Result

To assess whether participants' behavioral LE was affected by the level of visual fidelity, we analyzed secondary (vibrotactile) task performance and response times, as well as the percentage of correctly answered questions of the primary task. Data was modeled using generalized linear mixed-effects models (GLMMs). Regarding secondary task performance, the final GLMM included the fixed effect Condition (`Single-Task Baseline`, `Dual-Task (None)`, `Dual-Task (Static)`, `and Dual-Task (Animated)`) and random intercepts for Participant, Trial, and Vibration Pattern. This model was specified with a binomial distribution and logit link function, considering that the outcome variable was binary (i.e., either correct or false). Regarding response time, the final GLMM again included the fixed effect Condition and, random intercepts for Participant, Trial, and Vibration Pattern. This model was specified with a Gamma distribution and log link function. Post-hoc comparisons were conducted using the Tukey Method, based on estimated marginal means calculated with the *emmeans* package [15].

Table 1 shows the descriptive results for the **primary (HTR) task** of answering the text-related questions. While there were no significant main effects of fidelity or task, there was a significant interaction effect of both ($\chi^2(2) = 10.56$, $p = .005$). However, post-hoc tests did not reveal significant pairwise differences with only single-tasking `Static` vs `Animated` coming close ($p = .075$) and all other p's $> .35$.

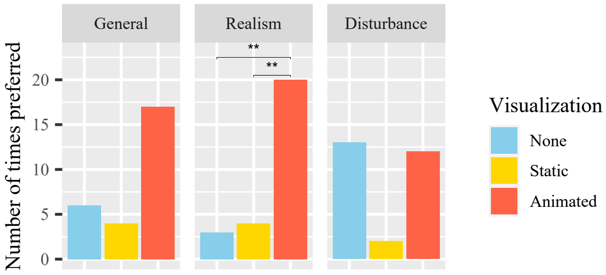

Figure 4: The number of times a visualization was picked as most preferred, in general, and with regard to subjectively perceived realism or disturbance. $**$ indicates $p < .01$.

Table 1: Primary (HTR) task results (percentage of correctly answered questions) as a function of fidelity and single- vs. dual-tasking

Fidelity	`Single-Tasking`	`Dual Tasking`
	Mean (SD)	*Mean (SD)*
`None`	59.96 (20.07)	50.04 (13.19)
`Static`	64.54 (24.53)	48.69 (16.06)
`Animated`	54.31 (22.51)	52.65 (16.19)

Regarding the **secondary (vibrotactile) task**, Fig. 3 depicts participants' performance (left) and response times (right) for the `Single-Task Baseline` condition, and the three visual fidelity levels when dual-tasking. Note that, in contrast to the primary task, the secondary task Baseline condition was not performed under each visual fidelity.

Statistically, secondary task performance varied significantly with the condition under which the task was performed (i.e., `Single-Task Baseline`, `Dual-Task (None)`, `Dual-Task (Static)`, and `Dual-Task (Animated)`) ($\chi^2(3) = 67.71$, $p < .001$). More precisely, participants' performance in the `Single-Task Baseline` condition was significantly better compared to their performance in any of the `Dual-Task` conditions ($p < .001$). However, the degree to which performance declined did not vary for visual fidelity, as revealed by pairwise comparisons conducted using Tukey's method for adjusting p-values (`None` vs. `Static`: z-ratio $= -0.17, p = 1.00$; `None` vs. `Animated`: z-ratio $= 0.96$, $p = .77$; `Static` vs. `Animated`: z-ratio $= 1.14$, $p = .67$).

Similar results were obtained for response time measures. Overall, response times also varied significantly across the conditions ($\chi^2(3) = 124.73$, $p < .001$). That is, participants responded fastest in the `Single-Task Baseline` condition but were significantly slower in each of the three `Dual-Task` conditions ($p < .001$). Again, however, the increase in response times when dual-tasking was unaffected by Fidelity Condition, as indicated by pairwise comparisons (`None` vs. `Static`: z-ratio $= -1.21$, $p = .62$; `None` vs. `Animated`: z-ratio $= -0.73$, $p = .88$; `Static` vs. `Animated`: z-ratio $= 0.46$, $p = .97$).

3.3 Subjective Evaluation

Following the subjective ratings in the questionnaires between the study blocks and at the end will be analyzed. If not stated differently, we performed 1-way repeated-measures ANOVAs and post-hoc

Bonferroni-corrected t-tests for statistic analysis. If the data violated the normality assumption (validated via Shapiro-Wilk's tests), Friedman tests were conducted with potential Bonferroni-corrected Wilcoxon signed-rank tests as post-hoc tests.

Regarding reported presence using the **igroup presence questionnaire** [27], there were no significant differences between the visualization conditions for all subscales (*sense of being there*: $p = .93$, $M = 4.17$, $SD = 1.51$; *Spatial Presence*: $p = .17$, $M = 4.03$, $SD = 1.18$; *Involvement*: $p = .17$, $M = 3.86$, $SD = 1.27$; *Experienced Realism*: $p = .48$, $M = 2.50$, $SD = 1.06$). The same is true for the **Godspeed's** Anthropomorphism scale [2] ($p = .57$, $M = 2.77$, $SD = 0.86$). Analyzing the answers to the single **social presence** question from the SPS [1] referring to the speaker only, a Friedman test revealed a significant effect of visualization ($\chi^2(27) = 7.58$, $p = .02$). Post-hoc tests showed a significant effect ($p = .01$) only between `Static` ($M = 2.41$, $SD = 1.12$) and `Animated` ($M = 2.89$, $SD = 1.25$) visualizations, with `None` scoring in between ($M = 2.56$, $SD = 1.16$).

After finishing the three study blocks, participants were asked for **preference** with regard to disturbance, realism, and in general, shown in Fig. 4. For each condition, a rating of 1 (preferred) to 3 (least preferred) was gathered. A Friedman test of preferences with regard to disturbance showed no significant difference ($p = .15$). However, for realism a significant effect was found ($\chi^2(27) = 18.1$, $p < .001$) and post-hoc tests showed that `Animated` ($M = 1.33$, $SD = 0.62$) was significantly preferred over `Static` ($M = 2.30$, $SD = 0.72$, $p = .001$) and `None` ($M = 2.30$, $SD = 0.72$, $p = .003$), while the latter two were not significantly different. For general preference, a similar trend emerged ($\chi^2(27) = 9.1$, $p = .018$), where `Animated` ($M = 1.56$, $SD = 0.80$) was preferred over `Static` ($M = 2.22$, $SD = 0.70$), approaching statistical significance ($p = .052$), and over `None` ($M = 2.22$, $SD = 0.80$, $p = .088$), although these differences did not reach conventional levels of significance.

In the post-study questionnaire, participants were also asked to rate how **disturbing** they experienced the non-continuous noise (background sounds were only scheduled during the presentation of the stories, not during questions), the static representations, or the missing representation. The ratings regarding the noise had a mean of $M = 2.59$ ($SD = 1.05$). Comparing the ratings for static visualizations ($M = 2.56$, $SD = 1.22$) and missing visualizations ($M = 1.96$, $SD = 1.06$) using a Wilcoxon signed-rank test revealed a significant difference ($z = 29$, $p = .04$) judging the latter as subjectively less disturbing. Again, all were rated on the identical 5-point Likert scale.

Before revealing in the next question which **background sounds** we included, we asked participants to state which sounds they remember, allowing multiple answers. Twenty-five (93%) remembered conversations between the peers, 22 (81%) mentioned mobile phones, 11 (41%) also stated coughing, and eight (30%) outside noises, or some more specifically cars passing or a dog barking. Furthermore, 7 participants (26%) mentioned a constant background noise or referred more specifically to the fan, while 7 (26%) remembered typing sounds and 2 (7%) specifically referred to yawning. Additionally, sounds that were not part of the simulation were mentioned: moving papers (4 participants, 15%), drinking water (3 participants, 11%), and sounds of chairs (2 participants, 7%). On the next page of the questionnaire, we then asked participants to rank the background sounds that they experienced (explicitly given here) by their annoyance. The results of this ranking can be seen in Fig. 5, with mean rankings being: Phone Ringing (2.2), Conversation (2.7), Phone Vibrating (3.3), Coughing (5.2), Laptop Typing (5.2), Throat Clearing (5.9), Yawning (7.0), Dog Barking (7.8), Car Passing (8.4), Fan (9.0), Window Blends (9.7), and Other (11.6).

Analyzing the ten questions asked on 5-point Likert scales regarding participants' **listening impression**, including perceived LE, and **realism of the sounds and scene**, only one significant effect of the visualization can be found, namely for "Was your mental performance negatively affected by the background noise?" ($\chi^2(27) = 10.7$, $p = .004$). Post-hoc tests showed that participants subjectively felt that their mental performance was significantly more disturbed when `Animated` representations were present ($M = 3.37$, $SD = 1.12$) compared to when no visualization (`None`) were shown ($M = 2.85$, $SD = 1.15$) with $z = 25.5$, $p = .014$. A non-significant trend ($p = .07$) was found for a second question ("Did you feel in the room, aside from the speaker, in the company of others?") with a post-hoc test showing a non-significant trend ($p = .075$) between `None` ($M = 2.37$, $SD = 1.18$) and `Animated` ($M = 3.04$, $SD = 1.16$) with `Static` scoring in between ($M = 2.52$, $SD = 0.96$). For all other questions, no significant effects of visual fidelity were revealed (all $p > .18$).

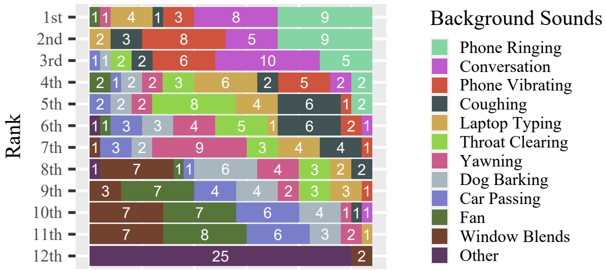

Figure 5: Ranking of the annoyance of the background sounds from most annoying (1st) to least (12th). Numbers indicate the number of occurrences of a raking.

4 Discussion

In our study, participants' subjective preferences across three visual fidelity levels of background source visualization (and thereby varying audiovisual coherence) were investigated while also taking the participants' performance in the dual-task paradigm into account.

When asked for their general preference, participants clearly preferred background sources being visualized in high fidelity (`Animated`), albeit only significantly with regard to realism. Additionally, the general preference also shows a clear trend towards the `Animated` level. Surprisingly, no clear preference emerged for `Static` or, in our case, partially abstract representations (wooden mannequins) over not visualizing the background noise sources at all (`None`). In fact, participants rated having static representations as more disturbing compared to their absence. These outcomes led us to only partially accept **H1** (expected preference P: P(`Animated`) $>$ P(`Static`) $>$ P(`None`)) for high-fidelity visualizations. Consequently, when embedding background sound sources, a vivid representation would be the most favorable choice. However, before embedding only placeholders (`Static`), it might be advisable to refrain from introducing any virtual representation (`None`). This is further supported by participants' responses when asked to choose the least disturbing condition: votes were equally distributed between `None` and `Animated`, while `Static` was only preferred by a much smaller fraction (see Fig. 4).

While the overall background soundscape varied between the different runs, the individual background sound sources were kept identical and only their visual representations were manipulated, altering the audiovisual coherence. Surprisingly, the different visual fidelities did not affect the perceived presence, contradicting our initial expectations. Yet, interestingly, the ECA, presented consistently in all conditions, was perceived as more sentient when surrounded by virtual peers resembling its appearance (`Animated`) rather than abstract peer representations (`Static`). The Godspeed-Anthropomorphism scale, containing a similar item to be ranked on a bipolar unconscious-conscious scale, did not reveal a similar outcome. Nonetheless, the observed difference in perceived social presence is very interesting, given that only the environment was manipulated, not the virtual speaker itself, leading us to partially accept **H2** (higher fidelity correlates with higher (social) presence).

Upon examining the most recalled background sounds, there is a clear tendency towards those generated by virtual peers. The same is true when looking at the participants' ranking of the background sounds in terms of perceived disturbance. Consequently, we hypothesize that human-made sounds induce a higher level of disturbance compared to those emitted by scene objects (e.g., blinds closing) or even animals within the scene (e.g., a dog barking). However, since we did not explicitly vary those across conditions, further research in this avenue is required.

Our analysis did not reveal a significant effect of visual fidelity on behavioral indicators of LE and thereby potentially user engagement. Although participants performed significantly weaker and gave slower responses in the secondary task during dual-task conditions, compared to the single-task conditions, this discrepancy was unaffected by the level of visual fidelity. Consequently, our results suggest that participants' LE during a listening task in VR appears to be independent of how accurately the prevailing background sounds are visually represented. This observation is particularly interesting as participants indicated that their subjectively perceived listening impression was negatively influenced by the animated peers (`Animated`), albeit als not significantly for the single perceived LE question. Although not entirely congruent, these ratings partly support our prior assumption that high visual fidelity might divert them from their cognitive task at hand. However, one potential explanation could be occasional glitches or imperfections observed in the animations of the high-fidelity peers (e.g., the back of a peer penetrating the back of the chair shown in Fig. 2, 2nd from left). We invested considerable time refining VA movements, yet occasional glitches arose due to inherent limitations in the motion capture method. Importantly, we deliberately chose a diverse array of movements over a limited set of highly refined animations. This decision prioritized a close-to-real-life simulation, emphasizing realistic animation that closely mirrors peers' behaviors. Nevertheless, this poses a **limitation** of the presented study. Another potential shortcoming was revealed by the fact that several participants reported remembering sounds of people drinking or chairs moving, which we did not include in our general soundscapes. Although one participant mentioned in the open-ended comments "The sounds from the workspace of the experimenter were transmitted quite loudly.", suggesting that the talk-back microphone, used for set-up communication with the participant inside the sound-proof booth, was inadvertently left active, these recollections can also be intrusions (false memories which are not uncommon in eyewitness testimony). Repeating the experiment more carefully avoiding inadvertently acoustic noise and examining the impact of audiovisual coherence on potential false memories stands as an intriguing avenue for future research. A third potential limitation of our study is the choice of wooden mannequins as peer representations in `Static`. Despite our intention to mitigate behavioral realism discrepancies, we introduced a visual incongruity between `Static` and `Animated`, particularly as the wooden mannequins' realism contrasted with the overall realistic IVE. This likely resulted in lower social presence ratings towards the ECA in the front in `Static`, suggesting an impact on participants' perceptions. This emphasized the need to carefully consider visual congruity in future studies for an unbiased participant experience.

We deliberately chose not to assess the social presence of the virtual peers in the initial intermediate questionnaires to avoid biasing participants towards them. However, including these assessments in **future work** might substantially deepen our understanding. Consonant with this, we plan to explore the possibility of employing a more interactive scenario to foster higher social presence. Furthermore, it would be interesting for further design of background sounds to differentiate more between the sound generated by VAs and those originating from the environment, for example, by introducing this as an additional variable.

5 Conclusion

We presented a within-subject study that evaluated whether and how background sound sources need to be visualized for optimal audiovisual coherence. Our results suggest that while participants preferred sound sources being visualized with high fidelity, only showing abstract placeholders is favored less than not showing them at all, at least for our scenario where wooden mannequins were chosen as abstract representations of the VAs. However, perceived presence was not influenced. Our explorative attempt to link behavioral LE, as measured through a dual-task paradigm, with visual fidelity preferences did not yield conclusive findings. Nonetheless, our study underscores the importance of coherent audiovisual representation in IVEs, particularly when incorporating human-made sounds.

Acknowledgments

This research was funded by the Federal Ministry of Education and Research (BMBF) and the Ministry of Culture and Science of the German State of North Rhine-Westphalia (MKW) under the Excellence Strategy of the Federal Government and the Länder. Projekt-ID G:(DE-82)EXS-SF-OPSF712. Isabel Schiller's and Sabine Schlittmeier's contribution to this study was also funded by a grant from the HEAD-Genuit-Foundation (P-16/10-W). We thank the student assistants Lalita Angsuphanich, Stine Düwel, Tom Jungbauer, Melchior Käppel, Malte Kögel, Mai Ly Tenberg, Sebastian Menne, Melf Sopha, and Judith Ziob for their support in implementing and conducting the study.

References

[1] 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*, 10:583–598, 2001. doi: 10.1162/105474601753272844

[2] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int J Soc Robot*, 1:71–81, 2009. doi: 10.1007/s12369-008-0001-3

[3] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents*. MIT Press, 2000.

[4] C. Dicke, V. Aaltonen, A. Rämö, and M. Vilermo. Talk to me: The Influence of Audio Quality on the Perception of Social Presence. In *Proc BCS HCI*, pp. 309–318, 2010.

[5] J. Ehret, A. Bönsch, L. Aspöck, C. T. Röhr, S. Baumann, M. Grice, J. Fels, and T. W. Kuhlen. Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Applied Perception (TAP)*, 18(4):1–15, 2021. doi: 10.1145/3486580

[6] J. Ehret, A. Bönsch, J. Fels, S. J. Schlittmeier, and T. W. Kuhlen. StudyFramework: Comfortably Setting up and Conducting Factorial-Design Studies Using the Unreal Engine. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW): Workshop "Open Access Tools and Libraries for Virtual Reality"*, 2024.

[7] J. Ehret, A. Bönsch, P. Nossol, C. A. Ermert, C. Mohanathasan, S. J. Schlittmeier, J. Fels, and T. W. Kuhlen. Who's next? Integrating Non-Verbal Turn-Taking Cues for Embodied Conversational Agents. In *ACM International Conference on Intelligent Virtual Agents (IVA '23)*, 2023. doi: 10.1145/3570945.3607312

[8] C. A. Ermert, C. Mohanathasan, J. Ehret, S. J. Schlittmeier, T. W. Kuhlen, and J. Fels. AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm, 2022. doi: 10.18154/RWTH-2023-05543

[9] J.-P. Gagné, J. Besser, and U. Lemke. Behavioral Assessment of Listening Effort Using a Dual-Task Paradigm: A Review. *Trends in Hearing*, 21, 2017. doi: 10.1177/2331216516687287

[10] C. Hendrix and W. Barfield. Presence in Virtual Environments as a Function of Visual and Auditory Cues. In *Proceedings Virtual Reality Annual International Symposium'95*, pp. 74–82, 1995.

[11] J. Hvass, O. Larsen, K. Vendelbo, N. Nilsson, R. Nordahl, and S. Serafin. Visual Realism and Presence in a Virtual Reality Game. In *3DTV Conference: The True Vision - Capture, Transmission, and Display of 3D Video*, pp. 1–4, 2017. doi: 10.1109/3DTV.2017.8280421

[12] A. C. Kern and W. Ellermeier. Audio in VR: Effects of a Soundscape and Movement-Triggered Step Sounds on Presence. *Front. Robot. AI*, 7:20, 2 2020. doi: 10.3389/frobt.2020.00020

[13] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *Proceedings of 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 105–114. IEEE, 10 2018. doi: 10.1109/ISMAR.2018.00039

[14] P. Laurienti, R. Kraft, J. Maldjian, J. Burdette, and M. Wallace. Semantic Congruence is a Critical Factor in Multisensory Behavioral Performance. *Experimental Brain Research*, 158(4), 2004. doi: 10.1007/s00221-004-1913-2

[15] R. V. Lenth. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.5. https://CRAN.R-project.org/package=emmeans. Accessed: January 5, 2024.

[16] J. Llorca-Bofí and M. Vorländer. IHTAclassroom. Multi-detailed 3D architecture model for sound perception research in Virtual Reality, 2021. doi: 10.5281/zenodo.4629716

[17] F. Mantovani, G. Castelnuovo, A. Gaggioli, and G. Riva. Virtual Reality Training for Health-Care Professionals. *CyberPsychology & Behavior*, 6(4):389–395, 2003. doi: 10.1089/109493103322278772

[18] C. Mohanathasan, J. Ehret, C. A. Ermert, J. Fels, T. W. Kuhlen, and S. J. Schlittmeier. Measuring Listening Effort in Adverse Listening Conditions: Testing Two Dual Task Paradigms for Upcoming Audiovisual Virtual Reality Experiments. In *22. Conference of the European Society for Cognitive Psychology , Lille , France , ESCoP*, 2022.

[19] F. Paas and P. Ayres. Cognitive Load Theory: A Broader View on the Role of Memory in Learning and Education. *Educational Psychology Review*, 26:191–195, 3 2014. doi: 10.1007/S10648-014-9263-5/METRICS

[20] T. Pejsa, S. Andrist, M. Gleicher, and B. Mutlu. Gaze and attention management for embodied conversational agents. *ACM Trans. Interact. Intell. Syst*, 5:3:1–34, 2015. doi: 10.1145/2724731

[21] I. S. Schiller, L. Aspöck, C. Breuer, J. Ehret, and A. Bönsch. Hoarseness among university professors and how it can influence students' listening impression: an audio-visual immersive VR study. In *Proceedings of the 1st AUDICTIVE Conference*, pp. 134–137, 2023. doi: 10.18154/RWTH-2023-08885

[22] I. S. Schiller, L. Aspöck, and S. J. Schlittmeier. The Impact of a Speaker's Voice Quality on Auditory Perception and Cognition: a Behavioral and Subjective Approach. *Frontiers in Psychology*, 14, 2023. doi: 10.3389/fpsyg.2023.1243249

[23] I. S. Schiller, A. Bönsch, J. Ehret, C. Breuer, and L. Aspöck. Does a Talker's Voice Quality Affect University Students' Listening Effort in a Virtual Seminar Room? In *Proceedings of the 10th Convention of the European Acoustics Association, Forum Acusticum 2023: Acoustics for a Green World*, 2023.

[24] S. J. Schlittmeier, C. Mohanathasan, I. S. Schiller, and A. Liebl. Measuring text comprehension and memory: A comprehensive database for Heard Text Recall (HTR) and Read Text Recall (RTR) paradigms, with optional note-taking and graphical displays, 2023. doi: 10.18154/RWTH-2023-05285

[25] A. Schmitz. A New Digital Artificial Head Measuring System. *Acustica*, 81:416, 1995.

[26] D. Schröder and M. Vorländer. RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments. In *Forum Acusticum*, pp. 1541–1546. Aalborg Denmark, 2011.

[27] T. Schubert, F. Friedmann, and H. Regenbrecht. The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments*, 10:266–281, 6 2001. doi: 10.1162/105474601300343603

[28] C. Spence. Audiovisual Multisensory Integration. *Acoustical Science and Technology*, 28(2):61–70, 2007.

[29] B. G. Witmer and M. J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, 1998. doi: 10.1162/105474698565686