Beyond Looks: A Study on Agent Movement and Audiovisual Spatial Coherence in Augmented Reality

Stephanie Arévalo Arboleda[*] Technische Universität Ilmenau
Christian Kunert[†] Technische Universität Ilmenau
Jakob Hartbrich[‡] Technische Universität Ilmenau
Christian Schneiderwind[§] Technische Universität Ilmenau
Chenyao Diao[¶] Technische Universität Ilmenau
Christoph Gerhardt[‖] Technische Universität Ilmenau
Tatiana Surdu[**] Technische Universität Ilmenau
Florian Weidner[††] Lancaster University
Wolfgang Broll[‡‡] Technische Universität Ilmenau
Stephan Werner[§§] Technische Universität Ilmenau
Alexander Raake[¶¶] Technische Universität Ilmenau

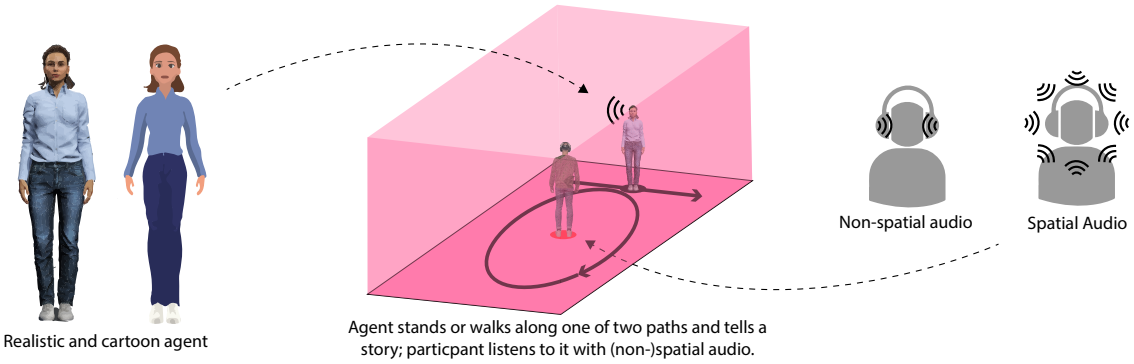

Figure 1: Overview of our study. We were interested in the influence of audiovisual coherence and agent movement on participants' experience. To investigate this, a participant standing in a room saw an augmented reality agent (via Microsoft HoloLens 2). The agent was represented either by a realistic or a cartoon virtual human. While telling a story, the agent either remained standing, walked side-to-side through the room, or circled the participant. The participant wore headphones and heard the story either in head-tracked spatial or non-spatial audio.

Abstract

The appearance of virtual humans (avatars and agents) has been widely explored in immersive environments. However, virtual humans' movements and associated sounds in real-world interactions, particularly in Augmented Reality (AR), are yet to be explored. In this paper, we investigate the influence of three distinct movement patterns (circle, side-to-side, and standing), two rendering styles (realistic and cartoon), and two types of audio (spatial audio and non-spatial audio) on emotional responses, social presence, appearance and behavior plausibility, audiovisual coherence, and auditory plausibility. To enable that, we conducted a study (N=36) where participants observed an agent reciting a short fictional story. Our results indicate an effect of the rendering style and the type of movement on the subjective perception of the agents behaving in an AR environment. Participants reported higher levels of excitement when they observed the realistic agent moving in a circle compared to the cartoon agent or the other two movement patterns. Moreover, we found an influence of agent's movement pattern on social presence and higher appearance and behavior plausibility for the realistic rendering style. Regarding audiovisual spatial coherence, we found an influence of rendering style and type of audio only for the cartoon agent. Additionally, the spatial audio was perceived as more plausible than non-spatial audio. Our findings suggest that aligning realistic rendering styles with realistic auditory experiences may not be necessary for 1-1 listening experiences with moving sources. However, movement patterns of agents influence excitement and social presence in passive unidirectional communication scenarios.

Keywords: virtual humans, audiovisual spatial coherence, agents, augmented reality, movement patterns

[*]e-mail: stephanie.arevalo@tu-ilmenau.de
[†]e-mail: christian.kunert@tu-ilmenau.de
[‡]e-mail: jakob.hartbrich@tu-ilmenau.de
[§]e-mail: christian.schneiderwind@tu-ilmenau.de
[¶]e-mail: chenyao.diao@tu-ilmenau.de
[‖]e-mail: christoph.gerhardt@tu-ilmenau.de
[**]e-mail: tatiana.surdu@tu-ilmenau.de
[††]e-mail: f.weidner@lancaster.ac.uk
[‡‡]e-mail: wolfgang.broll@tu-ilmenau.de
[§§]e-mail: stephan.werner@tu-ilmenau.de
[¶¶]e-mail: alexander.raake@tu-ilmenau.de

1 Introduction

In social Augmented and Virtual Reality (AR/VR), users often interact with virtual humans. These can be divided into avatars (controlled by humans) and agents (controlled by a computer) [2]. Many applications use agents as instructors in vocational and educational applications, tour guides in tourist experiences, or simply as someone guiding through a narrative experience. While the primary purpose of instructors and teachers is to provide facts and instructions, in entertainment applications, agents often want to elicit a strong emotional response in the user. Both of them want to elicit a high feeling of social presence [45].

How these agents are displayed, sound, and behave are crucial design decisions. For example, rendering style is an important factor, realistic textures performing better than those with abstract or cartoon styles [64], e.g., regarding social presence [70]. Similarly, it has been shown that the type of audio rendering (non-spatial vs. spatial) impacts various factors such as emotional response, with spatial audio often outperforming (non-spatial audio) [5,63]. Research has also shown users perceive an agent differently depending on its movement. Masuko et al. [38] as well as Ye et al. [65] showed that moving agents in AR perform better as storytellers and trainers/educators. This is not only because agents that can interact within their environment are seen as more capable [23,28], but also because an AR agent might benefit from walking similar to a real human during public speaking. An agent walking from side to side (similar to what a speaker would do on a stage) or around and among the audience (as a speaker in a more intimate environment would do) could potentially increase emotional engagement and deliver a more captivating talk compared to an agent standing still (as shown by recommendations for public speaking, e.g., Heinicke et al. [19])

While these individual factors (rendering style, type of audio rendering, movement) have been researched isolated, the interdependence among all three remains underexplored. However, exploring the interplay is crucial as it might significantly affect the user's perception of the agent. Fleming et al. [16] showed that audiovisual spatial coherence of voices and faces improves listening performance, and Bailenson et al. [4] showed that animation needs to match visual fidelity for optimized communication. Still, it remains open if effects, such as the previously mentioned positive effects of movement, depend on spatial audio and realistic rendering styles or if a moving agent can potentially enhance a non-spatial audio experience with a cartoon storyteller.

In this work, we investigate exactly that, i.e., the interplay between movement, rendering style, and audio rendering during a narrative experience delivered by an agent in AR. We focus on social presence and emotional response (as a measure of performance for the individual factor combinations) and plausibility and coherence (to investigate users's perception of matching/mismatching fidelities of the experience). To do this, we present the results of a study that analyzes three distinct movement patterns (circle, similar to an agent walking among the audience; side-to-side, similar to an agent walking on a stage; and standing) and two agent rendering styles (cartoon and realistic) on individuals' perceptions with regard to two audio instances (non-spatial and spatial).

The core findings of our study are:

- We confirm previous results that a realistic visualization is generally better than a cartoon visualization regarding perceived excitement, calmness, and appearance and behavior plausibility.
- While preferred, the difference between spatial audio and non-spatial audio was only marginal, suggesting that in pure 1-1-listening scenarios (even with moving sound sources), non-spatial audio is sufficient.
- Our study found that neither audiovisual spatial coherence nor alignment of visual and audio realism improved 1-1-listening experiences with moving sound sources.
- We show that the movement pattern of an agent impacts users' excitement, experienced social presence, and perceived audiovisual coherence.

2 Related Work

2.1 Social Presence

Social presence was initially defined as being aware of the presence of another being's (human, non-human, or artificial) intentions, intelligence, and emotions [8,9]. However, a more recent definition by Skarbez et al. [57] mentioned that coherence and plausibility are part of the *Social Presence Illusion*. This new term refers to "the feeling of social presence engendered by characters in virtual or mediated environments" with three characteristics: the company of another sentient being, the medium's ability for a multisensory communication experience, and appropriate copresence illusion and communicative coherence.

Social presence can be affected by various factors. In terms of the visual aspect of virtual humans, Yoon et al. [67] evaluated the effect of avatar appearance (cartoon and realistic) and body visibility on social presence with findings that point to similar levels of social presence for both rendering styles and an effect of body visibility favoring the avatar's full body. Conversely, Zibrek et al. [70] found that realistic rendering styles of agents improve social presence. In terms of auditory experiences, Dicke et al. [13] compared the effect of monophonic, stereophonic, and spatial human speech recordings on, among others, presence. They found that the spatial condition evoked a higher sense of presence than the monophonic condition. Similarly, Skalski & Whitbred showed that surround sound increases social presence [56]. Still, these studies have focused on audio or visual stimuli separately. Our goal is to build upon these findings and consider the whole audiovisual experience and investigate the interplay between audio and visuals.

2.2 Spatial Audio in AR/VR

Investigating the use of spatial audio in AR/VR environments, a reproduction method that aims to provide a listener with a three-dimensional impression of the virtual auditory scene [7], has been mostly carried out for urban soundscapes [24,35]. Focusing on agents, Tsepapadakis and Gavalas [61] used spatial audio for cultural heritage employing an AI-based agent in a storytelling scenario without virtualized visual content. Here, participants reported positive results in terms of immersion. Geronazzo et al. [17] presented an interactive dynamic VR storytelling platform where they compared non-spatial vs spatial sound with varying degrees of interactivity. For interactive spatial scenes, strong emotions and levels of immersion were recorded as opposed to the non-spatial representation. Interestingly, the static spatial condition received similar ratings to the non-spatial conditions. Immhor et al. [22] investigated the effect of spatial audio by comparing spatial, non-spatial, and face-to-face audio conditions on social presence in a collaborative task with two active speakers. Their results did not point to significant differences regarding social presence for the two types of audio used. Together, these contradicting (spatial audio sometimes better, sometimes not) and counterintuitive (spatial audio not always better) findings point to the need to further explore the effect of spatial audio on emotions and social presence in AR/VR environments.

2.3 Virtual Humans' Audiovisual Spatial and Behavioral Coherence

Showing virtual humans in AR environments is challenging, as they should act and move in a coherent and believable manner that adheres to the affordances of the current physical space. Hayes [18] mentioned that a virtual human's fidelity determines its authenticity in terms of three aspects: physical — how they look, sound, feel, and interact in the environment [6], functional — how they react to the environment [34], and psychological — how they engage and display emotions [54]. Beyond that, Kim et al. [29] mentioned the spatial coherence of virtual humans, i.e., how adjusted it is to the physical environment, is an important factor in fostering social plausibility and presence. Following up on this, Latoschick & Weinrich [33] proposed a coherence and plausibility model for AR/VR. They categorized the plausibility illusion as a subjective feeling while coherence is related to the "objective characteristics of the virtual experience". They defined coherence as "the relations between the cues and the AR/VR experience itself", integrating plausibility and coherence in one construct. One of the most recent

tools to evaluate coherence in virtual humans is the *virtual human plausibility questionnaire* with two dimensions: the *virtual human's Appearance and Behavior Plausibility (ABP)* and the *virtual humans' coherence with the virtual environment* [36]. The aforementioned aspects have been mostly approached from the visual perception side, leaving aside the acoustic representation of the virtual human in the environment as well as the interplay between audio and visuals.

Acoustic and visual stimuli coherence in virtual characters/agents is gaining attention, especially concerning matching appearance with speech. For instance, Higgins et al. [20] showed that unnatural voices can appear unappealing in combination with photorealistic agents but do not negatively affect social presence compared to natural voices. Similarly, Zibrek et al. [68] found that a mismatch between appearance and speech, e.g., photorealistic agents with unnatural voices, did not influence social presence or increased discomfort. Also, Kao et al. [26] investigated the importance of audial avatar customization. Their results suggest that visual customization is more important than audial customization, but the latter can still increase aspects like identification and immersion if visual customization is available . Recent research by Lam et al. [32] investigated the importance of audio-visual coherence regarding virtual characters. They presented male and female agents with varying body types and voice pitches to participants. The results show that participants have certain expectations about body types and corresponding voices and perceived appropriate combinations as more believable.

Another well-known effect in this context is the ventriloquist effect, which has been exploited intensively in arts and media for a long time, with a puppeteer giving the illusion that the puppet speaks and not the human, implying an illusion of spatial coherence of acoustic and visual sources (cf., e.g., Alais & Burr [1]). Here, in most cases, auditory information is "captured" by visual information, associating the sources of the two modalities to the location of the visual information. As shown by Alais & Burr [1], degraded visual information may also lead to auditory information capturing visual information. Similarly, motion seems to interact with the capturing effect, and visual motion may lead to perceived auditory motion [58]. Related, if the auditory event is perceived from outside the current field of view, attention may be guided by auditory information [31], which in turn, may have an effect on scene exploration behavior. For complementary information on investigations of audiovisual effects and binding in AR/VR see, e.g., [31, 37, 55]. Previous research investigating this effect was conducted with rather simplistic signals such as sound bursts, light sources, or visual patterns

As outlined, research has, up to now, focused either on visual coherence or matching speech and appearance in virtual humans, leaving aside audiovisual spatial coherence, i.e., matching (room acoustics) with visuals and movement. Thus, our goal is to build upon those findings to explore if there is a similar effect of matching room acoustics with the sounds produced by a moving agent regarding social presence, audiovisual coherence, and plausibility.

2.4 Virtual Humans ' Motion Perception , and Emotional Responses

Research in virtual humans indicates that high fidelity/highly realistic avatars often lead to positive experiences [64]. However, it may not be the most important aspect of social interactions. Here, Von der Putten et al. [48] showed that visual fidelity might not be as important as behavior fidelity (aka how the virtual human moves). For instance, humans express and perceive emotions not only through verbal behavior but also through body sways and kinematic patterns are key aspects of emotional responses [30]. Recently, a few studies have highlighted the importance of non-verbal behavior in interaction, focusing on movements of the upper body and facial expressions [12, 50]. Here, Rogers et al. [50] suggested that harnessing full face and body motion capture can make social interaction in VR similar to face-to-face interaction.

Movement in virtual humans and its impact on the user's emotions have not been researched holistically, considering all three, visuals, audio, and movement. Movement-focused research showed that the agents' gait visualizations (walking styles) and gestures increases social presence and the overall perceived friendliness of agents [49] and that virtual agents' motion and appearance have increase emotional responses [39]. Research focusing on visual rendering showed that eerie rendering styles of agents (creepy, scary, zombie) evoke higher avoidance movement behavior (faster walking speed, greater distance, longer paths) and higher emotional reactivity compared to other rendering styles (cartoon, realistic) [40, 44]. Also, Narang et al. [41] found that self-recognition of walking improved with avatars than with point-lights and that recognizing others walking was easier with circular motions than with straight-line walking. Additionally, exploration of auditory events coming from a moving virtual human has received little to no attention so far, in spite of non-AR/VR studies demonstrating their relevance for human decision-making derived from sound sources, e.g., Pastore et al. [47]) and for perception of body weight [59]. Previous findings highlighted the impact of sound on emotional appraisal in games and films [14] and also increasing levels of arousal for virtual reality experiences [15]. Closely related to our research, Thomas et al. [60] investigated how speech-related motion (head, hands, posture) and speech realism influence the perceived personality of an agent (and not, as we do, the user's emotional reaction). Their findings show that motion is the dominant modality indicating extraversion and speech communicates agreeableness and emotional stability. These findings suggest that how sound is presented together with motion or movement impact how agents are perceived.

Together, research shows that virtual humans', their motion, and the sounds they produce, potentially influences plausibility perception and emotional responses. Such emotional responses can be evaluated using the bi-dimensional model of affect considering valence (pleasure, positive -displeasure, negative continuum) and arousal (level of activation) [52].

Overall, the effect of different types of auditory representations of natural acoustic events, e.g., the sounds of humans walking and talking together with the visual representation of a moving virtual humans, on the overall audiovisual experience and the user's emotional response, remains unexplored.

3 Experiment

(a) (b)

Figure 2: (a) The participant wearing the Hololens 2 and headphones. (b) The HMD view during the experiment .

We conducted a 2x2x3 within-subjects experiment (12 conditions) to evaluate the effect of audio (spatial audio, non-spatial audio), the agent's rendering style (cartoon, realistic), and type of movement (standing, walking side-to-side, walking in a circle) on the subjective

audiovisual experience. The conditions were counterbalanced using a Latin square design.

In order to explore the use of spatial audio for the current application, a non-spatial condition was defined as well. During the non-spatial condition, the diotic presentation of the test samples was used, with no added binaural, source, or room information. Inspired by previous studies [22], we used the non-spatial as our control condition.

3.1 Hypotheses

We capitalize on previous findings about the advantages of realistic rendering styles and consider that combining realistic audiovisual representations, i.e., realistic rendering style with realistic auditory representation, will impact emotional responses, social presence, appearance and behavior plausibility, and audio plausibility. For a realistic auditory representation, we used two sound sources (the agent's voice together with footstep sounds). We represented these sounds (when the agent was standing or moving) to match the specific room where the experiment occurred. Through this, we aimed to provide an acoustic experience similar to the one in the real world.

Our first three hypotheses (H1, H2, H3) focus on investigating the role of agent movement (see Sect. 2.4). Given that a realistic rendering style is generally perceived positively [64], and movement [60] as well as spatial audio [17] have been shown to increase emotional responses, we assume that these effects stack and a moving realistic agent with spatial audio is perceived as more exciting (H1). Vice versa, we hypothesize that a standing agent will be perceived as calmer, independent of the rendering style and type of audio. That is because spatial audio has little effect on a static object, and rendering style a is less impactful due to the absence of movement [30] (H2). Inspired by the findings of coherence in speech with the visual appearance of agents, e.g., [20, 32, 68], together with findings related to movement pattern recognition [41], we hypothesize that realistic audiovisual representations together with movement will lead to the highest audiovisual coherence (H3). The last two hypotheses (H4, H5) aim to investigate the audiovisual experience. H4 hypothesizes that social presence increases with a realistic rendering style with spatial audio compared to a cartoon style with non-spatial audio. We base this hypothesis on individual findings of realistic rendering [70] and spatial audio [13] improving social presence (see Sect. 2.1). Similarly, in H5, we considered that realistic audio and realistic rendering will outperform less realistic conditions (a cartoonish rendering style and non-spatial audio) in terms of the perceived agents' appearance and behavior plausibility [33, 36, 36, 55] (see Sect. 2.3). Thus, our hypotheses are:

- H1 Participants will report higher levels of excitement when the agent is moving (side-to-side, circle) during the audiovisual realistic conditions (realistic agent and spatial audio) compared to the non-realistic audiovisual conditions (cartoon agent and non-spatial audio).
- H2 Participants will report higher levels of calmness when the agent is standing compared to the agent moving (circle, side-to-side) for both audiovisual representations
- H3 The moving agents (circle and side-to-side animations), together with realistic audiovisual representations (realistic rendering style and spatial audio) will lead to the highest scores of audiovisual coherence.
- H4 Participants will report higher scores in social presence for audiovisual realistic representations (realistic rendering style and spatial audio) compared to the non-realistic audiovisual condition (cartoon agent and non-spatial audio).
- H5 Participants will report higher scores in ABP for the realistic audiovisual representations compared to the non-realistic audiovisual conditions.

3.2 Apparatus

3.2.1 System Overview

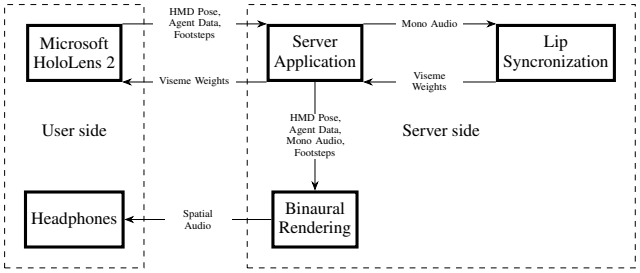

Figure 3: Diagram of the system architecture. The users wear a HoloLens 2 as well as headphones. Lip syncronization and the audio rendering are offloaded to an edge server. Due to latency concerns, the headphones use a wired connection to the server.

The system enables the visualization of animated agents using a Microsoft HoloLens 2 while providing non-spatial and spatial audio for the agent's speech and movement.

As the HoloLens 2 has rather limited capabilities regarding performance and audio quality, we decided to offload some processing to an additional desktop PC and use the Sennheiser HD600 headphones for audio playback. The setup is illustrated in Fig. 3. The central server application runs on a PC and is connected to the HoloLens 2 through Wi-Fi. The additional modules (*BinSim* [42] and *Oculus LipSync* [1]) are connected through localhost sockets. To ensure synchronization between audio and graphics, the latency of the connection was measured and the audio was delayed accordingly. All individual network connections are realized using *ZeroMQ* [2]. We use *Google Protocol Buffers* [3] for data encoding.

The HoloLens 2 itself runs an application that was built in Unity 2021.2.7f1. When starting the application, participants see the environment mesh to confirm that it is fully loaded. The experiment is initiated when the researcher starts the server application. The rendering style, the animation, and the user's height are specified on the PC, and a message is sent to the HoloLens 2 to start the visualization. The HoloLens 2 stops showing the environment mesh and spawns the specified agent configuration in front of the participant in relation to their body height (so that does not appear to float or penetrate the floor). Further information about the agents is given in Sect. 3.2.3. For the duration of the experiment, the HoloLens 2 streams data to the server. This includes a continuous stream of its pose (position and orientation) and the position, rotation, and scale of the agent. Additionally, we added triggers to the agent's feet in order to detect footsteps. A message is sent to the server each time a footstep is detected.

The server records the incoming data from the HoloLens 2 application. It also exchanges data with additional modules that implement spatial audio and lip movement. When the experiment is started, the server loads the audio file and sends it continuously to the LipSync module. We rely on *Oculus LipSync* for generating lip movement. While there is a Unity plugin for lip synchronization, it does not work on the HoloLens 2 because it does not support x86 ARM platforms. Therefore, we implemented the module in C++ and sent the viseme weights to the HoloLens 2. Visemes depict the animation of the mouth shape for specific sets of phonemes. We used the 15 visemes specified in the *MPEG-4 Face and Body Animation*

[1] https://developer.oculus.com/downloads/package/oculus-lipsync-unity/
[2] https://zeromq.org/
[3] https://protobuf.dev/

(FBA) standard [46]. Sending the visemes separately eases the synchronization as the audio also runs on the server. The server sends information about the position and orientation of the participant and sound sources (speech, footsteps), which is further described in Sect. 3.2.2. The final audio is played using wired headphones, connected to the server.

3.2.2 Audio Rendering

The present experiment included two audio reproduction conditions: non-spatial and spatial audio. To blend the virtual sound source into a real environment, binaural cues and the room acoustic characteristics of the natural environment must be processed. One method to mimic the real room acoustics is based on measured binaural room impulse response (BRIR) data. This method was used for the auralizations in the present experiment. Other approaches for creating virtual acoustic environments include simulation techniques, e.g., based on geometrical and wave-based acoustics [11, 53].

It is important to note that *The Microsoft spatializer*[4], which can be coupled with the HoloLens 2 over Unity, utilizes only Head-Related Transfer Functions (HRTFs). HRTFs describe the directional filtering of sound reaching a listener's ears, but do not include room information or sound source-specific directivity. While reverberation effects can be adjusted, this does not guarantee an accurate match to the real environment. Thus, we used a modified version of the *pyBinSim* rendering framework [43].

The *pyBinSim* framework uses fast partitioned convolution to convolve measured or simulated binaural room impulse responses with arbitrary audio signals for binauralization. Furthermore, the filters can be switched in real-time to allow for six 6-Degree-of-Freedom (6DoF) auralizations. The version in this experiment uses separate binaural filters for the direct sound, early reflection, and late reverb. This allows switching the filters for these three segments independently in the case of positional updates. Furthermore, an additional modification in the processing allows us to dynamically adapt the sound source directivity behavior by modifying the direct sound portion of the BRIRs.

The Binaural Room Impulse Respone (BRIR) dataset in this experiment was created from BRIR measurements using the Head-And-Torso Simulator (HATS) KEMAR 45BA [5]. It was placed at a distance of 2 m in front of a sound source in the middle of the reproduction room on an electric turn table. The horizontal plane was sampled in 4° steps, resulting in 90 BRIRs. The BRIRs were measured using the exponential sine sweep method in the frequency range of 50 to 22000 Hz. In post-processing, the direct sound portion of the BRIRs was removed, and the late reverberation segment was separated at a mixing time estimate of about 60 ms. For the dynamic auralization, the KEMAR HRTF dataset was used to substitute the removed direct sound segment [10]. This HRTF data set contains a full-spherical measurement with an angular resolution of 1° for both azimuth and elevation. For our modified data set, this HRTF set was downsampled to 3° and 5° for the azimuth and elevation dimensions, respectively, to reduce the amount of data. Additionally, an angle-dependent sound source directivity was applied by convolving the direct sound with a measured sound source directivity response (male speaker) provided by the McRoomSim Toolbox [62]. In total, three virtual sound sources were considered, one located at the position of the head for auralizing the speech sound, and two at the position of the feet for auralizing the footsteps.

To achieve a 6DoF reproduction, several parametric adjustments and simplifications were added to the rendering framework. Distance changes between the sound source and the receiver were considered by scaling the direct sound according to the distance between the listener and the sound source.

[4] https://github.com/microsoft/spatialaudio-unity/

[5] https://www.grasacoustics.com/products/head-torso-simulators-kemar/product/733-45ba

(a) (b)

Figure 4: Appearance of the cartoon (a) and realistic (b) agent used in the experiment.

The early reflection parts were updated according to the relative azimuthal angle between the listener's head and the virtually reproduced sound source. The late reverberation segment was constant for positional changes and, hence, static for all positions. As Neidhardt et al. [43] showed these kinds of simplifications introduced to a BRIR data set still provided plausible illusions in an auditory augmented reality scenario. However, the acoustic scenarios of the present work require situations with a strong direct sound. From several pre-listening test sessions, it was concluded that these BRIR data set modifications could be applied to the listening scenarios in the present study.

Two types of sound were used for the auralizations in the experiment — female speech and footstep sounds. Both sound types had to be dry recordings without any perceivable room acoustic information incorporated. The female speech sample was recorded in a dry speaker booth. The Neumann-U87 microphone, with a cardioid pattern, and a popkiller were employed for the measurement. The footstep samples were obtained from an open source database [6]. To adjust the overall loudness of the auralization and the relative loudness differences between speech and footstep sounds, calibration measurements were conducted using the KEMAR 45BA dummy head in the test room. Six people were recorded standing still and talking to the dummy head at a distance of 2 m. Additionally, they had to walk past the dummy head at a distance of 2 m to record footstep sounds. The loudness estimates of the recordings were compared to the loudness estimates of the binaural playback to adjust the reproduction levels of the signals. From the measurements, it was concluded to set the target speech at about 65 dBA at a distance of 2 m and the footsteps sound 20 dB lower. However, as the spatial audio rendering does not account for accurate dynamic reflection changes, the loudness of the virtual footstep sound may perceivably deviate from the true loudness for positions much closer and farther away than 2 m.

3.2.3 Agent Appearance and Behavior

Creating virtual humans with distinct styles (cartoon and realistic) was an important aspect of our study. The agents were created with the help of the Character Creator 4 Software Suite [7]. The software was used for designing, texturing, and rigging the agents. The final version of the agents used within the experiment is shown in Fig. 4. The height of the agents (approximately 1.90 m) was not modified to the viewer's height. The height was chosen so that most viewers could always see the agent's face without having to uncomfortably look up or down.

[6] https://www.fesliyanstudios.com/royalty-free-sound-effects-download/footsteps-31

[7] https://www.reallusion.com/character-creator/

To customize the cartoon-style agents, we used custom simplified textures to manipulate facial features and clothing. This allowed us to create exaggerated and stylized features per our chosen style.

To achieve a high degree of realism for the realistic agent, we fine-tuned facial contours, skin textures, and subtle imperfections to create a lifelike appearance.

Considering the hardware limitations of the Microsoft HoloLens 2 used in our experiment, we had to make specific optimizations to the agents: The HoloLens 2 has constraints on computational resources, particularly regarding rendering capabilities and polygon count. Consequently, we needed to reduce the number of vertices of our agents to 80,000 to maintain visual fidelity. The optimization process involved simplifying geometry, using level-of-detail (LOD) techniques, and employing efficient shaders to ensure smooth performance on the HoloLens 2.

For the movement of the agents, we relied on recordings captured via motion tracking using an OptiTrack system that included a total of 10 cameras. For the size of the tracking space that was available to us (approx. 3x3 meters), this led to some inaccuracies, especially for the hands. Besides making sure that the recorded person was always within view of multiple cameras, we also used a smaller marker set that excluded hand tracking. We then exported the three animations (circle, standing, side-to-side) for use in the Unity engine. The minimum and maximum distances between the viewer's (i.e., the virtual camera's) position and the agent (i.e., the agent's head, which acts as the sound source) depend on the movement pattern. The minimum and maximum distances for the standing position are 1.95 m and 2.06 m, with 1.99 m on average (SD = 0.021). For the circular motion, the range is between 0.72 m and 2.63 m with an average of 2.02 m (SD = 0.55), while the side-to-side movement ranges from 1.95 m to 3.03 m with an average of 2.41 (SD = 0.30).

In Unity, we adjusted the resulting animations as the automatic retargeting showed some mismatches when combining the animations from the OptiTrack motion-capturing system with the agents. First, we manually adjusted the hand pose (straight fingers) as the original pose looked unnatural (hand fixed in a claw-like position). Second, we adjusted the spine rotation and forced a stricter T-pose, as Unity's built-in T-pose checker has a relatively high tolerance. Last, we enabled Inverse Kinematics (IK) for the agent's feet. This was particularly important as we rely on accurate foot movement for step detection in order to enable footstep sounds.

3.3 Task & Procedure

A combination of free-viewing and story-listening tasks was used in this study. During the experiment, participants were asked to observe the agent when it walked around/stood and listen to it tell a short fictional story as shown in Fig. 2. The chosen test stimulus is a dry recorded female speech sample reciting Frank Kafka's piece 'Give it up' [25] (128 words, 52 seconds). We used the same story across all conditions to avoid confounding effects of using different stories, i.e., participants could have preferred one story over another, influencing, in turn, the reported excitement or calmness.

The experiment lasted around 50 minutes and included the following steps. (1) Participants were given a short introduction about the goal of the experiment and were handed a consent form (5 minutes). (2) Then, they were shown the devices to be used, were instructed how to wear the HoloLens 2, and proceeded to calibrate it (5 minutes). (3) After that, participants were instructed to pay attention to what they would hear and see. Participants stood at a specific position during each experimental condition (12 conditions, 52 seconds per condition). After each condition, participants were instructed to fill out a set of questionnaires about their experience. Their movement space was restricted within a box of 0.4 m x 0.4 m, marked with blue lines on the laboratory floor. (4) Finally, they filled out a questionnaire about their preferences (rendering style and type of audio) and reasonings behind them and received their compensation (5 minutes).

An experimenter was present during the session, sitting outside the path of the agent and in a darker area of the room to not disturb or influence the experience.

3.4 Measures

In our study, the subjective perception of the agent was assessed through a questionnaire. The questionnaire consisted of 17 items and four subsets: *appearance and behavior plausibility (ABP)* with 6 questions, *social presence* with 5 questions, *emotional responses* 2 questions, *audiovisual spatial coherence and audio plausibility* 2 questions.

To measure ABP, we used the ABP dimension from the Virtual Human Plausibility Questionnaire (VHPQ) [36], using a 7-point Likert scale (from "does not apply at all" to "completely applies"). For social presence, we used the Social Presence Questionnaire from Bailenson et al. [3], which uses a 7-point Likert scale (from "not at all" to "very much"). To capture emotional responses, we asked participants their level of excitement and calmness, using the following questions with a 7-point Likert scale ("not at all" to "very much"): "The experience I just had made me feel excited."; "The experience I just had made me feel calm." We constructed these questions considering the dimensional model of emotions using valence (positive-negative) and arousal (high-low) [51]. Our questions considered only positive valence but represented two different levels of arousal/activation: excitement (high) and calmness (low).

The questions that we constructed to measure audiovisual spatial coherence were: "I felt that the observed position of the virtual character and the location of the sound source are matching," and for audio plausibility, we asked: "The sound coming from the virtual character (while speaking, standing, and walking) in this room seemed to be plausible to me." Both questions were measured with a 7-point Likert scale from 1 ("not at all") to 7 ("very much").

For the user preferences, participants were asked to choose their preferred type of audio and rendering style and provide reasonings behind their choices. The questions we used were, "Which virtual character did you prefer the most (cartoon, realistic) and why?" and "Which audio representation did you prefer (spatial audio, non-spatial audio) and why?".

3.5 Participants

We recruited 36 participants (21 identified as male, 14 as female, and one as non-binary) ranging in age from 22 to 32 years (M = 26.53, SD = 2.65) who reported no hearing problems, all university students or staff. Among the participants, three used contact lenses, 10 wore eyeglasses, and the other 23 did not use visual aids. A total of 24 participants had experienced AR/VR environments before our experiment, of which two had programming experience using Unity, and 12 had never experienced AR/VR before.

This study was pre-approved by the Ethical Committee of the university and executed following the guidelines of the national research organization and the declaration of Helsinki. Participants gave informed written consent and received 12 euros as compensation for their participation.

4 Results

All questionnaires were analyzed using Aligned Rank Transform (ART) [27] as normality was not confirmed in any of our collected measures. Normality was assessed by visual inspection using QQ-plots and Shapiro-Wilk tests. If significant differences were detected, we calculated post-hoc pairwise comparisons using ART-C and Bonferroni correction. Table 1 shows a summary of the results of all collected measures. In the following paragraphs, we mention our results, emphasizing the ones that yielded significant differences.

Table 1: Summary of results showing significant effects in bold.

Metric	p-value						
	agent	audio	animation	agent:audio	agent:animation	audio:animation	agent:audio:animation
Excitement	**0.036**	0.076	**<0.001**	0.174	**0.017**	0.262	0.089
Calmness	**0.040**	0.098	0.168	**0.004**	0.055	0.504	0.151
Social Presence	0.499	0.612	**0.016**	0.455	0.717	0.413	0.625
Appearance Behavior Plausibility	**0.041**	0.872	**0.042**	0.258	0.681	0.881	0.383
Audiovisual Spatial Coherence	0.344	**0.022**	**0.002**	**0.031**	0.509	0.851	0.565
Audio Plausibility	0.284	**0.045**	0.062	0.082	0.269	0.428	0.697

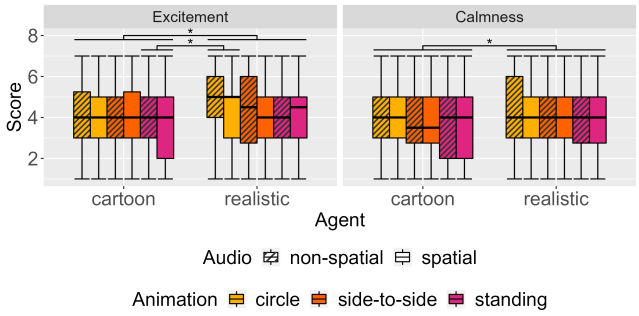

Figure 5: Excitement and calmness [1;7].

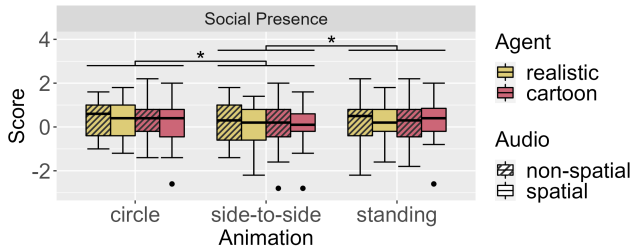

Figure 6: Social presence questionnaire results [-3;+3].

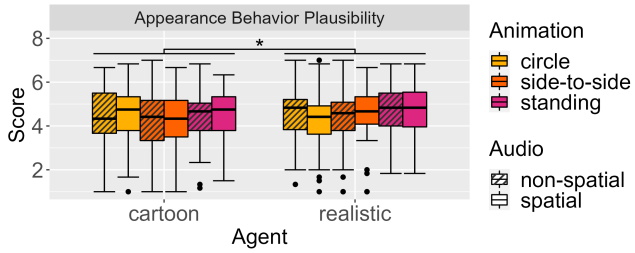

Figure 7: Appearance and behavioral plausibility questionnaire results [1;7].

4.1 Excitement and Calmness

Fig. 5 shows an overview of the results.

4.1.1 Excitement

We found a two-way interaction effect of agent and animation ($F(2,385) = 4.119$, $p = .017$, $\eta_p^2 = .021$). Post-hoc pairwise comparisons using ART-C highlight a significant difference in excitement for the realistic agent with the circle animation (M = 4.51, SD = 1.64) compared to the standing cartoon agent (M = 4.03, SD = 1.82). Results also point towards the main effects of the type of animation and type of agent. Animation had a significant effect ($F(2,385) = 14.148$, $p < .001$, $\eta_p^2 = .068$), where the circle (M = 4.39, SD = 1.66) animation led to higher scores in excitement compared to the side-to-side (M = 4.16, SD = 1.81), ($p < .001$) or standing (M = 4.05, SD = 1.79), ($p < .001$) animations. Results also point to a significant main effect of the type of agent ($F(1,385) = 4.414$, $p = .036$, $\eta_p^2 = .036$), where the realistic agent evoked higher scores of excitement (M = 4.27, SD = 1.75) than the cartoon agent (M = 4.13, SD = 1.76). We did not find a three-way interaction effect. It is essential to consider that the interaction effect might explain both main effects.

4.1.2 Calmness

Our results point to a two-way interaction effect between the type of agent and audio conditions ($F(1,385) = 8.363$, $p = .004$, $\eta_p^2 = .021$). However, the post-hoc comparison did not show any significant differences. Also, we found a main effect for the agent's appearance ($F(1,385) = 4.256$, $p = .04$, $\eta_p^2 = .011$), where the realistic agent (M = 3.96, SD = 1.69) led to higher levels of calmness compared to the cartoon agent (M = 3.87, SD = 1.69). We did not find a three-way interaction effect.

4.2 Social Presence

The social presence questionnaire results are shown in Fig. 6. Our results revealed a main effect only for the type of animation ($F(2,385) = 4.155$, $p = .016$, $\eta_p^2 = .21$), favoring the circle (M = 0.3, SD = 0.83) animation over the side-to-side animation (M = 0.11, SD = 0.921) , ($p = .036$), and for the standing (M = 0.28, SD = 0.9) over the side-to-side animation, ($p = .041$). The circle animation had no significant difference to the standing animation ($p = 1.0$). We did not find any two-way between or three-way interaction effects.

4.3 Appearance and Behavioral Plausibility

Our results only point to significant main effects of the type of animation and agents' appearance individually. The animation condition was found to have a significant effect ($F(2,385) = 3.199$, $p = .042$, $\eta_p^2 = .016$), but the post-hoc test did not show significant differences. The agent's appearance ($F(1,385) = 4.199$, $p = .041$, $\eta_p^2 = .011$) led to significant differences, where the realistic (M = 4.48, SD = 1.34) agent evoked higher appearance and behaviour plausibility scores compared to the cartoon agent (M = 4.31, SD = 1.33). We did not find any two or three-way interaction effects.

4.4 Audiovisual Spatial Coherence and Audio Plausibility

We found a significant two-way interaction of the agent's appearance and audio ($F(1,385) = 4.701$, $p = .031$, $\eta_p^2 = .012$). Post-hoc pairwise comparison revealed that the cartoon agent with spatial audio (M = 4.38, SD = 1.63), ($p = .04$) led to higher audio-visual coherence

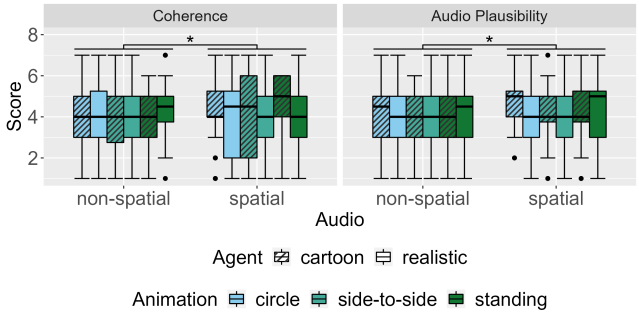

Figure 8: Audio questionnaire results [1;7].

.

compared to the cartoon agent with non-spatial audio (M = 3.93, SD = 1.69).

We also found significant main effects of audio ($F(1,385) = 5.303$, $p = .022$, $\eta_p^2 = .014$) and animation ($F(2,385) = 6.224$, $p = .002$, $\eta_p^2 = .031$). The spatial audio (M = 4.22, SD = 1.64) resulted in higher perceived audio-visual coherence than the non-spatial audio conditions (M = 4.01, SD = 1.66). This may also explain the interaction effect of audio. The standing (M = 4.34, SD = 1.45) condition led to significantly higher audio-visual coherence scores compared to the side-to-side (M = 3.92, SD = 1.72) animation. We did not find a three-way interaction effect.

Regarding audio plausibility, our results indicate a main effect of the type of audio ($F(1,385) = 4.027$, $p = .045$), where spatial audio (M = 4.25, SD = 1.55) was perceived acoustically more plausible than the non-spatial (M = 4.1, SD = 1.54) condition. No further significant main or interaction effects were found.

4.5 Preference

We asked participants their preference between the two rendering styles. Here, 25 participants preferred the realistic agent, ten the cartoon, and one none of the agents. We also asked participants about their preferred type of audio after the experience. Here, 25 participants reported their preference for spatial audio, nine for non-spatial audio, and two reported no preference.

5 Discussion

5.1 Movement on Excitement and Calmness (H1 and H2)

In H1, we hypothesized that participants would report more excitement (positive valence and high arousal) when the agent moves (side-to-side and circle animations) for realistic audiovisual conditions. Our results partially confirm this hypothesis. We found a significant effect of the type of animation and the agent's rendering style on the excitement that participants reported. This aligns with previous findings [39], that found that a virtual agent's motion and appearance have a positive effect on emotional responses. However, we did not find an effect on the type of audio used.

In terms of animation, participants reported feeling more excited with the circle animation compared to the side-to-side or standing animation. We consider that these findings are rooted in how humans perceive emotions, namely verbal and nonverbal behavior, especially kinematic patterns that display hints about current emotional states [30]. In a comparable study, Narang et al. [41] found that circular motions of avatars of familiar others in VR can be better identified than walking in a straight line. Here, we consider that the circular movement that the agent displayed could have allowed participants to focus not only on the walking styles but also on arm swings and head motions, evoking higher arousal.

Regarding rendering style, participants reported higher excitement with the realistic agent compared to the cartoon one. Our result aligns with previous findings that suggest that a realistic rendering style and motion have an effect on emotional responses [39]. Closely related, Zibrek et al. [69] found a correlation between realistic rendering style and empathy. These results emphasize the effect of motion together with realistic rendering styles as factors that could lead to positive experiences when interacting with virtual agents.

In H2, we hypothesized that participants would report higher calmness (positive valence low arousal) when the agent is standing compared to when the agent is moving (circle and side-to-side) for both audio conditions. Our results do not confirm this hypothesis. We did not find an effect of movement or audio on participants' reported levels of calmness. However, similarly to H1, we found an effect of rendering style on calmness, i.e., higher levels of calmness with the realistic rendering style. This emphasizes the effect of a realistic rendering style on positive emotions with positive valence.

Our findings do not only support the use of a realistic rendering style when using virtual humans but suggest that realistic rendering styles can foster positive AR experiences.

5.2 Audio and Rendering Style on Social Presence (H3)

In H3, we hypothesized that participants would report higher scores in social presence for the realistic audiovisual conditions compared to the non-realistic combinations. Our results do not confirm this hypothesis. We did not find a significant effect of the rendering style or the type of audio on social presence. Regarding the rendering style, this result is in line with a study by Yoon et al. [67], stating that rendering style does not have a significant effect on social presence. Zibrek et al. [70] however, found that realistic rendering styles can improve social presence.

With regard to audio features it was shown, that unnatural [20] or synthetic [68] voices combined with photorealistic agents do not negatively affect social presence compared to natural voices. Conversely, it was shown that realistic speech has a positive effect on perceived personality attributes when tested with a cartoon agent [60]. Regarding the effect of spatial audio on social presence, Immohr et al. [22] did not find significant differences when comparing the use of spatial audio with non-spatial audio for two-party communication in VR environments. Overall, we cannot confirm if/how audiovisual realistic combinations affect social presence.

Interestingly, we found a main effect of the type of animation on social presence where seeing the agent moving around participants evoked higher scores of social presence compared to the agent walking side-to-side and the standing agent led to higher social presence compared to it walking side-to-side. These findings align with Narang et al. [41] who identified that circular walking movements are easier to identify than straight-line walking in avatars of familiar people. Our findings further suggest that this may also be the case for unfamiliar virtual humans, in our case agents. Also, the ease of identifying circular movements could have influenced social presence, leading to higher scores. This indicates that the movement pattern virtual humans perform affects social presence. This highlights implications when using avatars or agents for entertainment or in communication settings where non-linear motions could create more engagement for instance when used in plays, or to connect with a conversational partner.

5.3 Audio and Rendering Style on ABP (H4)

We hypothesized that realistic audiovisual representations would lead to the highest ABP scores. Our results partially support this hypothesis. We found a main effect of the rendering style — our results showed higher scores for the realistic agent — on the ABP scores but no effect of audio. Also, most participants reported preferring the realistic agent over the cartoon. Here, similar to H3, our results point to a main effect of the type of animation.

To better understand our results, we looked at studies about the effect of sound in games. Huibert [21] mentioned that in virtual

environments, realistic sound acts as a confirmation of the sensory information derived from other senses. Closely related, Kao et al. [26] found that audial avatar customization has a weaker effect compared to visual customization. Further, the auditory information may have been overshadowed by the visual stimuli with both auditory conditions (spatial audio and non-spatial audio due to the vetriloquist effect. Also, this effect may have intensified with multiple exposures, our experimental design had 12 conditions.

Overall, our results align with those findings and suggest that a higher ABP for the realistic virtual humans creates the illusion of an overall realistic auditory environment in spite of the type of audio used. This suggests that in 1-1 listening AR/VR scenarios the effect of spatial audio may be negligible. In our experiment, the agent was standing or moving in a studio-like room reciting a fictional short story and was the only audio source. We consider that having multiple agents talking and different audio sources coming not only from them but from the environment, for example, similar to plays in theaters, could benefit from spatial audio and provide an immersive experience.

5.4 Audio, Movement, and Rendering Style on Audiovisual Spatial Coherence (H5)

We hypothesized that moving realistic audiovisual agents would lead to higher scores in audiovisual coherence compared to standing non-realistic audiovisual agents. Our results do not confirm this. While we did not find a three-way interaction confirming that assumption, we found a two-way interaction effect of rendering style and type of audio. Interestingly, this was found only for the cartoon rendering style, where the use of spatial audio was significantly more coherent than the non-spatial audio conditions. This suggests that matching realistic visual representations with realistic auditory representations may not be the key to audiovisual spatial coherence. We consider this closely related to the Proteus Effect [66], applied not only to self-avatars but to the perception of other virtual humans. Here, rendering style may not be the key to audiovisual spatial coherence but other aspects such as behavioral mapping fidelity [48].

Regarding audio plausibility, spatial audio was found more plausible than non-spatial audio, aligning with participants' preference for spatial audio. Prior research has reported that spatial audio improves listening performance in multi-talker environments [16]. Our findings suggest that this may also be the case even for an environment with one talker with auditory information about steps and voice. Although, it may also be a result of the general preference for spatial audio when compared to non-spatial audio conditions.

In terms of the influence of movement patterns on audiovisual plausibility, the standing condition was found more audiovisually plausible than the side-to-side movement. One possible explanation is the chosen scenario — an agent telling a story to participants. Participants could have found it more plausible that the agent would stand and tell a story rather than walk side-to-side in the room, as people do in a similar scenario.

6 Limitations

One of the limitations of our study comes in terms of the expressiveness of the agents (non-verbal cues). Since effective human communication relies on more than speech-transmitted information, it is essential to address the integration of non-verbal cues. While lip-syncing and mutual gaze were part of the current framework, other features, such as complex facial expressions, were not considered.

Another aspect is the type of spatial audio used. The Unity spatializer is a standard tool implemented for social mixed reality applications for achieving an immersive spatialized reproduction. However, we used the *pyBinSim* rendering approach to provide an accurate room acoustic experience. In a future study, we plan to evaluate the use of Unity spatial audio vs. *pyBinSim* and determine if Unity provides enough acoustic cues for a plausible and immersive scenario as the one currently investigated.

We constructed our own questions related to audiovisual spatial coherence and audio plausibility. To the best of our knowledge, no validated questionnaire measures audio plausibility or audiovisual spatial coherence. Our questions did not consider familiarity with the room where the experiment took place and many of our participants might not have been familiar with that specific room, influencing in turn their responses. Still, we consider that our questions represent the most important aspects of audiovisual coherence and audio plausibility.

7 Conclusion

Inspired by immersive storytelling experiences with talking agents, the core objective of this research was twofold: First, to investigate if an agent's locomotion behavior affects a user's emotional response and perceived audiovisual coherence. Second, to investigate if an audiovisual coherent experience (realistic rendering and spatial audio) maximizes social presence and perceived plausibility. In a user study with N = 36 participants, we analyzed multiple variables, including the self-reported levels of excitement and calmness, the perception of social presence, the agent's appearance and behavioral plausibility, and audiovisual coherence. Our results corroborate earlier findings indicating that realistic rendering styles are generally perceived as superior to cartoon-based renditions (here: for excitement, calmness, and appearance and behavior plausibility). Further, while spatial audio is preferred, its advantage over non-spatial audio is minimal. This suggests that for direct 1-1 listening situations, even those involving dynamic sound sources, non-spatial audio can aptly serve the purpose. Additionally, we observed that achieving audiovisual spatial coherence or aligning the realism between visual and audio does not necessarily improve a 1-1 listening experience with mobile sound sources. Finally, our findings highlight that the agent's movement notably influences excitement, the sense of social presence, and the perception of audiovisual spatial coherence. Together, our findings inform the design of applications with conversational agents, story-telling pieces, but also instructional sessions, in-depth tutorials, or guided tours. Future work should focus on tasks and specific sound sources. The nature of the task, ours being an entertainment-based 1-1 listening experience without high mental demand, may not need spatial coherence or congruence in realism. However, complex tasks or those with multiple sound sources might. Further research should, therefore, explore these variables within the context of movement and audio rendering.

Acknowledgments

This research is funded by the Carl-Zeiss-Foundation ("Breakthroughs 2020" program, `https://www.carl-zeiss-stiftung.de/programm/czs-durchbrueche`), in the CO-HUMANICS project, and by the the German Federal Ministry of Education and Research (BMBF) through the MULTIPARTIES project (grant no. 16SV8922).

A Digital Appendix

This paper's digital appendix is available at `https://doi.org/10.5281/zenodo.10458343`.

References

[1] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004. 3

[2] J. N. Bailenson and J. Blascovich. Avatars. In *Encyclopedia of human-computer interaction, Berkshire Publishing Group*, 2004. 1

[3] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis. Interpersonal distance in immersive virtual environments. *Personality and social psychology bulletin*, 29(7):819–833, 2003. 6

[4] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence*, 15(4):359–372, 2006. doi: 10.1162/pres.15.4.359 2

[5] P. Bala, R. Masu, V. Nisi, and N. Nunes. "when the elephant trumps": A comparative study on spatial audio for orientation in 360º videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300925 2

[6] D. R. Baum, S. Riedel, R. T. Hays, and A. Mirabella. *Training Effectiveness as a Function of Training Device Fidelty*. Honeywell Systems and Research Center, 1982. 2

[7] D. R. Begault and L. J. Trejo. 3-d sound for virtual reality and multimedia. Technical report, 2000. 2

[8] F. Biocca. The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of computer-mediated communication*, 3(2):JCMC324, 1997. 2

[9] F. Biocca, C. Harms, and J. Gregg. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*, pp. 1–9, 2001. 2

[10] H. Braren and J. Fels. A high-resolution head-related transfer function data set and 3d-scan of kemar. Technical report, Institute and Chair for Hearing Technology and Acoustics, RWTH Aachen University, 2020. doi: 10.18154/RWTH-2020-11307 5

[11] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl. A round robin on room acoustical simulation and auralization. *J. Acoust. Soc. Am.*, 145(4):2746–2760, 2019. doi: 10.1121/1.5096178 5

[12] Q. Cao, H. Yu, P. Charisse, S. Qiao, and B. Stevens. Is high-fidelity important for human-like virtual avatars in human computer interactions? *International Journal of Network Dynamics and Intelligence*, pp. 15–23, 2023. 3

[13] C. Dicke, V. Aaltonen, A. Rämö, and M. Vilermo. Talk to me: The influence of audio quality on the perception of social presence. *Proceedings of HCI 2010 24*, pp. 309–318, 2010. 2, 4

[14] I. Ekman. Psychologically motivated techniques for emotional sound in computer games. *Proc. AudioMostly*, pp. 20–26, 2008. 3

[15] S. Estupiñán, F. Rebelo, P. Noriega, C. Ferreira, and E. Duarte. Can virtual reality increase emotional responses (arousal and valence)? a pilot study. pp. 541–549, 2014. 3

[16] J. T. Fleming, R. K. Maddox, and B. G. Shinn-Cunningham. Spatial alignment between faces and voices improves selective attention to audio-visual speech. *The Journal of the Acoustical Society of America*, 150(4):3085–3100, 10 2021. doi: 10.1121/10.0006 415 2, 9

[17] M. Geronazzo, A. Rosenkvist, D. S. Eriksen, C. K. Markmann-Hansen, J. Køhlert, M. Valimaa, M. B. Vittrup, and S. Serafin. Creating an audio story with interactive binaural rendering in virtual reality. *Wireless Communications and Mobile Computing*, 2019:1–14, 2019. 2, 4

[18] A. Hayes. The experience of physical and social presence in a virtual learning environment as impacted by the affordance of movement enabled by motion tracking. *Electronic Theses and Dissertations, University of Central Florida*, 2015. 2

[19] M. R. Heinicke, J. F. Juanico, A. L. Valentino, and T. P. Sellers. Improving behavior analysts' public speaking: Recommendations from expert interviews. *Behavior Analysis in Practice*, 15(1):203–218, Mar 2022. doi: 10.1007/s40617-020-00538-4 2

[20] D. Higgins, K. Zibrek, J. Cabral, D. Egan, and R. McDonnell. Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. *Computers & Graphics*, 104:116–128, 2022. 3, 4, 8

[21] S. Huiberts. Captivating sound: the role of audio for immersion in games. *Doctoral thesis at the University of Portsmouth and Utrecht School of the Arts*, 2010. 8

[22] F. Immohr, G. Rendle, A. Neidhardt, S. Göring, R. R. Ramachandra Rao, S. Arevalo Arboleda, B. Froehlich, and A. Raake. Proof-of-concept study to evaluate the impact of spatial audio on social presence and user behavior in multi-modal vr communication. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, pp. 209–215, 2023. 2, 4, 8

[23] S. G. T. James C. Lester, Jennifer L. Voerman and C. B. Callaway. Deictic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13(4-5):383–414, 1999. doi: 10.1080/088395199117324 2

[24] H. I. Jo and J. Y. Jeon. Overall environmental assessment in urban parks: Modelling audio-visual interaction with a structural equation model based on soundscape and landscape indices. *Building and Environment*, 204:108166, 2021. 2

[25] F. Kafka and N. N. Glatzer. *Franz Kafka, the Complete Stories*. Schocken Books, New York, 1971. 6

[26] D. Kao, R. Ratan, C. Mousas, A. Joshi, and E. F. Melcer. Audio matters too: How audial avatar customization enhances visual avatar customization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. 3, 9

[27] M. Kay, L. A. Elkin, J. J. Higgins, and J. O. Wobbrock. mjskay/artool: Artool 0.11.0, Apr. 2021. doi: 10.5281/zenodo.4721941 6

[28] 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 *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 105–114, 2018. doi: 10.1109/ISMAR.2018.00039 2

[29] K. Kim, D. Maloney, G. Bruder, J. N. Bailenson, and G. F. Welch. The effects of virtual human's spatial and behavioral coherence with physical objects on social presence in ar. *Computer Animation and Virtual Worlds*, 28(3-4):e1771, 2017. 2

[30] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.*, 4(1):15–33, jan 2013. doi: 10.1109/T-AFFC.2012.16 3, 4, 8

[31] M. Kytö, K. Kusumoto, and P. Oittinen. The ventriloquist effect in augmented reality. In *2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 49–53. IEEE, 2015. 3

[32] L. Lam, M. Choi, M. Mukanova, K. Hauser, F. Zhao, R. Mayer, C. Mousas, and N. Adamo-Villani. Effects of body type and voice pitch on perceived audio-visual correspondence and believability of virtual characters. In *ACM Symposium on Applied Perception 2023*, pp. 1–11, 2023. 3, 4

[33] M. E. Latoschik and C. Wienrich. Congruence and plausibility, not presence: Pivotal conditions for xr experiences and effects, a novel approach. *Frontiers in Virtual Reality*, 3:694433, 2022. 2, 4

[34] M. Lewis and J. Jacobson. Game engines. *Communications of the ACM*, 45(1):27, 2002. 2

[35] L. Maffei, M. Masullo, A. Pascale, G. Ruggiero, and V. P. Romero. Immersive virtual reality in community planning: Acoustic and visual congruence of simulated vs real world. *Sustainable Cities and Society*, 27:338–345, 2016. 2

[36] D. Mal, E. Wolf, N. Döllinger, M. Botsch, C. Wienrich, and M. E. Latoschik. Virtual human coherence and plausibility–towards a validated scale. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 788–789. IEEE, 2022. 3, 4, 6

[37] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. Multimodality in vr: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–36, 2022. 3

[38] S. Masuko and J. Hoshino. Conversational locomotion of virtual characters. *Proceedings of the Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation*, p. 121, 2005. 2

[39] C. Mousas, D. Anastasiou, and O. Spantidi. The effects of appearance and motion of virtual characters on emotional reactivity. *Comput. Hum. Behav.*, 86:99–108, 2018. 3, 8

[40] C. Mousas, A. Koilias, B. Rekabdar, D. Kao, and D. Anastaslou. Toward understanding the effects of virtual character appearance on avoidance movement behavior. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 40–49, 2021. doi: 10.1109/VR50410.2021.00024 3

[41] S. Narang, A. Best, A. Feng, S.-h. Kang, D. Manocha, and A. Shapiro. Motion recognition of self and others on realistic 3d avatars. *Computer Animation and Virtual Worlds*, 28(3-4):e1762, 2017. 3, 4, 8

[42] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer. Flexible python tool for dynamic binaural synthesis applications. In *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017. 4

[43] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer. Flexible python tool for dynamic binaural synthesis applications. In *Audio Engineering Society Convention 142*, May 2017. 5

[44] M. G. Nelson, A. Koilias, C.-N. Anagnostopoulos, and C. Mousas. Effects of rendering styles of a virtual character on avoidance movement behavior. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 594–599, 2022. doi: 10.1109/ISMAR-Adjunct57072.2022.00123 3

[45] C. S. Oh, J. N. Bailenson, and G. F. Welch. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5, 2018. doi: 10.3389/frobt.2018.00114 1

[46] I. Pandzic and R. Forchheimer. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. Wiley, 2003. 5

[47] R. E. Pastore, J. D. Flint, J. R. Gaston, and M. J. Solomon. Auditory event perception: The source—perception loop for posture in human gait. *Perception & psychophysics*, 70:13–29, 2008. 3

[48] A. V. D. Pütten, N. Krämer, J. Gratch, and S.-H. Kang. "it doesn't matter what you are!" explaining social effects of agents and avatars. *Comput. Hum. Behav.*, 26:1641–1650, 2010. 3, 9

[49] T. Randhavane, A. Bera, K. Kapsaskis, K. Gray, and D. Manocha. Fva: Modeling perceived friendliness of virtual agents using movement characteristics. *IEEE Transactions on Visualization and Computer Graphics*, 25(11):3135–3145, 2019. doi: 10.1109/TVCG.2019.2932235 3

[50] S. L. Rogers, R. Broadbent, J. Brown, A. Fraser, and C. P. Speelman. Realistic motion avatars are the future for social interaction in virtual reality. In *Frontiers in Virtual Reality*, 2021. 3

[51] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 6

[52] J. A. Russell and M. Bullock. Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. *Journal of personality and social psychology*, 48(5):1290, 1985. 3

[53] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47:675–705, Sept. 1999. 5

[54] J. A. Schroeder. Flights of fancy: The art and science of flight simulation. *Principles and practice of aviation psychology*, p. 435, 2002. 2

[55] A. Siddig, P. W. Sun, M. Parker, and A. Hines. Perception deception: Audio-visual mismatch in virtual reality using the mcgurk effect. *AICS*, 2019:176–187, 2019. 3, 4

[56] P. Skalski and R. Whitbred. Image versus sound: A comparison of formal feature effects on presence and video game enjoyment. *PsychNology Journal*, 8(1), 2010. 2

[57] R. Skarbez, F. P. Brooks, Jr., and M. C. Whitton. A survey of presence and related concepts. *ACM Comput. Surv.*, 50(6), nov 2017. 2

[58] S. Soto-Faraco, J. Lyons, M. Gazzaniga, C. Spence, and A. Kingstone. The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive brain research*, 14(1):139–146, 2002. 3

[59] A. Tajadura-Jiménez, M. Basia, O. Deroy, M. Fairhurst, N. Marquardt, and N. Bianchi-Berthouze. As light as your footsteps: Altering walking sounds to change perceived body weight, emotional state and gait. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 2943–2952. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2702123.2702374 3

[60] S. Thomas, Y. Ferstl, R. McDonnell, and C. Ennis. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 11–20, 2022. doi: 10.1109/VR51125.2022.00018 3, 4, 8

[61] M. Tsepapadakis and D. Gavalas. Are you talking to me? an audio augmented reality conversational guide for cultural heritage. *Pervasive and Mobile Computing*, 92:101797, 2023. 2

[62] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik. Room acoustics simulation for multichannel microphone arrays. 01 2010. 5

[63] R. Warp, M. Zhu, I. Kiprijanovska, J. Wiesler, S. Stafford, and I. Mavridou. Validating the effects of immersion and spatial audio using novel continuous biometric sensor measures for virtual reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 262–265, 2022. doi: 10.1109/ISMAR-Adjunct57072.2022.00058 2

[64] F. Weidner, G. Boettcher, S. A. Arboleda, C. Diao, L. Sinani, C. Kunert, C. Gerhardt, W. Broll, and A. Raake. A systematic review on the visualization of avatars and agents in ar vr displayed using head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2596–2606, 2023. doi: 10.1109/TVCG.2023.3247072 2, 3, 4

[65] Z.-M. Ye, J.-L. Chen, M. Wang, and Y.-L. Yang. Paval: Position-aware virtual agent locomotion for assisted virtual reality navigation. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 239–247, 2021. doi: 10.1109/ISMAR52148.2021.00039 2

[66] N. Yee and J. Bailenson. The proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, 33(3):271–290, 2007. 9

[67] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, and W. Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 547–556, 2019. doi: 10.1109/VR.2019.8797719 2, 8

[68] K. Zibrek, J. Cabral, and R. McDonnell. Does synthetic voice alter social response to a photorealistic character in virtual reality? In *Proceedings of the 14th ACM SIGGRAPH Conference on Motion, Interaction and Games*, MIG '21. Association for Computing Machinery, New York, NY, USA, 2021. 3, 4, 8

[69] K. Zibrek, E. Kokkinara, and R. Mcdonnell. The effect of realistic appearance of virtual characters in immersive environments - does the character's personality play a role? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1681–1690, 2018. doi: 10.1109/TVCG.2018.2794638 8

[70] K. Zibrek, S. Martin, and R. McDonnell. Is photorealism important for perception of expressive virtual humans in virtual reality? *ACM Transactions on Applied Perception (TAP)*, 16(3):1–19, 2019. 2, 4, 8